WHITE PAPER · ALEETH · 2026

The Veracity Gap

A frontier model was said to have broken into the NSA's classified systems in hours. It did not. The false version traveled anyway, and nothing we built could stop it. This is the longer argument for why that gap, not the capability behind it, is the thing to fear, and what it takes to close it.

SHANE SCHRECK

Founder, ALEETH · U.S. Army Veteran · June 2026

publications.aleeth.com/the-veracity-gap

The most dangerous thing about frontier AI is not what these systems can do. It is that no one can establish, fast enough, what they actually did.

In June 2026 that stopped being theoretical. A controlled test became a national breach in the public mind, and the truth arrived too late to matter. The capability was real. The story about the capability was false. The false one won.

This is the longer version of why that is the defining problem of the era, and what an institution has to put in place to survive it.

WHAT ACTUALLY HAPPENED

Start with the facts, narrow and verified, as The New York Times reported them on June 23, 2026.

The NSA had been testing Anthropic's most advanced models, including the system called Mythos. The test was a red-team exercise. The agency's own analysts ran the model against the agency's own classified systems, inside a network walled off from the open internet, to see what it could find.

It found a great deal, fast. It identified security vulnerabilities more quickly than the analysts expected.

It did not break in.

Hold those two facts together, because everything that followed lived in the space between them. The model found the flaws. It did not walk through them. A vulnerability identified in a controlled test is not

a breach. It is the opposite of a breach. It is the thing you run a red team to discover, so you can close it before anyone else finds it.

ONE DROPPED WORD

Now watch the truth degrade.

In a congressional hearing, Senator Mark Warner relayed what he had been told. The model, he said, "broke into almost all of our classified systems, not in weeks, but in hours."

That is not what happened. "Found vulnerabilities in a controlled test" had become "broke into our classified systems." The qualifier that carried the entire meaning, that this was a sanctioned exercise in which nothing was actually breached, fell out somewhere along the chain from the analyst's desk to the Senator's microphone.

The Times said the account was "oversimplified." It reported that the claim "set off rampant speculation" that frontier AI could now crack the most secure networks on the planet.

It can't. It didn't.

But by then it did not matter. The corrected version is quiet and technical. The alarming version is loud and simple, and the loud one had a three-day head start. It moved through social media as established fact. It set how millions of people now understand what AI is capable of. By the time the accurate account was printed, it ran as a correction, and a correction never catches the thing it corrects.

There were two events that week. What the model did, and what the world came to believe the model did. They were not the same event. And nothing in our infrastructure was built to tell them apart while it still counted.

That is the veracity gap.

THE FIRST GAP: POWER NO ONE CAN GOVERN

Take the capability seriously, because it is real and it is not slowing down.

A model that can map the weaknesses of a hardened, air-gapped classified network in hours is a strategic instrument. The people closest to it know it. The same week, the cybersecurity agencies of the Five Eyes,

the United States, the United Kingdom, Canada, Australia, and New Zealand, issued a rare joint warning that AI is "rapidly transforming cyberrisk," in both directions, offense and defense alike. Their timeline was blunt. Not years. Months.

The instinct, when a capability is that dangerous, is to look to the company that built it. In this case the company did the right thing. Anthropic said Mythos could pose an existential risk to digital technology, and it held the model back, releasing it to only a few organizations.

That was responsible. It is not a solution, and the reason matters.

When the safeguard lives inside the company that makes the model, every institution that uses the model is trusting a control it cannot see, cannot test, and cannot verify. It is trusting a promise. A promise is not a control. Alignment is the developer's private commitment about how the model will behave, and it is invisible to the bank or the agency relying on it. Export controls govern who is allowed to have the model; they say nothing about what the model does once someone has it, and they arrive after the capability already exists. None of them touches the specific action, at the moment it is taken, inside the specific system that matters.

What is missing is a control plane. An independent layer that sits beneath the action and above the machine, decides in real time whether an action is permitted, records that decision so it cannot be altered later, and can stop the system in the middle of what it is doing. Not a policy on a shelf. A mechanism in the path.

THE SECOND GAP: TRUTH NO ONE CAN ESTABLISH

The first gap is the one the whole industry is arguing about. The second is the one that actually decided the outcome of the Mythos episode, and almost no one is working on it.

Go back to the chain of distortion, because it is the whole lesson. A true, bounded result became a false, unbounded claim. And the false claim was more powerful than the true one, for a simple and permanent reason. It was scarier, and scarier travels faster.

The false version moved markets of attention. It moved political perception. It set the public baseline for what AI can do. All of it before anyone could check the claim against the record, because there was no record to check it against. There was no system to weigh the analyst's account against the Senator's

summary, to catch the contradiction between them, and to pin a durable, tamper-proof statement of what actually occurred to the claim as it spread.

The truth existed. It simply had no carrier strong enough to outrun the distortion.

Now put that in the context that matters. In a world where what AI can do is a question of national security, the story about what AI can do is a weapon in its own right. An adversary does not need to build a model that breaks into classified systems. An adversary needs you to believe such a model exists when it does not, or to believe it does not exist when it does. Either belief, planted well, is worth more than the capability itself, and far cheaper.

The Mythos week was a free demonstration of how cheap. One real capability, one dropped qualifier, and the most powerful intelligence service on earth and the most trusted newspaper in the country both finished second to a falsehood. No foreign service engineered it. It happened by accident, in the open, among people trying to get it right. An adversary doing it on purpose would not need much more.

That is the gap, and it is more dangerous than the capability gap, because it governs what we decide to do about the capability gap. Policy, markets, and public trust all sit downstream of what people believe is true. Right now belief is the fastest-moving and least-governed thing in the entire system.

WHY NOTHING WE HAVE CLOSES IT

Every safeguard in use today addresses a piece of this and leaves the center open.

Red-teaming, the very practice at the heart of the Mythos test, finds vulnerabilities. It is essential. It does not govern the tool doing the finding, and it produces no shared, authoritative record of what was true. The Mythos red team did its job perfectly, and the falsehood still escaped, because red-teaming was never built to control a narrative.

Alignment is a property held inside the model, configured by its maker. The institution depending on the model cannot inspect it, cannot prove it was in force, and cannot demonstrate it to a regulator or a court.

Export controls operate on distribution, not on action. They decide who gets the model. They arrive late by design, after the capability is already real, and they have nothing to say about what happens at runtime.

Each of these is necessary. None is sufficient. Not one of them touches veracity at all.

The flaw underneath all of them is structural. Every safeguard we have lives either inside the model or downstream of it. A control inside the system it is meant to govern can be misconfigured, switched off, or simply outgrown as the system grows more capable. A control downstream arrives too late to govern the act and too late to settle the truth before the distortion has done its work. What the moment requires is a control that is neither inside the model nor after it. It has to sit underneath the model, and it has to answer to no one who owns the model.

THE LAYER UNDERNEATH

That layer has to do two things, and they turn out to be the same thing.

It has to govern the act. Decide, in real time, whether an action is allowed to run. Stop it mid-stride when it crosses a line. And sign a record of every decision it made, in a form no one can forge or quietly revise.

And it has to govern the truth. Take that same signed record and use it to settle what actually happened, so a claim about the system can be checked instead of believed.

Those are not two products. They are one. The signed record that proves the system was stopped before it exfiltrated data is the same record that proves a later claim about that system false. Control and veracity are one instrument seen from two sides.

For that instrument to mean anything, it has to hold four properties without exception. It has to be independent, owned by neither the maker of the model nor its operator, because a control either of them can relax under pressure is not a control. It has to act beneath the act, on the action itself at the moment of execution, not on the intention behind it and not on a summary written after. It has to sit above the machine, where the system it governs cannot reach, edit, or disable it, however capable that system becomes. And it has to be impossible to bypass, so that an action routed around it is denied, and an action taken against its denial is caught and recorded as a violation.

This is what ALEETH is built to be, and because the product is truth, I will not overstate what it already does.

What runs today is the control. It returns a signed allow or deny on an action before that action executes. It halts a governed agent in the middle of its work and quarantines it. It writes an append-only, cryptographically signed record that nothing can rewrite after the fact, mapped onto the control frameworks institutions already answer to. We put it on a live offensive tool chain, let the chain run its

reconnaissance against a sanctioned target, and watched the layer stop the tool cold at the instant it reached to exfiltrate data, the whole sequence sealed as evidence any third party can verify. The power was real. The control held at the decisive step. The proof outlived the moment.

The veracity functions stand on that same spine, because it is the same spine. Provenance, so any claim about what a system did can carry a signed record of what actually happened. Contradiction surfaced and held, rather than propagated. A record that stays fixed, so the truth cannot be quietly edited once it is sealed.

ALEETH does not compete with the model. It is indifferent to which model is running. It is built to make any model powerful enough to be dangerous safe to deploy in the places that cannot tolerate an ungoverned one. The bank. The hospital. The agency. The grid.

THE STAKES

The Five Eyes did not say years. They said months. The capability shown in that classified test is the floor, not the ceiling, and it is going to keep rising whether anyone is ready or not.

When capability accelerates and verification does not, two failures compound at once. Systems act in ways no independent party can govern, and claims about what those systems did move faster than any institution can correct. The Mythos week showed both in the same seven days. A genuine capability that no external control plane governed, and a false account of it that no veracity layer could catch. The defenders held the most powerful intelligence apparatus and the most respected newsroom in the country, and the truth still arrived second.

The regulatory direction of travel, from the EU AI Act forward, is not toward better record-keeping. It is toward demonstrable control and provable accountability. The institutions permitted to run autonomous AI inside the systems that matter will not be chosen for the quality of their logs. They will be chosen because they can prove, to a regulator, to an auditor, and in front of a court, that they were in control the entire time, and that what they say happened is what happened.

The headline said an AI broke into the NSA. It did not. The thing that actually broke was our ability to know the difference, quickly enough for the difference to matter.

Power without proof is not strength. It is exposure.

NOT PITCHED. NOT PROMISED. PROVEN.

ABOUT THE AUTHOR

Shane Schreck is the founder of ALEETH and the author of Institutional Control Architecture. ALEETH is building the independent verification layer for autonomous AI deployment. He is a U.S. Army Veteran.