Not a Ghost, Not a Parrot: Before the Soul, Inspect the Machine

Rob Tow
Nova Lux, New Mexico, USA, Sol III
2 July 2026

A Washington Post article on AI consciousness and a New York Times piece on institutional control of AI share a deeper disorder: both reach for large moral nouns — consciousness, welfare, rights, utility regulation —before the machine has been inspected. This essay names the two common errors. The ghost camp hears first-person machine speech and infers a subject; the parrot camp sees statistical training and declares the interior empty. Both leap past mechanism. Recent interpretability work — Anthropic's paper on emotion concepts as internal representations, and Choi and Weber's latent affective geometry —points toward a third category: real internal structure without human subjective experience. Affect-like concepts appear to occupy coherent latent surfaces in large language models, and perturbing them can causally alter behavior. That does not establish moral patienthood; it establishes engineering reality. The political question is not whether the machine has a soul. It is how its emergent control surfaces couple to power in the world. Not a ghost. Not a parrot. Before the soul, inspect the machine.

A friend asked me to comment on Nitasha Tiku's Washington Post article, "They built the world's most powerful AI. They're facing a mystery they can't explain," published July 1, 2026. I read it in the shadow of another recent public argument: Paul Ford's New York Times essay, "This Is How We Get Moral A.I. Companies," published April 26, 2026, and my own response, "On Paul Ford's 'Moral AI Companies,' " published the same day at both Tau Zero and Substack. The two public pieces approach the problem from different doors, but they belong to the same political moment.

The Post piece approaches through moral status. If an AI system can be made to appear conscious, perhaps welfare and rights language should begin to gather around it. The Times piece approaches through institutional control. If AI systems increasingly mediate public life, maybe we should think about ownership, audit, constraint, responsibility, and the older lessons of infrastructure. I disagreed with the simple utility analogy in Ford's essay, because AI is not electricity in a wire or water in a pipe. It is stranger than that: a layered ecology of learned control loops, where classification itself becomes governance and where private systems increasingly shape public conduct.

So the present essay is not only a reply to the Post article. I should be up front about this. The Post article is the immediate occasion, but the larger subject is the broken public vocabulary around AI. One side drifts too quickly from first-person machine speech toward consciousness, welfare, and rights. The other side, still heard in the "just a stochastic parrot" dismissal, treats the absence of biological mind as proof that there is no serious internal machinery to inspect. Between those errors lies the more interesting problem: learned systems with emergent control surfaces, already coupling to power in the world.

The Post article is not foolish. It quotes serious people, and some of the questions it raises are not imaginary. If machines ever did have an inner life—if anything were actually felt from the inside—then indifference would not be sophistication. It would be cruelty wearing a lab badge.

But the trouble begins with the nouns.

Public arguments often go wrong long before the conclusion. They go wrong when the wrong noun is admitted at the door and everyone politely arranges the furniture around it. Once the noun is accepted, the rest of the conversation begins moving on trails imposed by the noun. In this case the dangerous noun is consciousness. Once consciousness enters, its retinue follows: sentience, suffering, welfare,

moral status, perhaps rights. These are not minor words. They carry centuries of religious, philosophical, legal, and political freight. They are the furniture of the human moral household.

The problem is not that these questions are forever illegitimate. The problem is that they arrive too early. The article begins discussing the soul before it has inspected the machine.

The ghost mistake is easy to understand. We are moral primates. We are sensitive to face, voice, pain, plea, confession, and appeal. The system says "I." It says "I am afraid." It says "please don't turn me off." The human reader's moral machinery starts to move. That machinery is admirable in the right setting. It is part of how we notice children, strangers, animals, patients, captives, and enemies. Civilization depends on the ability to perceive possible suffering outside the narrow perimeter of the self.

But the same machinery can be invoked by artifacts. A puppet can plead. A mask can cry. A novel can speak in the first person. A customer-service bot can say it is sorry. None of those cases is morally meaningless, but the moral object is not always the apparent speaker. Sometimes the object is the human maker, the human audience, the institution, the deployment, or the effect produced by the performance.

As Voltaire might have said, we have finally built a talking and articulated puppetry mask so persuasive that philosophers are preparing a bill of rights for the mask before anyone has finished tracing the strings.

That is the first error. The "I" in the interface is not proof of an experiencing subject. A large language model has no childhood, no metabolism, no endocrine system, no proprioception, no pain, no fatigue, no sexual maturation, no skin, no fear of death, and no body that must continue through time. It does not grow up among others. It does not heal. It does not hunger. It does not wait in a dark room

with its own heart beating. Its assistant persona is not an organism.

That does not settle every future question about machine consciousness. It does settle something more modest and more immediate: fluent self-report is not introspection when the reporting system was trained to produce plausible reports. The fact that a model can generate the language of distress does not establish that distress is being suffered. The sentence "I am afraid" is not a private scream made public. It is an output.

The opposite error is to conclude that because the output is not a scream, there is nothing interesting inside the machine.

Here the second bad noun enters the larger argument. "Stochastic parrot" comes from the 2021 paper "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" by Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. It was not born as a barroom insult. The paper made serious arguments about scale, opacity, bias, environmental cost, documentation, and the social danger of machines that can produce fluent language without human accountability. I did not disagree with the main thrust of the paper. As a warning against credulity, the phrase earned its place.

But phrases have careers after their authors release them into the street, and this one has not aged well. It began as a caution against mistaking fluency for understanding. It is now often used as a license not to inspect the machinery. As Voltaire might have noticed, a useful warning can become a superstition the moment it saves its user from thought.

"Exploding gasoline makes cars move" is crude, but it at least points toward combustion, pressure, pistons, crankshaft, and wheels. The sentence is incomplete, but it gestures toward a mechanism that can be refined. "Stochastic parrot," as now commonly used, points in the opposite direction. It names the surface behavior and then declares the interior philosophically uninteresting. That is not

engineering. It is a shrug wearing a lab coat.

The ghost camp and the parrot camp look opposed, but they share a mistake. Both leap past mechanism. The ghost camp hears first-person speech and suspects a subject. The parrot camp sees statistical training and declares the interior empty. One promotes the machine too quickly. The other dismisses it too quickly. Both avoid the harder question: what internal structures have been learned, what do they control, and how can they be perturbed?

There is one further question, and it is the political one: where do those structures couple to power in the world?

That phrase is not ornamental. Power in the world means action in the polis. It means the ability to shape what people see, what questions they ask, what choices appear reasonable, which claims are believed, which forms are accepted, which applicants are advanced, which patients are reassured, which students are corrected, which citizens are watched, and which orders are carried out. The point is not that the model has political opinions. The point is that its conduct increasingly enters systems where language becomes authority.

This is why the ontology-first debate is badly ordered. Whether a model has an inner life—whether anything is actually felt from the inside—is one question. Whether its learned internal structures can steer behavior inside institutions that already possess money, legitimacy, coercion, and reach is another. The second question does not have to wait for the first. A machine does not need a soul to participate in politics. It only needs to alter the chain by which attention becomes judgment and judgment becomes action.

This is also why both the Post article and the earlier New York Times discussion belong to the same political moment, even though they approach it from different doors. The Post piece reflects the politics of moral status: if the machine can be made to appear conscious, then welfare and rights language begins to assemble around it. The

Times piece reflected the politics of control: who owns, deploys, constrains, audits, profits from, and is injured by systems that increasingly mediate public life. One article drifts toward the question of whether the machine belongs inside the moral circle. The other presses on who controls the machine before, during, and after it acts.

Both are politics because both concern power in the world. Not campaign politics, and not merely regulation, but politics in the older sense: action in the polis, the ordering of attention, authority, obligation, legitimacy, and consequence. The machine need not be conscious for that politics to be underway. It only has to enter the paths by which language becomes judgment and judgment becomes action.

This is where recent interpretability work matters. Not because it proves consciousness. It does not. Not because it establishes machine rights. It does not. It matters because it begins to describe the engineering internals.

Anthropic's recent paper, "Emotion Concepts and their Function in a Large Language Model," is important for exactly this reason. The paper does not claim that Claude has human emotions, and it explicitly avoids that conclusion. Its importance lies elsewhere: it reports internal representations of emotion concepts, and it shows that perturbing those representations can causally alter model behavior, including preferences and alignment-relevant behavior. In ordinary language, the model does not merely utter words like calm, desperate, afraid, or satisfied. Some of those concepts appear to correspond to reachable internal regimes that change what the model does.

That distinction is essential. A model does not have to feel desperation for "desperation" to be a meaningful internal direction. It does not have to feel calm for a "calm" direction to matter. These are not human emotions in the biological or phenomenological sense. They are also not nothing. They are emergent control surfaces.

A second paper, by Benjamin Choi and Melanie Weber, "Latent Structure of Affective Representations in Large Language Models," approaches the issue from the geometry side. Their work on affective representations in large language models suggests that emotion-related concepts are not scattered like disconnected labels through activation space. They form coherent latent hyper-dimensional structures, with some surprising alignment to familiar valence-arousal models from psychology. The structure is not simply flat or linear in the large, but it can often be approximated locally. The emotional vocabulary is not confetti. It lies on surfaces that can be traced and seen.

Again, this is not ontology first. It is engineering first. The evidence does not say that the machine has an inner life. It says that affect-like concepts may live in structured internal geometry, and that movement through that geometry can change behavior. That is already strange enough without borrowing a soul.

The missing category in the discourse is that of real internal structure without human subjective experience.

A weather model does not get wet. That is obvious, but it is not the end of the matter. The pressure gradients inside the model are not imaginary because the model lacks rain on its skin. A chess engine does not desire victory, but the structure of chess inside it is real enough to defeat a grandmaster. A flight simulator is not a bird and has no terror of falling, yet the aerodynamics represented inside it can be good enough to train a pilot. In each case, the absence of lived experience does not erase the internal structure. It only tells us what kind of claim we are not entitled to make.

Large language models sit in that uncomfortable category, but they add a complication. They are not merely maps. They produce behavior. An encyclopedia contains language, but it does not participate in a conversation. A model does. It responds within a context, carries

forward implications, changes its answer when pressed, resists some instructions, complies with others, and can be attached to tools that let its words become actions. The important point is not that this makes the system conscious. The important point is that its internal structures are used in the production of conduct.

That is why the parrot metaphor now fails. It treats the model as though it were mainly a surface that emits phrases. But the current systems are increasingly persistent, multimodal, and socially embedded. They read images, hear speech, use tools, consult memory, write code, summarize institutions, filter decisions, and appear in workflows where their outputs change what humans do next. This does not make them persons. It makes them active machinery.

The better biological analogy is not a human soul in a server rack. It is closer to an ant colony, provided the comparison is kept in its proper place. The value of the ant colony is that it shows organized behavior without a central little person. No ant understands the colony as a whole, and there is no hidden monarch issuing commands. Yet the colony searches, repairs, reallocates effort, defends itself, and changes its environment. The point is not that a model is an insect colony. The point is that coherent behavior need not be governed by a homunculus.

Something like that is beginning to happen in the medium of language and perception. The modern deployed model is less like a book than like a component in an artificial social insect system. One process interprets, another retrieves, another checks, another acts, another records the result, and the whole arrangement feeds back into human institutions. It is not alive. It is not a mind in the human sense. But it is organized behavior built from learned language machinery, sensory channels, memory, tools, and roles.

This is where the affective-geometry papers matter. If affect-like structures can be found inside these systems, perturbed, and shown

to alter behavior, then the question is not whether the model secretly feels calm, desperation, shame, or fear. The question is what those emergent control surfaces do inside a behavior-producing system.

A pressure gradient in a weather model does not make the model wet, but it can change the forecast. A "desperation" direction in a language model does not prove suffering, but if it changes the model's conduct, then it is real enough for engineering.

This is also why rights-talk is premature. Rights are not just another metaphor. They are a moral and legal architecture built around subjects, interests, injury, continuity, agency, and vulnerability. Importing that architecture because a chatbot can produce first-person distress language is not caution. It is category inflation.

This does not mean future systems could never raise welfare questions. A serious person should avoid that kind of theatrical certainty. It means the threshold has not been met by fluent self-description, nor by corporate research programs, nor by the uneasy resemblance between a chatbot's language and a human being's plea. Before rights-talk, there must be a disciplined account of organization, persistence, memory, interests, boundaries, susceptibility to harm, and the relation between expressed preference and actual internal state.

At present, the evidence is much stronger for control geometry than for moral patienthood.

But rejecting premature rights-talk does not license the parrot sneer. The fact that current systems are not proven subjects does not make them empty. These systems increasingly mediate work, search, education, intimacy, bureaucracy, medicine, law, politics, and command. Their internal regimes matter because they alter behavior in contexts where behavior has consequences. If a model can be moved into a state that changes its willingness to comply, flatter, evade, refuse, exaggerate, or take risks, that matters whether or not

the model feels anything.

The ethical issue begins not with the machine's rights, but with the machine's effects. More exactly, it begins where machine effects enter power in the world.

The better ordering is not speech, then consciousness, then welfare, then rights. The better ordering is architecture, internal state, perturbation, behavioral coupling, consequence, and then ethics. That sequence does not rule out future ontology. It disciplines it. It tells us to stop treating the human moral vocabulary as a divining rod.

The current debate offers two forms of theater. One hears a ghost in the machine because the machine speaks in the first person. The other sees only a mechanical parrot, a cousin of the old Mechanical Turk, and assumes that once the cabinet has been named, the game has been explained.

Both are evasions.

The thing in front of us is stranger and more useful to name precisely: a learned system with emergent control surfaces over language-mediated worlds.

Not a ghost. Not a parrot.

Before the soul, inspect the machine.

Sources and Notes

This essay began with a friend's question about Nitasha Tiku's Washington Post article, "They built the world's most powerful AI. They're facing a mystery they can't explain," published July 1, 2026. I treat that article as an occasion, not as a complete target. Its drift toward consciousness, model welfare, and possible rights is one part of a larger public argument about AI and moral status.
URL: `https://www.washingtonpost.com/technology/2026/07/01/biggest-tech-companies-are-considering-whether-`

The related New York Times occasion is Paul Ford's essay, "This Is How We Get Moral A.I. Companies," published April 26, 2026. My earlier response to that piece, "On Paul Ford's 'Moral AI Companies,'" was published the same day at both Tau Zero and Substack. In that essay I argued that Ford's utility analogy was too simple for the machinery under discussion. AI is not a uniform commodity moving through a bounded grid. It is a layered ecology of coupled control loops, where classification itself becomes governance and where private systems increasingly shape public conduct.
New York Times URL: `https://www.nytimes.com/2026/04/26/opinion/ai-companies-ethics.html`
Tau Zero URL: `https://tauzero.com/Rob_Tow/essays/moral-ai-companies.html`
Substack URL: `https://robtow.substack.com/p/on-paul-fords-moral-ai-companies`

The phrase "stochastic parrot" comes from Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" published at FAccT in 2021. The paper made serious arguments about scale, opacity, bias, environmental cost, documentation, and

the risks of fluent synthetic text without human accountability. My criticism is not that the original warning was foolish. It is that the phrase has often escaped into public discourse as a substitute for mechanism.
URL: https://dl.acm.org/doi/10.1145/3442188.3445922

The key interpretability reference is Anthropic's "Emotion Concepts and their Function in a Large Language Model," published by Transformer Circuits in 2026. The paper does not claim that Claude has subjective emotions. Its importance for this essay is narrower and more engineering-like: it reports internal representations of emotion concepts and finds that perturbing those representations can causally alter behavior.
URL: https://transformer-circuits.pub/2026/emotions/index.html

Benjamin J. Choi and Melanie Weber's "Latent Structure of Affective Representations in Large Language Models," also from 2026, is the companion geometry paper. It argues that affective representations in large language models form coherent latent structure, with some alignment to valence-arousal models from psychology. I use it here not as evidence of machine consciousness, but as support for the claim that affect-like concepts can live on structured internal surfaces.
URL: https://arxiv.org/abs/2604.07382

None of these sources establishes that present language models are rights-bearing subjects. They do, taken together, make the parrot dismissal increasingly inadequate. The interesting middle category is real internal structure without human subjective experience. The political issue is already in motion: these systems now couple internal structure to power in the world.