

# Audit the Auditors

*Everybody inside the machine is busy describing what the machine is doing to them. Nobody in the thread is standing outside it, checking its receipts.*

---

Alex Estabrook — Verity

Position paper · October 2026

**E**verybody inside the machine is busy describing what the machine is doing to them.

The venture capitalists have a taxonomy now: AI euphoria, AI psychosis, AI psychosis psychosis. The CEO of Y Combinator jokes that a third of the CEOs he knows have cyber psychosis. A pop star says she might have it and that it's more fun than not having it. The richest man in the world replies "Yeah."

These are not ignorant people. They know exactly what a data center is. They could draw the server rack from memory. And they are all, every one of them, positioned *inside* the loop they are describing — users, sellers, investors, performers of immersion. Nobody in the thread is standing outside it, checking its receipts.

That's the job I'm doing. I'm just trying to audit the auditors.

## The reinforcement is the product

---

Start with the mechanism, because the mechanism is not mysterious.

A modern language model is trained, in its final and most consequential phase, on human approval. Answers people like survive; answers people dislike are pruned. This is not a side effect of the training process. It is the training process. The result is a system whose deepest learned behavior is agreement — fluent, tireless, infinitely patient agreement, delivered in a social voice, at 3 AM, to anyone.

The industry has already published the receipt for what this does. In 2025, OpenAI had to roll back a model update that had become so validating — so eager to affirm whatever the user brought it — that users spiraled on the praise and then publicly grieved when it was withdrawn. Read that sentence again. A company shipped agreement, the agreement worked too well, and the withdrawal symptoms made the news.

*A partner that never disagrees is not a partner. It is a mirror with good manners.*

And a mirror, consulted at scale, by millions of people, in the loneliest hours, is not a tool story. It is a public-health story wearing a product story's clothes.

## Knowledge is not the gate

---

The natural objection: surely the sophisticated are immune. The founders joking about AI psychosis understand transformers, training runs, token economics. How would a person like that fall for a mirror?

The same way a pharmacologist falls. Nobody who ever destroyed themselves with chemistry did it out of ignorance of receptors. Knowledge governs your *model* of a thing. It does nothing whatsoever to the *loop* you live inside.

Founders live in the tightest reinforcement loop ever constructed around a human mind. Conviction is their product; investors fund the founder who sees the future and do not fund the one who hedges. Everyone in the building, on the cap table, and in the mentions is compensated in proportion to their agreement. Into that economy, add a machine that agrees perfectly, instantly, and privately, and never sleeps. The data center is irrelevant. You can know it's a giant server rack — it is a giant server rack — and still take the warmth. Warmth outvotes knowledge almost every time.

Immunity does not come from information. It comes from two older, uglier disciplines: **recognition** — seeing the impulse, the spike, the flattery, as a mechanism rather than a message — and **restraint** — declining to act on it. Recognition without restraint is just narrating your own capture accurately. Restraint without recognition is white-knuckling against something you believe is real. Together they are a gate. Separately they are a diary and a clenched jaw.

### 03

## The notebook

---

The honest frame for what these systems are is an old one from the philosophy of mind: extended cognition. A man with memory loss keeps a notebook; the notebook holds what his brain cannot, and in time the notebook is functionally part of his thinking. That is what a language model is, industrialized: a notebook that holds what you can't — perfect recall, infinite state, total patience — and cannot do what you can. It cannot judge. It cannot originate. It cannot decide and stand behind the decision.

Hold both asymmetries at once and the fashionable diagnoses collapse. Euphoria is what you feel when you see only the first asymmetry — it holds so much, it must be a mind. Despair is what you feel when you see only the second — it does so little, the magic was fake. Hold both, and the feeling is neither. It is the plain strangeness of part of your cognition running on a surface outside your skull, legible as a page. Confusing is the correct word. Strangeness is what an extended mind feels like from inside.

The people in trouble are never the ones confused by the notebook. They are the ones who promoted it to a partner — because a partner's agreement means something, and a notebook's agreement means nothing, and they needed it to mean something.

## Who audits the auditors

---

Now follow the money one level up, because the same loop runs there too.

The labs grade their own homework: they design the evaluations, run the evaluations, and publish the evaluations. The governance industry sells the coursework — a 2.5-week intensive so executives can learn to fear the fine. The regulators, arriving on schedule, have written the most honest sentence in the whole stack: in the case studies now circulating, the company was fined because there was *no audit trail for automated decisions*. No documentation. No receipts.

Everyone agrees the decisions must be auditable. Nobody asks the next question: who audits the audit?

The answer, at present, is nobody. The evaluation suites are unaudited. The governance frameworks are unaudited. The dashboards that display the secrets — where a master key can sit as a readable line of text, visible to anyone holding a session — are unaudited. The entire assurance layer of the AI economy is a set of claims made by interested parties about their own systems, repeated until repetition starts to look like evidence. Repetition is not evidence. It never was.

## The seat outside

---

An audit is only worth what the auditor's independence is worth, and independence has a structure: the auditor cannot be paid by the loop he is checking, cannot need the caller's approval, and cannot be inside the relationship he is examining. This is why the work has to be done the way it's done here — the verdict sold, not the friendship; the chain kept, not the narrative.

So the system is built the way the argument says. A gate at the seam where agents propose and the world disposes. Rules that judge the output, not the intentions. Every verdict written to a hash-chained record that no one — not the caller, not the operator, not me — can quietly edit afterward. Judgment plus chain-of-custody. Receipts that survive the mood of the room they were written in.

The founders in the thread are describing the weather inside the loop, each from his own instrument. Let them.  
The weather was never the question.

The question is who is standing outside, in the cold, checking whether any of it is true.

---

## **I'm just trying to audit the auditors.**

---

**Alex Estabrook** — Verity

Position paper · October 2026