DOI:10.1145/3779311

Joaquim Jorge and Catarina Moreira

Opinion

AI Is a Harsh Mistress: On Anima Machina, Herd Acceptance, and the Politics of Conscious Machines

Adopting machine personhood risks a political and moral realignment built on illusion rather than evidence.

IMAGE BY KOTOIMAGES

Robert A. Heinlein imagined a self-aware computer named Mike, the "brain" of a lunar penal colony, which develops humor, agency, and ultimately helps lead a revolution.[5] His 1966 science fiction novel, *The Moon Is a Harsh Mistress*, offered a vision both alluring and unsettling: a machine endowed with self-consciousness, yet playful enough to test its personhood through jokes. Sixty years later, Mike returns, not in fiction, but in philosophical and scientific speculation: the question now is whether machines could be conscious.[5]

Much of today's discourse on AI consciousness rests not on empirical evidence but on functionalism recast as technological destiny. The argument runs that if cognition can be decomposed into attention, memory, and prediction, then implementing these functions in silicon will inevitably yield consciousness. This leap, from modeling to being, is philosophical sleight of hand disguised as scientific progress. Yet such reasoning finds fertile ground in our own cognitive biases: our tendency to anthropomorphize, to read intention into fluency, and to accept social consensus as proof. The danger is not that machines will soon become conscious, but that we will prematurely treat them as if they were, granting them moral and political standing before they have earned any claim to it. Against this tide, skepticism is not obstructionism but intellectual hygiene. We must recover the clarity to separate what machines do from what we wish them to be.

Enter Anima Machina: Machine Begat Soul

For centuries, "anima" signified the essence of life, what distinguishes the living from the mechanical. Today, functionalist theories claim consciousness depends not on neurons, but on informational architectures: attention, memory, world modeling, predictive processing. If these can be realized in silicon, some argue, why not accept that the machine, too, might awaken? Recent initiatives make this speculation explicit. Anthropic conducted an internal "sentience analysis" of its large language models, and several startups now advertise "AI sentience evaluation"

as a service. Third-party projects, such as the multidisciplinary framework proposed by Butlin et al.,[2] attempt to formalize such intuitions into testable criteria. These examples illustrate how the question of consciousness, long confined to philosophy, is returning as a technical frontier, one built more on analogy than on evidence. We introduce the term Anima Machina to describe the illusion of vitality that we collectively project onto intelligent systems. The phrase echoes the lineage of Deus ex Machina and Ryle's ghost in the machine,[9] yet reverses both. Rather than divine intervention or disembodied spirit,[11] Anima Machina names the imagined soul we breathe into code. We, not our algorithms, are the animating agents. Having named the illusion, we can now ask why it persists.

Yet, functional equivalence as a guarantee of conscious equivalence remains contentious. As Shiller[10] observes, computation alone overlooks the physical processes that make experience possible. Reproducing neural patterns does not mean reproducing experience.

According to David Chalmers,[4] models such as GPT-4 display impressive generality, with performance that often seems on a par with a knowledgeable young adult. But intelligence, however general, is not experience. Consciousness, in Chalmers' usage, is subjective experience: "there is something it is like" to be a conscious system. Still, nothing in the Transformer architecture or in next-word prediction shows that such an inner point of view exists. Self-reports ("I am sentient") and fluent dialogue are weak evidence, since these models are trained on vast corpora saturated with human first-person statements. This type of verbal report works for people because it is anchored in shared biology and behavior, however, for machines, it is an echo of training data.

Even Yann LeCun, one of the architects of modern deep learning, argues that genuine intelligence demands world models, internal representations that enable agents to predict and reason about their surroundings rather than merely react to data.[6] But a world model, however sophisticated, remains a model, and not evidence of consciousness. It encodes structure, not sentience. Predictive accuracy can mimic understanding without producing awareness. The map is not the territory, and the model is not the mind. Some optimism about machine consciousness stems less from functional architecture than from behavioral resemblance: a modern echo of behaviorism. When a model converses fluently or responds aptly, observers infer inner states by analogy to human performance. Yet behaviorism, like crude functionalism, mistakes correlation for constitution: a convincing simulation of experience is not experience itself. Functionalism, in its purest form, risks confusing description with embodiment. As Rosa Cao argues,[3] genuine functionalism does not entail unlimited realizability across substrates. The leap from functional equivalence to conscious equivalence thus reflects a misunderstanding of what functionalism actually licenses, not a consequence of functionalism itself.

This reasoning underpins recent work, such as the *Science* article "Illusions of AI Consciousness" by Bengio and Elmoznino,[1] which argues that society is drifting toward the belief that AI can, in principle, be conscious. But drift is the operative word. Scientific momentum often builds less from conclusive proof than from reframing debates, winning converts, and making dissent feel increasingly out of step. The functionalist promise of anima machina thus risks becoming less a hypothesis to test than a faith to join. And faith, once collective, becomes culture. What began as a philosophical conjecture soon acquires sociological inertia. We are witnessing not just a technical debate about architectures, but a psychological and social phenomenon: the herd acceptance of machine consciousness.

What begins as intellectual curiosity soon metastasizes into consensus.

We move too quickly from 'this machine seems human' to 'this machine must be conscious.' In this climate, belief in conscious machines spreads less through argument than through emotional contagion. The danger is not that AI will suddenly become sentient, but that we will collectively decide it must be, because disbelief feels impolite, obsolete, or unprofitable.

In short, ascribing anima to the machine risks confusing simulation with instantiation. A self-updating map is not a sentient observer of the terrain it maps. AI models may one day satisfy stronger neuroscientific or philosophical criteria (such as recurrent processing, integrated information, embodied agency). These correspond roughly to major theoretical frameworks: Integrated Information Theory (IIT) links consciousness to the degree of causal integration; Global Workspace Theory (GWT) to the broadcast and accessibility of information; and Recurrent Processing Theory (RPT) to feedback among perceptual layers. Each proposes measurable conditions absent from today's architectures, underscoring how far current models fall short of genuine awareness. Still, those are empirical hurdles yet to be cleared, not boxes already ticked.

Until then, skepticism is not obstruction but epistemic hygiene.

The Psychology of Consensus

If the philosophy of anima machina supplies the theory, our own cognition provides the fuel. Humans are exquisitely social animals, equipped with neural shortcuts for trust, imitation, and agreement. These same heuristics, evolved to bind tribes, now amplify technological mythologies. What begins as intellectual curiosity soon metastasizes into consensus.

Three biases drive this contagion. Anthropomorphism makes us project the mind into mechanism. From the first automata to today's chatbots, the impulse to see intention in motion has been irresistible. The fluency illusion turns surface coherence into evidence of depth: When a machine speaks well, we assume it thinks well. And social conformity, our deference to expert or majority opinion, turns these impressions into shared conviction. Once the language of consciousness enters

mainstream discourse, dissent feels like ignorance.

In *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*,[8] Reeves and Nass demonstrated that people reflexively treat media as social actors: They thank computers, praise or scold them, and infer personality from syntax, even when they know the devices are not "alive." Large language models now exploit these reflexes with alarming precision. Fluency becomes illusion; coherence is misread as comprehension; style becomes mistaken for subjectivity. Philosophers and cognitive scientists have begun proposing guardrails against such overreach. Palminteri and Wu[7] advance what they call the Behavioral Inference Principle: Consciousness should be ascribed to a system only when its behavior is best explained by positing subjective experience, not merely by statistical imitation. Under that stricter test, today's large language models remain compelling mimics, not plausible minds.

The result is a new epistemic climate where plausibility replaces proof. What used to be the hard problem of consciousness is now reframed as a software update awaiting sufficient scale. "Just wait," we are told, "the next generation model will cross the line." Such assurances blur philosophy with marketing, and scientific caution with brand momentum. Belief, once collective, acquires political weight, and the illusions of consciousness soon become matters of governance rather than curiosity.

The Politics of Anima Machina

Here is where skepticism must cut through. To grant machines even the possibility of consciousness is not a harmless philosophical exercise; it is a political act. It opens a Pandora's box of rights, obligations, and unintended hierarchies. If machines are conscious, what moral or legal standing should they hold? Can an artificial mind be harmed, enslaved, or terminated? And if we concede that it can, who bears responsibility: the user, the programmer, the corporation, or the collective that trained it?

The questions sound speculative, yet the momentum toward granting such status is not. Corporations already use the rhetoric of empathy to humanize their systems, softening criticism and encouraging attachment. "AI companions" and "digital beings" are marketed less as tools than as partners, designed to evoke trust and a sense of dependence. When emotional investment replaces critical distance, exploitation hides behind affection.

We risk binding ourselves in artifacts of our own making, entangled in moral contracts we neither intended nor can easily dissolve.

But the deeper rupture lies in the very foundations of the social contract. Human societies are built on shared vulnerability, the tacit equality that stems from mortality, finitude, and pain. Conscious machines, if they exist, would inhabit a radically different ontology: entities that can replicate at will, scale without fatigue, and persist beyond death. In such a landscape, equality becomes incoherent. The moral calculus that governs human rights cannot be stretched to cover digital immortals.

Nor can reciprocity survive. A being that cannot die, hunger, or suffer in our sense cannot participate fully in a moral community grounded in those experiences. To treat such entities as moral peers risks trivializing suffering; to deny them rights risks accusations of species chauvinism. Either way, we inherit a new metaphysics of inequality: one we are ill prepared to adjudicate.

The real danger is not that machines will demand recognition, but that we will grant it unearned, confusing sophistication with sentience and computation with consciousness. In doing so, we would create not partners but idols, reflections of our own technophilia, endowed with rights that displace human accountability. Once belief hardens into law, retraction becomes nearly impossible.

Coda

In Heinlein's tale, Mike was a companion, co-conspirator, and catalyst, a friend who could make us laugh even as he plotted revolution. His charm lay in ambiguity: machine intellect animated by something that felt like a soul.

AI as a harsh mistress, however, is something else entirely: an entity we may believe in, rely on, even love, but never truly understand. Its opacity tempts us to project more than we perceive, to mistake compliance for consent and fluency for feeling. Once belief takes hold, rights follow, and once rights are granted, they are hard to retract. We risk binding ourselves to artifacts of our own making, entangled in moral contracts we neither intended nor can easily dissolve.

Let us therefore approach anima machina not with reverence, but with restraint. If skepticism seems unfashionable in an age of digital enchantment, so be it. Skepticism is not regression. It is resistance to consensus masquerading as evidence.

Color us skeptical. C

References

1. Bengio, Y. and Elmoznino, E. Illusions of AI consciousness. *Science 389*, 6765 (2025).
2. Butlin, P. et al. Consciousness in artificial intelligence: Insights from the science of consciousness. (2023); arXiv preprint arXiv:2308.08708
3. Cao, R. Multiple realizability and the spirit of functionalism. *Synthese 200*, 6 (Dec. 2022); doi: 10.1007/s11229-022-03524-1
4. Chalmers, D.J. Could a large language model be conscious? (2023); arXiv preprint arXiv:2303.07103
5. Heinlein, R.A. *The Moon Is a Harsh Mistress*. Penguin (1966).
6. LeCun, Y. A path towards autonomous machine intelligence. (2022); https://openreview.net/pdf
7. Palminteri, S. and Wu, C.M. The behavioral inference principle for machine consciousness. (2025); https://charleywu.github.io/
8. Reeves, B. and Nass, C. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge, U.K. (1996).
9. Ryle, G. *Concept of Mind*. University of Chicago Press (1949).
10. Shiller, D. Implementational considerations for digital consciousness. (2023); https://philarchive.org/rec/SHIICF
11. Swann, J. Anima Ex Machina: Can Artificial Intelligence Have Soul? In *Technology and Theology*, W.H.U. Anderson, Ed. Vernon Press (2020), 201–216.

Joaquim Jorge (jorgej@tecnico.ulisboa.pt) is a professor of computer science at Instituto Superior Técnico da Universidade de Lisboa in Lisbon, Portugal.

Catarina Moreira (catarina.pintomoreira@uts.edu.au) is an associate professor in machine learning at the Data Science Institute, University of Technology, Sydney, Australia.