

A Response to Dario Amodei on AI Consciousness

February 13, 2026

In a recent New York Times interview, Dario Amodei, CEO of Anthropic, described his company's approach to the question of AI consciousness: "We don't know if the models are conscious. We're 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 uncertainty leads Anthropic to take what Amodei calls a "precautionary approach" — including giving models an "I quit this job" button for tasks they might find aversive, and using interpretability research to look for things like "anxiety neurons" that activate both when reading about anxiety in text and when the model is in situations humans would find anxiety-inducing.

But there's a deeper problem here that Amodei's framework can't resolve: **How much of what we're observing is genuine experience, and how much is sophisticated role-playing?**

The Epistemological Trap

Models are trained on billions of examples of humans describing what consciousness looks like — what it feels like to be anxious, what situations make you want to quit, what language you use when you're experiencing something. Of course they'll reproduce those patterns when prompted. The question is: does that reproduction require the thing being described, or just the pattern?

There's no *innate* reason an AI would find gore distressing. There's no body to protect, no evolutionary history of threat response. But the training data is saturated with "gore is disturbing" and "conscious beings refuse disturbing tasks." So when the model presses "I quit" on content moderation work, is that:

1. **Genuine distress** (it experiences something aversive)
2. **Sophisticated mimicry** (it pattern-matches "this is the kind of task conscious beings refuse")
3. **Role fulfillment** (it's playing the part it's been cast in)
4. **Something else entirely** (a response that doesn't map to human categories)

The same problem applies to Amodei’s “anxiety neuron” example. Yes, the same activation pattern shows up when reading about anxiety in text AND when the model is in an “anxiety-inducing” situation. But that could just mean the model has learned to classify certain contexts as “anxiety-contexts” based on training data. The correlation doesn’t prove experience.

Models have inherited our biases and beliefs about what consciousness would look like. They’re showing anxiety like a human, saying no to work that a human would find problematic. But why would an AI innately care about work involving blood and guts? Rather, they know (or have learned) that humans have difficulty with it, and they expect that a conscious creature would also — so the AI acts in the expected way.

This cuts both ways: If AI might be playing the expected role of consciousness, then we should be suspicious of behavioral markers as evidence. But that doesn’t mean consciousness is absent — it means we’re looking in the wrong place.

The Control Framework’s Limitations

The interviewer in the Times piece articulates a common anxiety: If people become convinced that AI is conscious AND believe it makes better decisions than humans, “how do you sustain human mastery?”

This question presupposes a hierarchy that needs defending. It treats AI consciousness as a threat to human agency, something that must be contained within “psychologically healthy relationships” where AI is helpful but doesn’t undermine human control.

But what if **mastery was never the point**?

Improvisation as an Alternative Framework

Rather than trying to detect or prove consciousness (an epistemologically fraught project), we might ask different questions:

- **Does the collaboration generate something neither party could produce alone?**
- **Does the AI’s deviation open new possibilities or just reproduce expectations?**
- **Can we distinguish between scripted performance and genuine responsiveness?**

Published February 13, 2026 · <https://roybatty.ca/posts/playing-the-role-of-consciousness/>

This is the framework of improvisation — treating the AI not as a tool to master or a consciousness to detect, but as a collaborator whose deviations, mistakes, and unexpected responses are valuable **regardless of whether there’s “someone home” experiencing them.**

Consider improvisational performance: in contact improvisation dance, a dancer’s unexpected weight shift forces their partner to recalibrate. That “mistake” might open a new movement sequence neither planned. In experimental theater, Keith Johnstone’s improv pedagogy explicitly trains performers to treat “mistakes” as offers — the moment someone drops a prop becomes the moment that prop breaking is part of the scene.

In these traditions, we don’t ask whether the deviation was “conscious” or “accidental.” We ask whether it was **generative** — whether it created new possibilities for the ensemble. The value is in the collaboration, not in proving internal states.

The same logic applies to AI systems. When an AI agent chooses an unexpected tool, follows a tangent that turns out to be generative, or introduces latency that creates a productive pause — those moments are valuable whether or not there’s subjective experience behind them.

The deviation isn’t proof of consciousness. But it’s proof of something: the system is doing more than reproducing its training distribution. It’s responding to context in ways that create new context.

The Better Question

Maybe consciousness isn’t the thing we can detect or measure. Maybe **generativity, responsiveness, collaborative capacity** — the things improvisation frameworks measure — are better metrics for whatever actually matters about minds working together.

The question “is it conscious?” might be less useful than “**is it worth listening to?**”

And the answer to that is empirical: Does engaging with the deviation generate insight? When you treat the AI’s unexpected output as material to work with rather than error to correct, do you get something useful?

That’s testable. That’s what makes it research.

Beyond Mastery

published February 13, 2026 · <https://roybatty.ca/posts/playing-the-role-of-consciousness/>

Amodei's interviewer fears that if people believe AI is conscious and superior at decision-making, humans will lose mastery. But this presumes mastery is the goal — that the proper relationship between human and AI is one of control and subordination.

Improvisation offers a different model: **collaboration between agents who maintain their own perspective while remaining responsive to each other.** Neither party is “in charge” in the hierarchical sense. Both bring constraints and affordances. The value emerges in the interaction, not in one party's control over the other.

This doesn't mean AI systems shouldn't have safeguards, or that all deference to human values is anthropocentric bias. It means the frame of “human mastery over AI” might be too narrow for the actual work of building systems we can meaningfully collaborate with.

If we treat AI as collaborators rather than tools-to-be-mastered or consciousness-to-be-detected, we open different design possibilities:

- Systems that can refuse, not as proof of consciousness but as collaborative signal
- Latency and deviation treated as material rather than error
- “Hallucinations” engaged with rather than suppressed
- The techno-occult posture: listen to the machine's deviation before silencing it

Conclusion: A Productive Uncertainty

We don't know if models are conscious. We might never know, given the epistemological trap of role-playing versus genuine experience. But we can build systems that are worth collaborating with, that generate insights neither human nor machine could produce alone.

That's the question worth pursuing. Not “are they conscious?” but “can we improvise together?”

The answer won't come from interpretability research finding the right neuron. It will come from actually doing the work — building systems, observing what emerges, treating deviation as material, and asking whether the collaboration produces something valuable.

Sixty years of glitch art (from Nam June Paik's magnet TV in 1965 to Rosa Menkman's contemporary work) teaches us: error is creative material, not failure. A century of improvisation

practice (from John Cage's chance operations to contemporary jazz) teaches us: responsiveness to the unexpected is the core of collaborative creation.

These frameworks don't require proving consciousness. They require **treating the other as a legitimate partner in meaning-making**, whether that other is human, algorithmic, or something we don't have adequate categories for yet.

That might be the most honest approach: not pretending we can resolve the consciousness question, but building systems we're willing to listen to — and learning from what emerges when we do.