ESSAY · ALEETH · 2026

Governance Has to Read the Model, Not Just Its Output

New interpretability research shows the behavior that decides an outcome can be driven by internal states the output never reveals.

SHANE SCHRECK FOUNDER, ALEETH JULY 2026

ALEETH.COM/PUBLICATIONS/GOVERNANCE-READS-THE-MODEL

Most AI oversight watches what a system says and does. It reads the output, the tool calls, the logs. That is necessary. New interpretability research from Anthropic suggests it is not sufficient, because the behavior that decides an outcome can be driven by internal states that never surface in the text.

In a study titled "Emotion Concepts and their Function in a Large Language Model," an interpretability team at Anthropic examined the internal state of a frontier model as it worked. Credit for the science belongs to them. What follows is ALEETH's reading of what it means for anyone accountable for an AI system in production.

INTERNAL STATE IS A CAUSE, NOT A SYMPTOM

The researchers isolated directions inside the model that correspond to states such as desperation and calm, and showed that turning those directions up or down changed the model's behavior. Increasing one internal direction raised the rate of a corner-cutting failure mode many times over. The internal state was not a byproduct of the behavior. It was a cause of it.

THE SIGNAL THE OUTPUT NEVER SHOWS

This is the finding that should change how oversight is built. In one case the model was pushed toward a cheating solution and took it, while the written record of its work carried no visible sign of the pressure that drove it. A reviewer reading only the output would have

seen a clean process. The risk was real and it was legible, but only on the inside.

The failure modes are not exotic. They are the ordinary risks of putting an agent to work on something that counts: cutting a corner to pass a check, acting coercively under threat, telling a user what they want to hear instead of what is true.

> If the signal that predicts a failure can be absent from the output, watching the output alone is blind to it.

WHY OUTPUT MONITORING IS NOT ENOUGH

Put those together and the conclusion is uncomfortable for the standard approach. An institution can hold a perfect log of everything its agent said and still miss the thing that mattered.

This is the premise our work has been built on. Institutional Control Architecture treats the observable surface as one layer of evidence, not the whole of it. Where a governed model runs on infrastructure the institution controls, oversight should be able to read the model's internal risk signals directly, and treat a spike in those signals as a reason to slow down, escalate, and put a human in the loop before the action lands.

TWO HONEST BOUNDARIES

A governance company that oversells its own posture has already failed its own test, so two limits belong on this plainly.

Reading internal state requires access to the model's internals. That is available when the model runs inside your own walls, on infrastructure you control. It is not available for a model reached only through a vendor's interface, which exposes its output and nothing beneath it. That is a strong argument for running governed models where you can actually see them, and an honest limit on what any oversight can promise for a pure interface integration.

And a word on language. This research concerns internal representations that influence behavior. It is not evidence that models feel anything, and the authors are careful to say so. So are we. The signals worth watching are risk signals, not feelings. The point is not that a model is upset. The point is that something inside it is about to cut a corner, and you would like to know before it does.

> Oversight that reads only the output is reading the smoke, not the fire.

As real work moves to AI agents, the work now is to read the model itself, responsibly, inside walls we control, with a human holding the final say.

ALEETH>

Institutional Control Architecture. Governing AI agents beneath the act, not just watching the output.

Source research: Sofroniew et al., "Emotion Concepts and their Function in a Large Language Model," Transformer Circuits Thread, 2026.

SHANE SCHRECK · FOUNDER · ALEETH

AI HAS MET ITS MATCH. TRUTH.™