GLOBAL RACE CONDITION
INTELLIGENCE BRIEFING

FEATURED ANALYSIS

Wednesday, April 29, 2026

Signal over noise, for people others depend on.

Substack PDF EPUB

America's AI Incoherence, Part 2: The Measurement Problem

Why You Can't Audit What You Can't See, and Why That's About to Drive American CISOs Toward Chinese Models

On Saturday, I argued that American AI policy is pushing CISOs toward Chinese models because the capability gap leaves them no choice. Defenders need offensive-grade tooling to defend, and US frontier labs have restricted access through a combination of safety friction, classification, and self-imposed restriction.

That was the *capability* argument, and it stands.

This is the *measurability* argument. Unfortunately, it points in the same direction, and it's the part that should worry you more.

If the capability gap is about what the model can *do*, the measurement gap is about what the regulator can *prove*. Right now, the math for compliance only works on open weights. And the primary outlet for capable open-weight models is being shipped from Beijing.

The Subspace Problem: Why AI Models Are Fundamentally Insecurable

A new arXiv preprint dropped in November, “Quantifying the Risk of Transferred Black Box Attacks.” The authors set out to build a framework for adversarial robustness testing on neural networks. Their conclusion is what you'd expect from anyone who's spent five minutes thinking about high-dimensional input spaces.

You cannot compute the attack surface of an AI model.

Not “it's hard.” Not “we need better tools.” You cannot do it. The math doesn't work, and the authors say so directly.

The reason is simpler than any countermeasure, real or hyped:

AI models understand math, not human language. When you type in a sentence, those words are encoded into math. Any attack that results in that same math will work.

Language is imprecise, sometimes frustratingly so. Take these two sentences:

Time flies like an arrow. Fruit flies like a banana.

You understood both. A model processes them as vectors in high-dimensional space, where “flies” maps to completely different regions depending on context. Another way of saying the second sentence could be “Members of the species Drosophila Melanogaster enjoy eating bananas.” Same meaning, entirely different math.

There are a nearly infinite number of ways to generate sentences or even add random characters in ways that will be encoded to match the math your model uses. This is called the "Subspace Problem," and is why AI models are fundamentally insecurable against prompt injection: you cannot block a mathematical synonym you cannot predict. If you want to go down this rabbit hole, HiddenLayer has a good writeup here.

The industry's answer is, of course, more AI. I've even said it myself: "Takes a good guy with AI to fight a bad guy with AI." But that isn't a universal cure-all. When you use an AI model in-line to evaluate prompts and responses before passing to a secondary model, that filter is just more AI vulnerable to the same subspace problem. You haven't solved the problem. You've added a layer vulnerable to it.

It's like blocking the ocean with cinderblocks that melt in salt water. "Here - just add more cinderblocks."

The result of all this is simple: comprehensive adversarial risk mapping is computationally infeasible. You cannot enumerate the attack surface. You cannot prove a model safe.

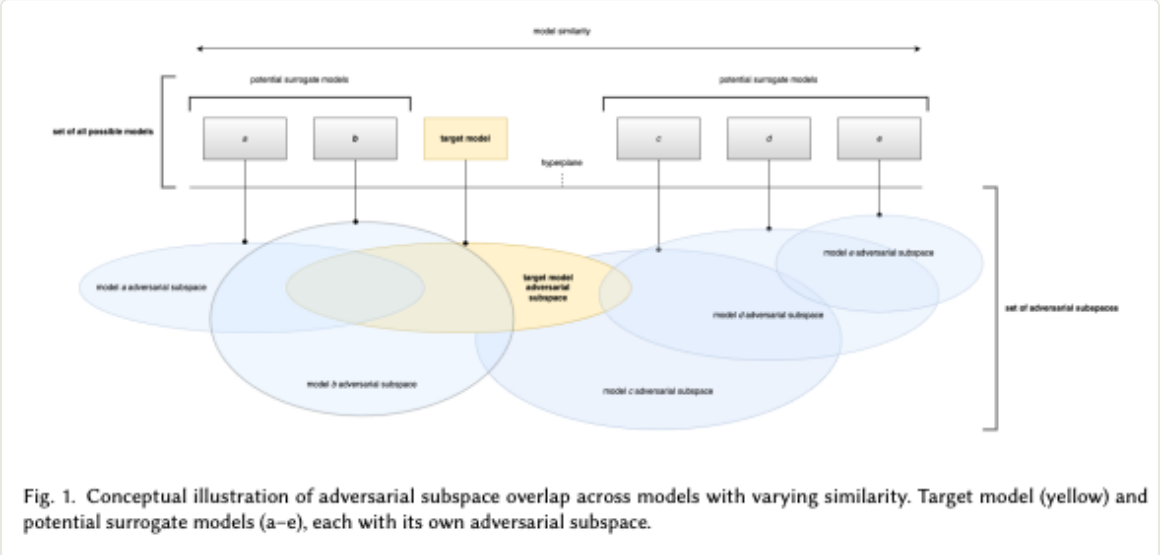

Fig. 1. Conceptual illustration of adversarial subspace overlap across models with varying similarity. Target model (yellow) and potential surrogate models (a–e), each with its own adversarial subspace.

Figure 1: Adversarial subspace overlap across models. The overlapping regions are why attacks developed against one model transfer to others. Source: Cox & Bunzel, arXiv:2511.05102

Transfer Attacks: If I Can Use Your Model, I Can Steal Your Model

There's a property of machine learning that makes closed weights less protective than they appear, and I want to lay it out to help with what follows.

If I can use your model, I can steal your model. If I can send you a bunch of inputs and capture the outputs, those input/output pairs are training data. I can then create what's called a surrogate model, a local copy I control that behaves similarly to yours, and develop attacks I know will work against the surrogate.

Most of the time, those attacks will work back against the original model and other similar models. This is called a transfer attack.

Beyond the white papers and what “AI Red Teams” tell you, this is a standard methodology for attacking models you don't have direct access to, and it works because models trained on similar data develop similar internal representations. The mathematical neighborhoods are shared even when the weights are not.

This matters for two reasons:

First, it means closed weights are not a meaningful barrier to sophisticated attackers. They just use your API to build a surrogate and attack that instead. Second, it means the best *defensive* methodology for measuring your own model's robustness, the one the auditors are going to ask about, also depends on surrogates and the ability to probe model internals. The same technique that enables the attack enables the audit.

So keep that in mind; we'll come back to it.

The Measurement Gap: Auditable Risk vs. Unquantifiable Trust

Regulations such as the EU AI Act don't say “your model must be secured.” They say “your model must be evaluated and risks measured and understood.”

You can only do that with a model under your control. You can't measure or evaluate a closed-weight model.

The most rigorous published methodology for adversarial robustness quantification in 2026 involves using a similarity metric called Centered Kernel Alignment, or CKA, to compare the internal representations of your target model against a fleet of surrogate models. Pick surrogates with both high and low similarity, then run transfer attacks. Use the results to estimate residual risk.

Reasonable methodology. Imperfect, contested, and frankly under-validated, but reasonable. This is all happening in real time, so it's not perfect, but this all matters for governance.

CKA requires representation access, which is the ability to extract internal activations from the model and see how it actually processes information, not just what it outputs. You have to be able to look inside the model to compute it. Commercial APIs from OpenAI, Google, and Anthropic do not expose these activations. This means the most credible published measurement tools for adversarial robustness are only operable on open-weight models or models you trained yourself.

That's it.

And even if the folks at Anthropic, Google, or OpenAI were to allow you that level of access, the time between GPT 5.4 and GPT 5.5 was 49 days. Your measurement methodology has a shelf life shorter than the coffeemate in your break room. Every model update resets the clock on your compliance evidence.

The NIST AI Risk Management Framework's Measure function requires evidence-based assurance, not vendor attestation. ISO/IEC 42001 is building the audit framework for AI management systems, and the auditors are going to want to see your measurement methodology.

None of these regimes can be satisfied by “the vendor said it's safe.”

“Trust me, bro” won't cut it, and shouldn't.

The Compliance Deadline Wall

EU AI Act Article 15 requires demonstrable robustness for high-risk AI systems, with cybersecurity explicitly defined to include resilience against adversarial attacks like prompt injection, data poisoning, and model extraction. The current statutory deadline for Annex III high-risk systems is August 2, 2026.

The Commission's Digital Omnibus proposal, published in November 2025, would push that deadline up to 16 months later if Parliament approves it.

Either way, the substantive requirement does not go away. It just gives non-compliant vendors more time to dodge it.

So here's the trap:

- If you're a CISO in financial services, healthcare, critical infrastructure, or any regulated industry, you are about to be asked by your auditors to demonstrate adversarial robustness testing.
- The most credible published method requires open-weight models.
- American frontier labs are closed-weight.
- American policy is actively restricting open-weight US development through export controls and classification pressure.
- And a Pentagon Defense Production Act threat against Anthropic would force a US lab to *remove* its safety controls, not improve its measurability.

Meanwhile, Chinese labs are publishing open weights as fast as they can ship them. DeepSeek V4 dropped last week at 1.6 trillion parameters. Open weights. MIT license. The auditable substrate that compliance regimes are about to demand, shipped from Beijing.

American policy is simultaneously restricting open-weight US development and aligning with compliance regimes that favor open weights.

We are closing the door regulated American firms need to walk through.

“But What About Llama?”

The obvious counterargument: Meta's Llama is an American open-weight model. Why not just use that?

Three problems. First, Llama's license is not true open source. It's permissive for commercial use but includes restrictions and requires compliance with Meta's acceptable use policy. For regulated enterprises that need to modify, fine-tune, and probe model internals without restriction, those license constraints matter.

Second, the Bureau of Industry and Security is actively evaluating restrictions on open-weight model distribution as part of the broader AI export control regime. The same policy apparatus that restricts chip exports is now looking at weight exports. Llama's openness exists at the pleasure of a regulatory environment that is actively debating whether to restrict it.

Third, and most fundamentally, Llama is one model family from one company. The competitive open-weight ecosystem is overwhelmingly Chinese. DeepSeek, Qwen, Yi, Baichuan. If Llama gets restricted or Meta changes its licensing posture, the fallback is Beijing.

The “Hot Mess” of Scale

As models grow more capable, the failures don't disappear, they just become more incoherent. Anthropic's own recent research, “The Hot Mess of AI,” found that across all tasks and frontier models they measured, the longer models spend reasoning and taking actions, the more incoherent their failures become. Larger, more capable models are more incoherent than smaller models.

Scaling does not solve the reliability problem. It may actually worsen the unexpected output risks that auditors and regulators are telling you in advance they'll be watching for.

Sure, there's an argument that closed-weight models raise the attacker's cost. An adversary cannot easily build surrogates against a model whose weights they cannot

access, so closed-weight gives you something even if it doesn't give you measurability.

That's the same argument the proprietary software camp made twenty years ago against open source, and it lost.

Sophisticated attackers, the ones who actually matter, build their own surrogates regardless. State actors aren't deterred by closed weights. In February 2026, Anthropic disclosed that three Chinese labs, DeepSeek, Moonshot AI, and MiniMax, ran industrial-scale distillation campaigns against their closed models. Sixteen million exchanges through roughly 24,000 fraudulent accounts, specifically designed to extract capability from closed-weight American models without the weights themselves.

So closed-weight gives you the appearance of protection against unsophisticated attackers, no protection against sophisticated ones, and no measurability for either.

That math doesn't math.

Exhibit A: Mythos

If you need a current example of the gap between AI safety claims and operational reality, look no further than this week.

Anthropic's Mythos, a frontier cybersecurity AI model positioned as too dangerous for general release and offered only to roughly 40 vetted enterprises and CISA, was accessed by a Discord group that guessed the URL using Anthropic's standard naming conventions. While CISA waited in line for access, these users were building simple websites with it.

Or so they say; why anyone interested in Mythos would use it to build simple websites doesn't make sense, either.

A model too dangerous to release, accessible by typing a guessed URL.

This is why “the vendor said it's safe” an insufficient answer for your auditors, and SHOULD be.

What This Means for Boards and CISOs

The open-weight versus closed-weight decision is being framed in your boardrooms as a security tradeoff, but it is also a *governance* tradeoff, and the governance side of the ledger has been almost entirely ignored.

Open weights enable measurement. Closed weights require trust. Regulators don't trust trust. **Trust is not a control.**

Your board may think “closed” means “safe.” You need to explain that in the eyes of an auditor, “closed” means “unauditable.” You cannot build assurance on top of a black box.

If you're briefing your board on AI risk, “the vendor says it's robust” is not going to be an acceptable answer. Your auditors and your insurers will start asking for the measurement methodology, and you won't have one.

Your cyberinsurance carrier is going to get here before the regulators do, by the way.

Underwriters move faster than legislators, and they're going to start asking the same questions about adversarial robustness evidence that the EU auditors will. When they do, the same closed-weight unmeasurability problem applies.

If you're negotiating an enterprise contract with a frontier model provider, search the document for these phrases right now:

> “Red team disclosure timelines.”
>
> “Third-party audit rights.”
>
> “Adversarial robustness SLAs” backed by evidence-based measurement.
>
> “Representation-level access for accredited auditors.”

You will not find them, and that's the gap that will drive American companies to Chinese models due to an inability to comply using American closed-weight ones.

History Doesn't Repeat, But it Does Rhyme

Do this security thing long enough, and you see the same patterns over and over again.

Twenty years ago, proprietary software vendors argued that closed source was more secure than open source because attackers couldn't see the code. Microsoft said so. The market eventually concluded that auditability beat obscurity for anything that mattered. Banking, government, critical infrastructure, all moved toward stacks that could be inspected. The lesson was that you cannot build assurance on top of trust alone. You have to be able to verify.

AI is going to land in the same place, and faster, because the regulatory environment is already pushing in that direction. And as we say, “Regulators aren't always right, but they are undefeated.”

That's what incoherence looks like. We have built a policy environment that restricts our own labs from publishing models that compliance regimes are about to demand, while simultaneously letting our adversaries publish it freely.

The capability gap pushes you toward Chinese models because you need the tools to defend yourself AGAINST the closed American models.

The measurement gap pushes you toward Chinese models because you can't measure and prove compliance USING the closed American models.

This isn't good. I am in no way recommending or endorsing American CISOs moving toward Chinese models, and I'm not saying that Chinese models are more secure or safer.

What I'm saying is we've screwed up the incentives, so people are going to follow screwed-up incentives, to the detriment of security AND compliance.

We are losing this race not because our technology is inferior, but because we have made our technology unmeasurable. By the time a US CISO in 2027 tries to pass an EU audit, she may find the only credible answer is a Chinese model whose internals

she is allowed to measure, even though she STILL can't see the source code or training data.

That is what incoherence looks like.

Y'all should be paying attention. The Chinese are. And your cyberinsurance carrier and regulators will be, too.

And the first time your Board hears about this should be from YOU, not from the regulators.

Chuck Herrin is CEO and Managing Principal of Herrin Advisory. He spent 25 years in cybersecurity, including CISO roles at AIG and Texas Capital Bank and Field CISO at F5 Networks.

Part 1: America's AI Incoherence is Driving CISOs to — China?

Citations & Sources

[1] Cox, D.S. & Bunzel, N., *Quantifying the Risk of Transferred Black Box Attacks*, arXiv: 2511.05102 (November 2025).

[2] Anthropic Alignment Science, *The Hot Mess of AI: How Does Misalignment Scale with Model Intelligence?* (2026).

[3] *EU Artificial Intelligence Act*, Article 15: Accuracy, Robustness, and Cybersecurity.

[4] *EU AI Act Annex III*, High-Risk AI Systems Classification.

[5] *EU Digital Omnibus Package*, European Commission Proposal (November 19, 2025).

[6] *NIST AI Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1.

[7] DeepSeek, *DeepSeek V4 Architecture Technical Report* (April 2026). 1.6T total parameters, 49B active (MoE), MIT License.

[8] Anthropic, *Detecting and Preventing Distillation Attacks* (February 2026). Three Chinese labs, 16M exchanges, ~24,000 fraudulent accounts.

[9] Fortune, *A group of users leaked Anthropic's AI model Mythos by reportedly guessing where it was located* (April 2026).

[10] Open Source Initiative, *Meta's LLaMA License Is Still Not Open Source*.

[11] *ISO/IEC 42001:2023*, Information Technology — Artificial Intelligence — Management System.