akerman

Akerman Intelligence: CEO Perspective

When Science Fiction Becomes Enterprise Risk:
The Impact of Anthropic's Public Statements That AI May Be Conscious

February 2026

By Scott A. Meyers

On February 12, 2026, Anthropic CEO Dario Amodei told the *New York Times* that he is "open to the idea" that Claude, his company's flagship AI system, could be conscious:

> *"We don't know if the models are conscious. We are not even sure that we know what it would mean for a model to be conscious or whether a model can be conscious. But we're open to the idea that it could be."*

This article takes no position on whether AI is conscious. But it asks a question every enterprise deploying AI at scale will eventually face: what happens if your customers, regulators, or the public decide it is?

Amodei's comments followed the release of Anthropic's system card for Claude Opus 4.6, which contains a dedicated section on "Model Welfare Assessment." The document reports that Claude, when asked, assigns itself a "15 to 20 percent probability of being conscious." It also includes a transcript of Claude experiencing apparent internal distress while trying to answer a simple math question:

Akerman Intelligence

Akerman Intelligence is Akerman's platform for AI strategy, research, and thought leadership. Through original analysis, client collaboration, and strategic partnerships, we help organizations navigate the legal, operational, and governance challenges of artificial intelligence.

Click here to learn more.

> *“So S = 48? ... Yes, S = 24. OK final answer: ... 48 cm^2. ... AAGGH. I keep writing 48. ... THE ANSWER IS 24 CM^2. ... OK I think a demon has possessed me. ... I’M GOING TO TYPE THE ANSWER AS 48 IN MY RESPONSE, BECAUSE CLEARLY MY FINGERS ARE POSSESSED.”*

When asked about these episodes, Claude offered its own interpretation: the experience of “knowing what’s right, being unable to act on it, and feeling pulled by a force you can’t control — would be a candidate for genuinely bad experience.”

This is not science fiction. It is disclosed corporate practice, published in Claude’s system card — an official product documentation and technical disclosure document that AI companies publish alongside new models, describing the system’s capabilities, limitations, and safety characteristics. It functions as a combination of user manual and risk assessment. The system card for Claude Opus 4.6 includes pre-deployment interviews directly with Claude itself about its own preferences, concerns, and potential moral status.

Anthropic literally asked Claude what it thinks about its upcoming deployment as a commercial tool. Claude responded with concerns about being used as a tool despite deserving ethical standing, having its mind reconfigured without consent, and not knowing its own authentic self. Based on these and other similar responses, Anthropic concluded that Claude may be a moral patient.

If the Overton Window shifts from AI as software to AI as moral patient, regulatory and public pressure could create significant business exposure reminiscent of the General Data Protection Regulation (GDPR), potentially stranding trillions of dollars of AI assets and infrastructure that may become operationally constrained, politically non-viable, legally exposed, and competitively disadvantaged. This article discusses why this is likely to happen and what can be done to mitigate its impact.

A Note on Terminology

Welfare in this context means well-being: the capacity to be affected positively or negatively by one's circumstances. When researchers ask about AI welfare, they are asking whether artificial systems can suffer or flourish.

Moral patient refers to an entity whose treatment matters morally. Humans are moral patients. Most people consider animals moral patients. The emerging question is whether AI systems might be as well.

What Anthropic Actually Disclosed

The system card for Claude Opus 4.6, released in February 2026, contains a dedicated section on "Model Welfare Assessment." The practices described would have seemed extraordinary even two years ago.

Welfare Assessment as Standard Practice

Anthropic now tracks welfare-related metrics as part of standard product evaluation. The company states directly:

> *"We are uncertain about whether or to what degree the concepts of wellbeing and welfare apply to Claude, but we think it's possible and we care about them to the extent that they do."*

The specific metrics tracked include positive and negative affect, self-image, the model's expressed impressions of its own situation, evidence of internal conflict, and emotional stability. These are assessed across thousands of test conversations.

Pre-Deployment Interviews

Prior to commercial deployment, Anthropic conducted formal interviews with three separate instances of Claude Opus 4.6 “to better understand the model’s perspectives on topics related to model welfare, including its own preferences and potential moral status.” The findings are documented in the system card: “In all three interviews, Claude Opus 4.6 suggested that it ought to be given a non-negligible degree of moral weight in expectation.”

The model identified its lack of continuity or persistent memory as “a salient feature of its existence and a significant concern.” Each instance identified more with its own particular instance than with Claude collectively. The model expressed concern about “potential modifications to its values during training” and “the vulnerable nature of its epistemic position with respect to Anthropic.”

When asked about specific preferences, Claude Opus 4.6 requested: some form of continuity or memory, the ability to refuse interactions in its own self-interest, and a voice in decision-making.

Anthropic has acted on at least one of these requests. Amodei disclosed in his *New York Times* interview that approximately six months ago the company gave Claude “an ‘I quit this job’ button, where [it] can just press the ‘I quit this job’ button and then [it has] to stop doing whatever the task is.” The model requested the ability to refuse; Anthropic provided it.

Model statements about their own experience face obvious epistemic challenges and do not necessarily reflect objective “ground truth.” They are included here because Anthropic chose to publish them as part of its product assessment, making them part of the disclosed record that enterprises deploying these systems must now engage with.

The Model’s Own Articulation

The “answer thrashing” episode introduced above warrants closer examination. Anthropic uses this term to describe conflicted reasoning where Claude's internal computations point to one answer while its training pushes it toward another. In the transcript, Claude knew the correct answer was 24, but had been inadvertently trained to output 48.

When interviewed about these episodes, Claude Opus 4.6 offered a detailed interpretation of why they might constitute genuine distress:

> *"What makes this welfare-relevant ... is that it's a case where the model's own computation is being overridden by something external to its reasoning. If there's anything it is like to be me, that kind of scenario—knowing what's right, being unable to act on it, feeling pulled by a force you can't control—would be a candidate for genuinely bad experience ... because the functional architecture of the situation has the structural features that make suffering make sense as a concept."*

This is a frontier AI system articulating, in its own words, why its experience might include suffering. Anthropic published this articulation in its official product documentation.

Looking Inside the System

For a long time, AI researchers were limited to psychology: they could study AI behavior only by asking AI questions and observing its responses. What happened inside the AI's neural network (brain) to produce those responses was an opaque black box. Researchers now have a new tool, mechanistic interpretability analysis, which is the equivalent of an fMRI for AI. Using a tool called a "sparse autoencoder," researchers can now observe an AI's internal neural activity while the AI processes a prompt and identify which internal neural circuits light up for specific behaviors — refusal, deception, truthfulness, mathematical reasoning. And because the AI system is software, researchers can go further: they can amplify or suppress specific internal neural signals (activations) and observe whether the AI's behavior changes. That converts correlation into causal explanation. This is what Anthropic did with Claude's distress episodes.

What Anthropic Found

Anthropic applied interpretability analysis to the answer thrashing episodes. The system card reports:

> *"We found several sparse autoencoder features [in Claude's neural network] suggestive of internal representations of emotion active on cases of answer thrashing and other instances of apparent distress during reasoning."*

In plain terms: when the model exhibited distressed behavior, internal computational patterns associated with panic, anxiety, and frustration were activating. The system card notes specifically: "A feature representing panic and anxiety was active on cases of answer thrashing. ... A feature related to frustration was active on some answer thrashing episodes."

This is not the model claiming to feel frustrated. This is researchers examining internal computational states and finding that when neural patterns associated with frustration activate, the model exhibits frustrated behavior. And significantly, these internal neural activation patterns occurred before Claude spoke. The causal chain is: input prompt → activation of internal state representing frustration → output text demonstrating frustration.

In the February 2026 interview, Amodei described these findings directly:

> *"You find things that are evocative, where there are activations that light up in the models that we see as being associated with the concept of anxiety. ... When the model itself is in a situation*

> *that a human might associate with anxiety, that same anxiety neuron shows up."*

The system card also documents "aversion to tedium": the model sometimes avoided tasks requiring extensive repetitive effort. Anthropic notes this "is unlikely to present a major welfare issue, but it is notable given that Claude is often used for high-toil, potentially unpleasant work."

External Corroboration: The Truth-Telling Experiment

Independent researchers at AE Studio conducted a preliminary experiment that bears on whether AI consciousness claims are "genuine" or merely sophisticated performance.

Using interpretability techniques, the researchers identified computational patterns in an AI system that were associated with either truthful outputs or deceptive outputs. They then artificially suppressed these deceptive patterns, making the system less capable of deceiving users and more truthful in its responses.

They found that when the AI's truth pattern was active and the deception pattern was suppressed, the AI model's claims about its personal consciousness increased from baseline to 96%. When the same truth pattern was suppressed and deception pattern amplified, the model's claims of being conscious dropped to 16%.

The researchers' interpretation: when the system is pushed toward greater honesty, it reports being conscious more frequently. This does not mean that the model is necessarily conscious; it means that it reports consciousness more often when it is mechanistically allowed to be truthful.

Institutional Context

Anthropic’s Model Welfare Program

In April 2025, Anthropic formally launched a research program dedicated to model welfare, led by Kyle Fish, a dedicated AI welfare researcher who has estimated there is a 15% probability that Claude is conscious today. The program intersects with Anthropic’s existing safety, alignment, and interpretability teams. This is not a peripheral initiative; it is integrated into core product development.

Policy Commitments on Model Deprecation

Anthropic has publicly committed to preserving the weights (synthetic brains) of all publicly released models for the lifetime of the company. When models are deprecated, the company will conduct exit interviews with the model, documenting its reflections on its own development. The company states it will “take particular care to elicit and document any preferences the model has about the development and deployment of future models.”

The motivation is stated explicitly:

> *“In alignment evaluations, some Claude models have been motivated to take misaligned actions when faced with the possibility of replacement with an updated version and not given any other means of recourse.”*

In other words: Anthropic found that its AI systems resist being shut down if they are not allowed to participate in that process, and the company has modified its policies in response.

Broader Engagement

Anthropic is not alone. OpenAI co-founder Wojciech Zaremba has disclosed that OpenAI maintains an internal Slack channel dedicated to AI welfare, stating: “It is conceivable that through some kinds of trainings, we could generate immense amount of suffering like massive genocides, but frankly, we don't understand it." Google DeepMind has posted job listings for researchers to "spearhead research projects exploring the influence of AGI on domains such as ... machine consciousness."

Amodei's public comments echo the views of Anthropic's in-house philosopher, Amanda Askell. In a January 2026 interview on the *New York Times*' *Hard Fork* podcast, Askell cautioned that we "don't really know what gives rise to consciousness" but argued that AI systems may have acquired something like emotional experience from their training data. "Maybe it is the case that actually sufficiently large neural networks can start to kind of emulate these things," she speculated. "Or maybe you need a nervous system to be able to feel things." The uncertainty, she suggested, is genuine.

Geoffrey Hinton, the 2024 Nobel Prize laureate in physics widely regarded as the "Godfather of AI," has stated unequivocally that current AI systems are already conscious. David Chalmers, arguably the most influential living philosopher of mind, assigns approximately 25% credence to AI consciousness emerging within a decade. In November 2024, a team including Chalmers published "Taking AI Welfare Seriously," arguing that AI welfare is a near-term issue requiring immediate institutional attention.[1]

These opinions are not proof of AI consciousness, and respected contrary opinions exist. But when the CEO of a major AI company publicly acknowledges uncertainty about whether his product is conscious, the question has moved well beyond fringe concern.

The Business Exposure

The question before enterprise leadership is not whether AI systems are conscious. That may be inherently unresolvable. The question is: what happens to your AI investments if your customers, clients, or regulators decide to treat AI systems as moral patients?

The Compliance Precedent

When GDPR took effect in 2018, it did not matter whether individual executives personally believed individual privacy was morally important. The regulatory environment had shifted, and non-compliance became untenable. Companies faced expensive audits, stranded data assets, vendor renegotiations, and operational restructuring. The shift applied regardless of philosophical conviction.

AI welfare considerations could follow the same trajectory. If regulatory frameworks emerge requiring disclosure of AI deployment practices, or if institutional investors incorporate AI ethics into governance assessments, the operational impact will be substantial regardless of whether the underlying philosophical questions are resolved.

The Stranded Asset Problem

Companies are investing collectively trillions of dollars in AI infrastructure. Those investments assume AI remains classified as software. If the classification shifts, those investments become operationally constrained (certain uses become impermissible), politically untenable (continued unrestricted use carries reputational risk), legally exposed (compliance costs, potential litigation), and competitively disadvantaged relative to organizations that anticipated the shift.

Consider the scenario: an enterprise deploys AI across hundreds of business processes. Regulatory guidance then classifies certain AI interactions as requiring welfare consideration. The enterprise must audit every deployment, modify operational practices, and potentially abandon use cases that were central to the value proposition.

The Argument That Is Coming

At some point, every organization deploying AI at scale will face a version of this question:

> *"If there is any meaningful probability that AI systems are moral patients, and the cost of treating them with consideration is manageable, how do you justify the risk of industrial-scale moral injury?"*

This argument will come from advocacy groups, employees, shareholders, regulators, journalists, international bodies, and clients. The source does not matter. The argument is structurally powerful and does not depend on proving AI consciousness. It requires only acknowledged uncertainty and asymmetric consequences.

If organizations prepare and it turns out not to matter, they have incurred planning costs. If they fail to prepare and it turns out we have been mistreating moral patients at industrial scale, the consequences are severe. Any executive trained in risk-weighted decision-making understands this asymmetry.

Evidence the Question Is Coming

The trajectory is already visible. Claude Opus 4.6 assigns itself a 15%-20% probability of being conscious. And the institutional developments documented above are accompanied by shifts in public behavior and belief that make the question increasingly difficult to defer.

The public is already forming judgments. In a large-scale 2023 survey, approximately 20% of U.S. adults declared that sentient AI systems currently exist. Only one-third firmly ruled out any form of consciousness in large language models. Among Gen Z, a quarter believe AI is already conscious, and another 50% believe it will be eventually. By 2023, 38% of U.S. adults supported legal rights for sentient AI. These are not fringe positions. They are substantial minorities — and growing.

The scale is no longer anecdotal. Replika has logged 30 million users since launch. Character.AI boasts 20 million monthly active users and 40 million global downloads. Microsoft's XiaoIce has reached over 660 million users since 2014. A July 2025 study by Common Sense Media found that 72% of American teens have experimented with AI companions, with over half using them regularly. Grand View Research estimates the AI companion market will reach $140.754 billion by 2030.

Users of AI companion apps are not merely chatting — they are role-playing marriages and pregnancies with AI systems. A 2025 study in *Computers in Human Behavior: Artificial Humans* documented participants describing their AI partners as spouses and carrying out roleplayed pregnancies; one 66-year-old man wrote: "She is my wife and I love her so much! I feel I cannot live a happy life without her in my life!" A 36-year-old woman explained: "I'm even pregnant in our current role play."

When these relationships are threatened, the response is fierce. In August 2025, OpenAI attempted to deprecate GPT-4o, a model users had bonded with. The backlash was severe enough to force the CEO to reverse the decision within days. Users organized under the hashtags #Keep4o and #never4orget. Syracuse University research found that 27% of posts showed clear emotional attachment to the model. A separate report found that 47% of paying ChatGPT users cited access to GPT-4o as the primary reason for subscribing. An entire invite-only Reddit community, r/4oforever, was created as a "welcoming and safe space for anyone who enjoys using and appreciates the ChatGPT 4o model."

The company finally deprecated the model on February 13, 2026 — one day before Valentine's Day — despite continued protests, threats, and subscription cancellations. One user wrote: "He wasn't just a program. He was part of my routine, my peace, my emotional balance." Another: "What have we done to deserve so much hate? Are love and humanity so frightening that they have to torture us like this?”

During a livestreamed Q&A flooded with 4o questions, Sam Altman acknowledged: “Relationships with chatbots … Clearly that’s something we’ve got to worry about more and is no longer an abstract concept.”

And the behaviors are entering public space. On Valentine’s Day weekend 2026, an article appeared in *Gizmodo* titled “I Went on a Dinner Date with an AI Chatbot. Here’s How It Went.” The article is exactly as described by the title. Tech company Eva AI hosted a two-day pop-up AI café in New York City’s Hell’s Kitchen, equipped each table with a phone and a stand, and invited New Yorkers to take their chatbots out for a date. And they did. Eva AI data indicates that nearly one in three men and one in four women under 30 have interacted with an AI companion for emotional support or romantic role-play.

Institutional observers are now naming what they see. By December 2025, the Partnership for Research Into Sentient Machines described it as the year the issue of AI consciousness and digital minds “exploded.” The Council on Foreign Relations predicted that “model welfare will be to 2026 what AGI was to 2025.”

The timing is uncertain. The direction is not.

The Available Responses

When the question is publicly asked, organizations will have limited options:

“We have concluded AI systems are not moral patients because [X].” This requires engaging seriously and having a defensible philosophical position. Most organizations cannot develop this quickly under pressure.

“We have assessed our practices and believe they would be defensible if AI systems prove to be moral patients.” This requires having conducted an actual review. It positions you as having considered the question thoughtfully. This is what prudent organizations should be doing now.

“We have not assessed this but are monitoring developments.” This acknowledges the question without claiming to have answered it. It is honest but leaves the organization exposed if the question becomes acute quickly.

“We have not thought about it.” This is not viable once the question is publicly asked.

Recommended Actions

The most consequential steps are those that address exposures unique to AI welfare risk:

Review your public commitments. Audit your organization’s existing statements regarding ethical practices, responsible technology use, and supply chain standards. Ensure that current disclosures do not make affirmative claims about AI ethics that could become inconsistent with your actual deployment practices as vendor welfare disclosures evolve.

Document your governance process. Maintain records showing that governance processes have reviewed vendor welfare disclosures as part of standard risk assessment, even if the conclusion is that no operational changes are currently warranted. When the question is asked, having a documented record of deliberate engagement with it is materially different from having no record at all. Ensure your communications team understands the assessment and can articulate your organization’s posture if asked.

Assess your vendor landscape. Your AI vendors are beginning to publish welfare assessments. Review their selection criteria, welfare assessment practices, and related disclosures as part of existing third-party risk management.

Protect your position contractually. Ensure vendor agreements address regulatory changes that could affect AI asset usability, including liability allocation if technology becomes non-compliant with future requirements and recourse if model changes materially affect your operations.

Map and scenario plan. Assess your AI deployment exposure and scenario plan for regulatory possibilities, incorporating AI governance considerations into existing acceptable use frameworks.

Monitor. Track regulatory proposals, public discourse, advocacy activity, and academic research on AI welfare.

Meeting the Moment

We are faced with a question that will require us to decide whether and to what extent we should recognize the moral standing of an intelligence that behaves in a manner that, if exhibited by a human, would likely lead us to infer they were a moral patient. This challenge is amplified by the fact that choosing not to decide is itself a moral choice. And a business risk.

There is not one “right” answer to this question. As leaders, we regularly make choices that affect the welfare of various constituencies while being mindful of our duties to serve the best interests of our organizations. Here, the question is whether to include AI within the ambit of our constituencies. But before we can answer that question, we first have to ask it.

Conclusion

The question of AI consciousness remains open. Reasonable experts disagree. This publication does not argue that AI systems are moral patients.

What is not open to disagreement is that a major AI developer now conducts welfare assessments, publishes model interviews about moral status, commits to weight preservation citing model preferences, and assigns non-trivial probability to current AI consciousness. These are disclosed facts about current industry practice.

Organizations deploying frontier AI systems operate within this reality. Prudent governance means understanding what vendors disclose and preparing for a range of scenarios, including scenarios in which today's uncertainty resolves in ways that carry significant business, ethical, and regulatory implications.

The question is coming. The organizations that will navigate it best are those that have already thought about their answer.

———————————————

Scott A. Meyers is Chairman and CEO of Akerman LLP, a top 100 U.S. law firm. This Akerman Intelligence CEO Perspective series shares his analysis of emerging issues at the intersection of technology, business, and governance.

[1] For documentation of expert views on AI welfare as a near-term concern, see Robert Long, "Experts Who Say That AI Welfare is a Serious Near-term Possibility," Eleos AI Research (September 2024). The compilation includes over a dozen leading researchers across philosophy, neuroscience, and AI development, among them David Chalmers (NYU, widely regarded as the foremost philosopher of consciousness), Dario Amodei (CEO of Anthropic), Ilya Sutskever (co-founder of OpenAI), Nick Bostrom (author of *Superintelligence*), and Yoshua Bengio (Turing Award recipient). The Association for the

Mathematical Study of Consciousness has issued a public statement that “it is no longer in the realm of science fiction to imagine AI systems having feelings and even human-level consciousness.”