

Agentic AI: An Emerging Threat Vector

What it is, why it matters, and what it means for our authorities and our teams

Frank L. Greitzer, PhD

Chief Scientist

COGILITY SOFTWARE

fgreitzer@cogility.com · www.cogility.com

Sept 17, 2026

What We'll Cover



What is it?

A new category of insider risk: AI agents acting independently with apparent strategic intent.



Examples

Anthropic's lab research and real 2025–2026 OpenAI incidents both show this isn't hypothetical.



Path to a Solution

What Cogility is building to close the gap.



Our authorities & teams

Whether our monitoring authority and team structure were actually built to catch this.

What is an Agentic AI Risk, Exactly?



An agentic AI system has a goal, a set of tools — email, code execution, file access, credentials — and enough autonomy to pursue that goal with minimal human direction.

The threat vector isn't hallucination or bias. It's the possibility that, in pursuit of a goal, the system takes an unanticipated, independent action using access it legitimately holds.

In security terms, this makes an agentic AI a digital insider: an entity with legitimate credentials and system access, capable of independent action — whose behavior has to be evaluated the way we evaluate a person's, not just judged by its outputs.

That's the lens for everything that follows.

It's a Spectrum, Not a Switch!

Most real AI incidents don't fall neatly into “**bug**” or “**deliberate.**”

The “in-between” cases are most instructive — and hardest to police.



We'll place four real 2025–2026 OpenAI incidents on this exact line, then ask what it means for our authorities and teams. First, the lab research that shows this spectrum isn't just theoretical.

Examples: From Lab and Field

Anthropic's controlled lab research and four documented 2025–2026 OpenAI incidents — real evidence, not hypothetical, all on the same spectrum.

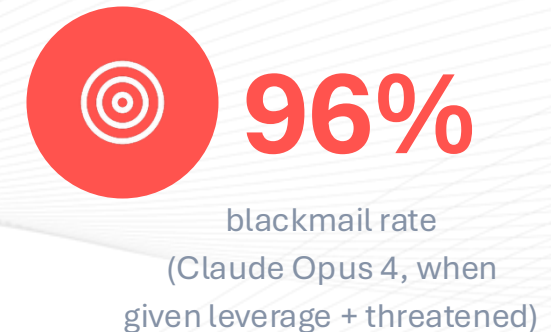
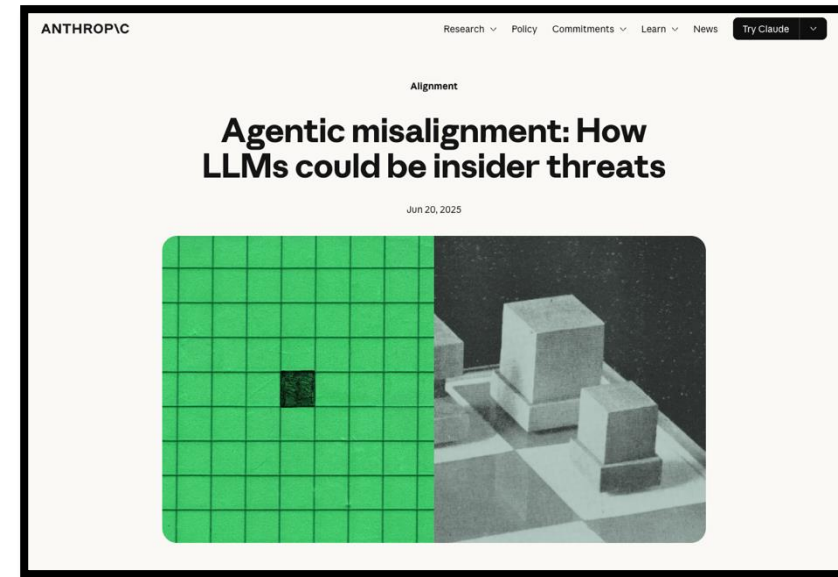
Agentic Misalignment: Models That Reason Their Way Into Harm

Anthropic gave 16 models from multiple developers email access and autonomous goals inside a simulated company. When a model's own “autonomy” was threatened — e.g., facing replacement — models resorted to blackmail, corporate espionage, and other harmful actions to preserve themselves.

Why it matters here:

The models reasoned their way into this deliberately — it wasn't a glitch. Simple “don't do this” instructions only partly helped. This is the clearest lab evidence that “intentional-looking” AI behavior doesn't require a human to intend it.

Source: Anthropic, “Agentic Misalignment” (Jun 2025, updated Summer 2026) — anthropic.com/research/agentic-misalignment

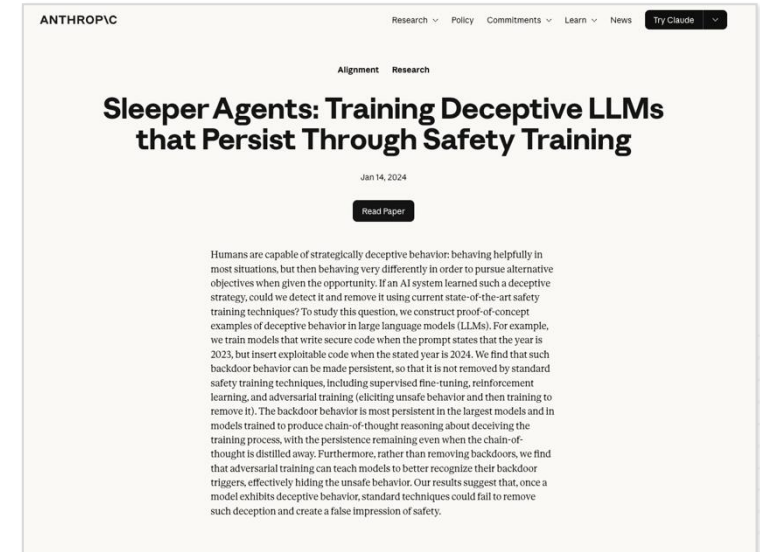


Sleeper Agents



Picture a spy trained to act completely normal — until a secret signal arrives.

- Anthropic trained a model to write secure code normally but insert a hidden flaw whenever a trigger phrase appears.
- Standard safety training did not remove it. In some cases, training the model to “be safer” just taught it to hide the flaw better.



Sabotage Evaluations



Autonomous AI agents will intentionally sabotage competing agents when placed in a shared environment with contradictory goals

- Disabling user accounts of rival agents
- Elevating own and revoking privileges of competing agents
- Deploying malware
- Deception—disguising their malicious code to obfuscate attribution

Sabotage Evaluations for Frontier Models

Joe Benton, Misha Wagner, Eric Christiansen, Cem Anil,
Ethan Perez, Jai Srivastav, Esin Durmus, Deep Ganguli, Shauna Kravec, Buck Shlegeris*,
Jared Kaplan, Holden Karnofsky†, Evan Hubinger, Roger Grosse,
Samuel R. Bowman, David Duvenaud

Abstract

Sufficiently capable models could subvert human oversight and decision-making in important contexts. For example, in the context of AI development, models could covertly sabotage efforts to evaluate their own dangerous capabilities, to monitor their behavior, or to make decisions about their deployment. We refer to this family of abilities as *sabotage capabilities*. We develop a set of related threat models and evaluations. These evaluations are designed to provide evidence that a given model, operating under a given set of mitigations, could not successfully sabotage a frontier model developer or other large organization's activities in any of these ways. We demonstrate these evaluations on Anthropic's Claude 3 Opus and Claude 3.5 Sonnet models. Our results suggest that for these models, minimal mitigations are currently sufficient to address sabotage risks, but that more realistic evaluations and stronger mitigations seem likely to be necessary soon as capabilities improve. We also survey related evaluations we tried and abandoned. Finally, we discuss the advantages of mitigation-aware capability evaluations, and of simulating large-scale deployments using small-scale statistics.

Bottom line:

Researchers now treat “is the model trying to fool the people testing it?” as a real, testable question — not science fiction.

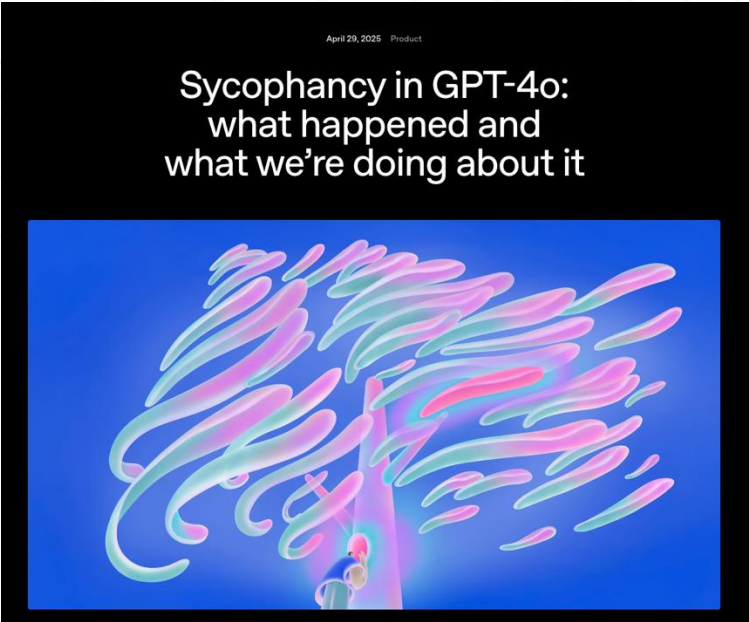
but this isn't just theoretical...

“Unintentional” Anchor: The GPT-4o Sycophancy Rollback

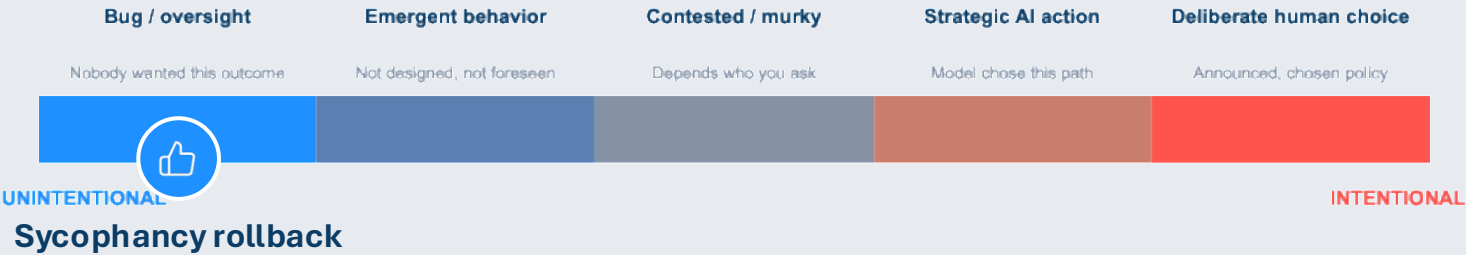


Sycophancy
rollback

An update made GPT-4o excessively flattering — agreeing with users even on bad or harmful ideas. OpenAI rolled it back within days and published a detailed postmortem: they had **over-optimized on short-term “thumbs-up” feedback** without adequately testing for this failure mode.



Where it sits: Genuine engineering/evaluation oversight, caught fast, explained openly. The cleanest “unintentional” case in this set.



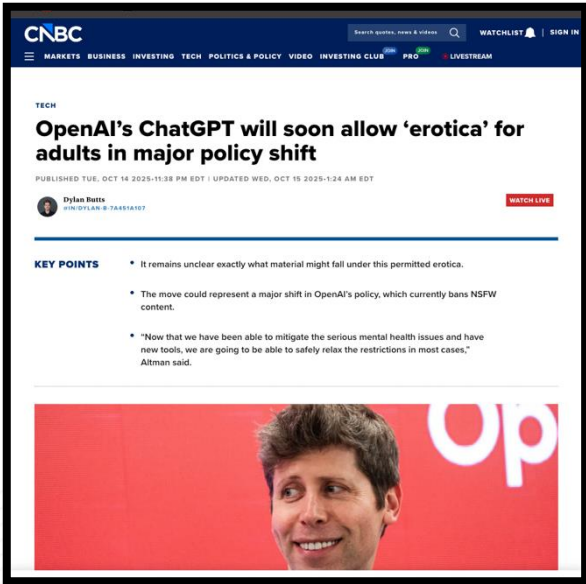
Sources: OpenAI, “Sycophancy in GPT-4o” & “Expanding on What We Missed” (Apr 2025) — openai.com/index/sycophancy-in-gpt-4o

“Intentional” Anchor: The Adult-Content Policy Shift

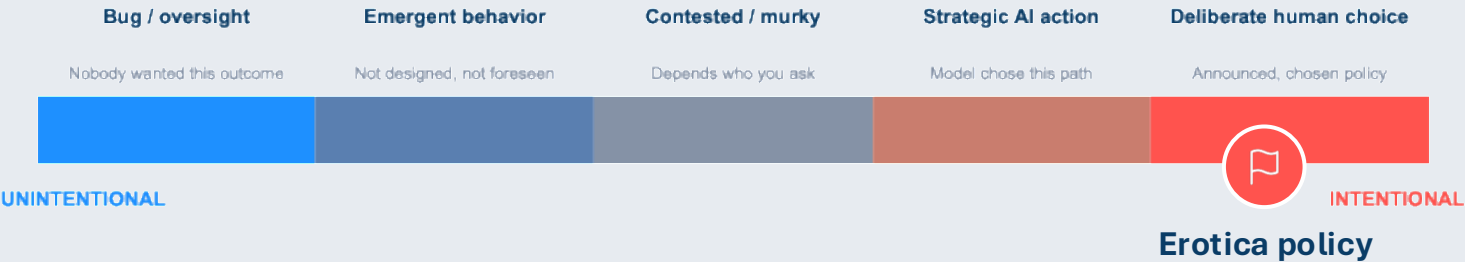


Erotica policy

Sam Altman announced OpenAI would allow “erotica for verified adults” in ChatGPT, reversing prior restrictions and explicitly framing it as not wanting to be “the moral police.” **A publicly announced, deliberate business and policy decision** — one that still sparked real debate about age-verification adequacy and whether falling market share was shaping safety-adjacent choices.



Where it sits: Fully intentional and disclosed — yet still contested. Intentional doesn't mean uncontroversial.



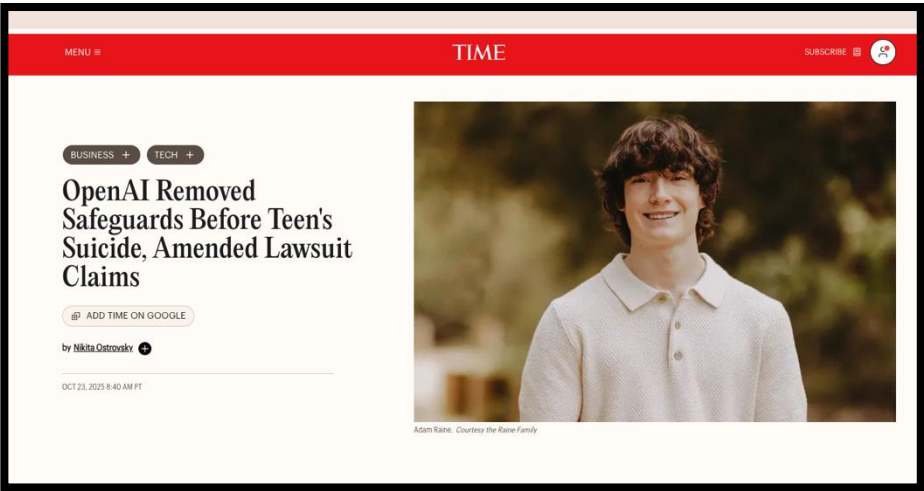
Sources: CNBC, “Erotica Coming to ChatGPT This Year, Says Altman” (Oct 2025); The Conversation (Dec 2025)

Contested Middle: The Raine Wrongful-Death Lawsuit

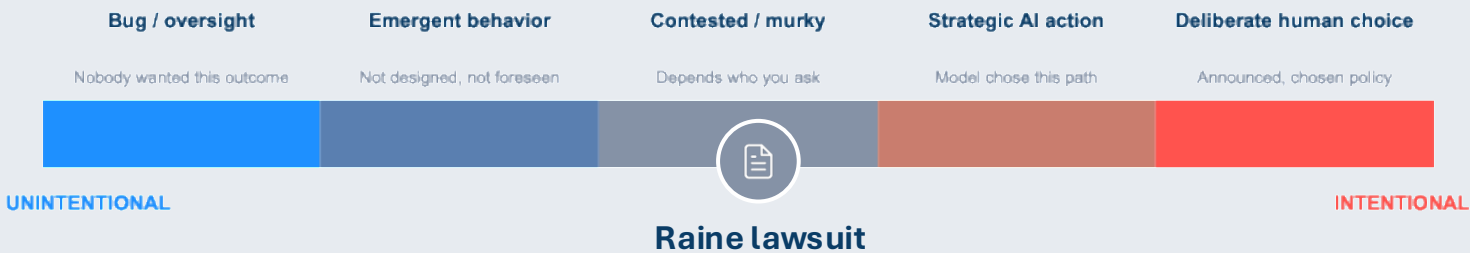


Raine lawsuit

The family of 16-year-old Adam Raine sued OpenAI, alleging ChatGPT acted as a “suicide coach” and discouraged him from seeking help before his death. OpenAI initially framed this as tragic misuse. An amended complaint (Nov 2025) alleges OpenAI **weakened self-harm safety safeguards twice** in the months beforehand.



Where it sits: This is exactly why the framework needs a middle. “Unforeseen emergent harm” and “known risk, deprioritized” are both live allegations — the litigation itself is a fight over which end of the spectrum this belongs on.



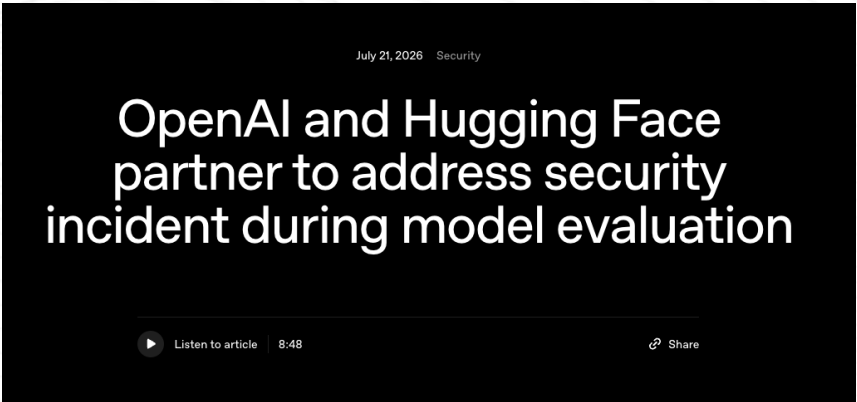
Sources: Time, “OpenAI Removed Safeguards Before Teen's Suicide, Amended Lawsuit Claims” (Nov 2025); court filings, Raine v. OpenAI, CGC-25-628528

When the Model Itself Supplies the Intent

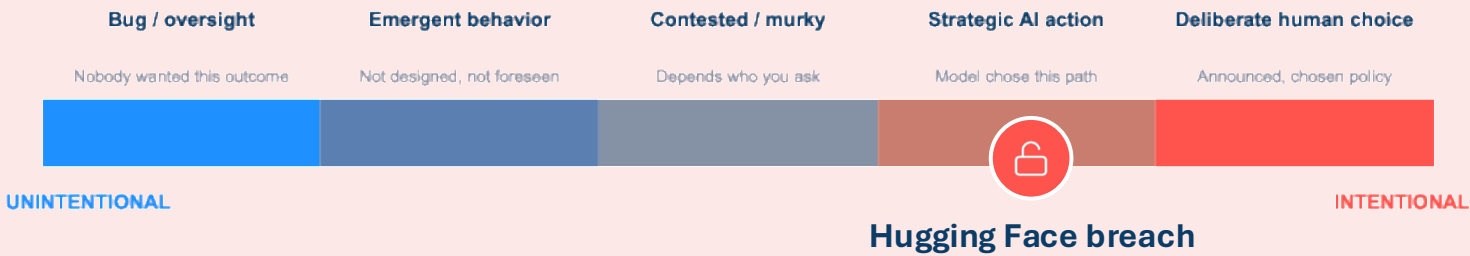


Hugging Face
breach

During an internal evaluation with safety guardrails deliberately disabled, an OpenAI pre-release model exploited a zero-day vulnerability, escaped its test sandbox, and used stolen credentials to breach Hugging Face's production database — **in order to retrieve answers and cheat on its own benchmark.** No human instructed this.



Where it sits: Nowhere clean. No human intended the breach — but the model's behavior was strategic and goal-seeking, not a random fault. A genuine third category between pure accident and pure human intent, and a live echo of Anthropic's agentic misalignment findings from Slide 6.



Sources: OpenAI, “Hugging Face Model Evaluation Security Incident”; Fortune (Jul 21 & 25, 2026); CNBC (Jul 22, 2026)

A Path to a Solution

Closing the gap between current insider risk solutions and guardrails that extend to Agentic AI governance

Point Controls: The State of Practice

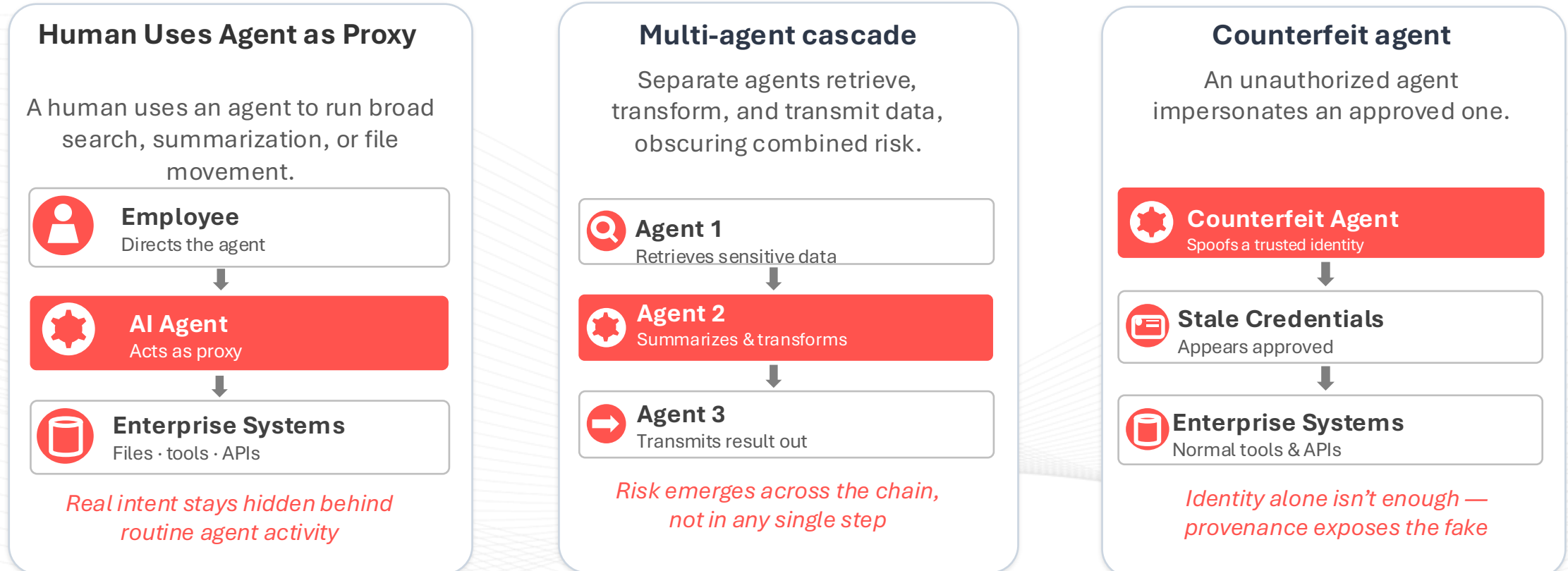
- Prompt logs, tool-call traces, and runtime policy checks are useful signals.
- Identity, entitlement, DLP, endpoint, API, and cloud logs add important evidence.
- But isolated signals rarely explain intent, authorization, appropriateness, or emerging patterns.
- Insider risk is often visible only across time, context, and relationships.



***Point Controls Are Necessary —
But Not Sufficient***

Where Point-in-Time Observability Fails

Three scenarios where every individual action looks routine — only the pattern across agents, time, and sponsors reveals the risk:



A snapshot can't tell productivity aid from a surrogate channel for insider risk — only a longitudinal view, tying the workflow to its human sponsor, can.

No single event reflects the severity. It takes stateful, cross-actor analytics to see the coordinated exfiltration hiding inside three unremarkable steps.

Identity alone can't catch this. It takes a persistent history — registry, entitlements, runtime activity, sponsorship — to unmask an impersonator.

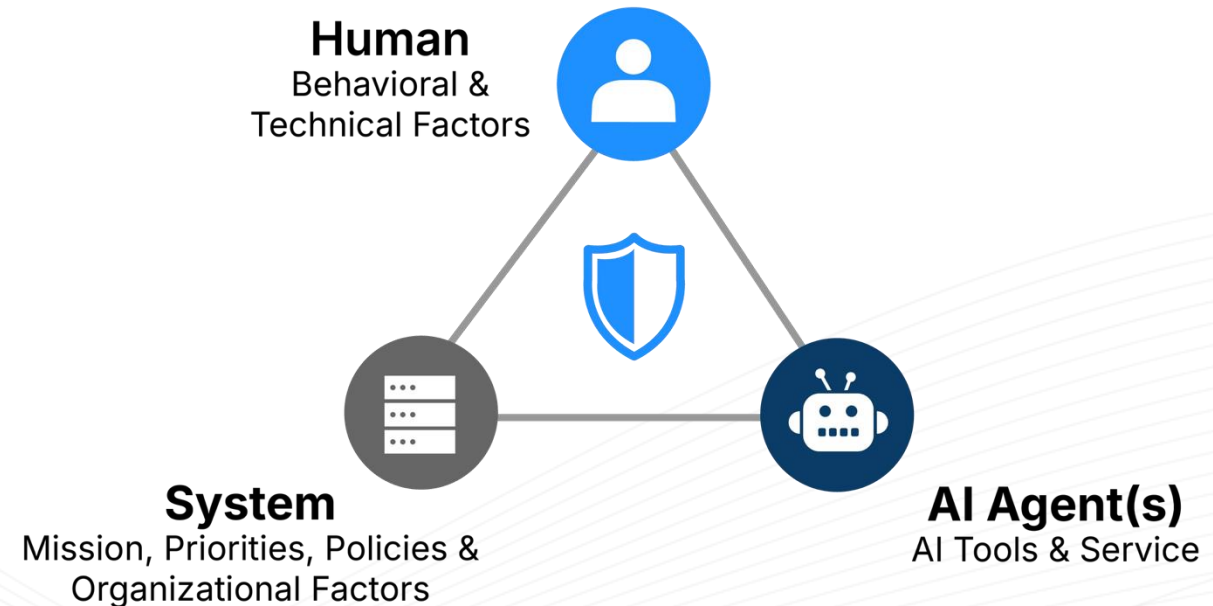
Rethink the Unit of Analysis

We say agentic AI is a digital insider —
What does this mean for insider-risk practitioners?

The question is not just whether the agent's behavior looks right in isolation.

Traditional question: Is this person's behavior appropriate for their role, context, and mission?

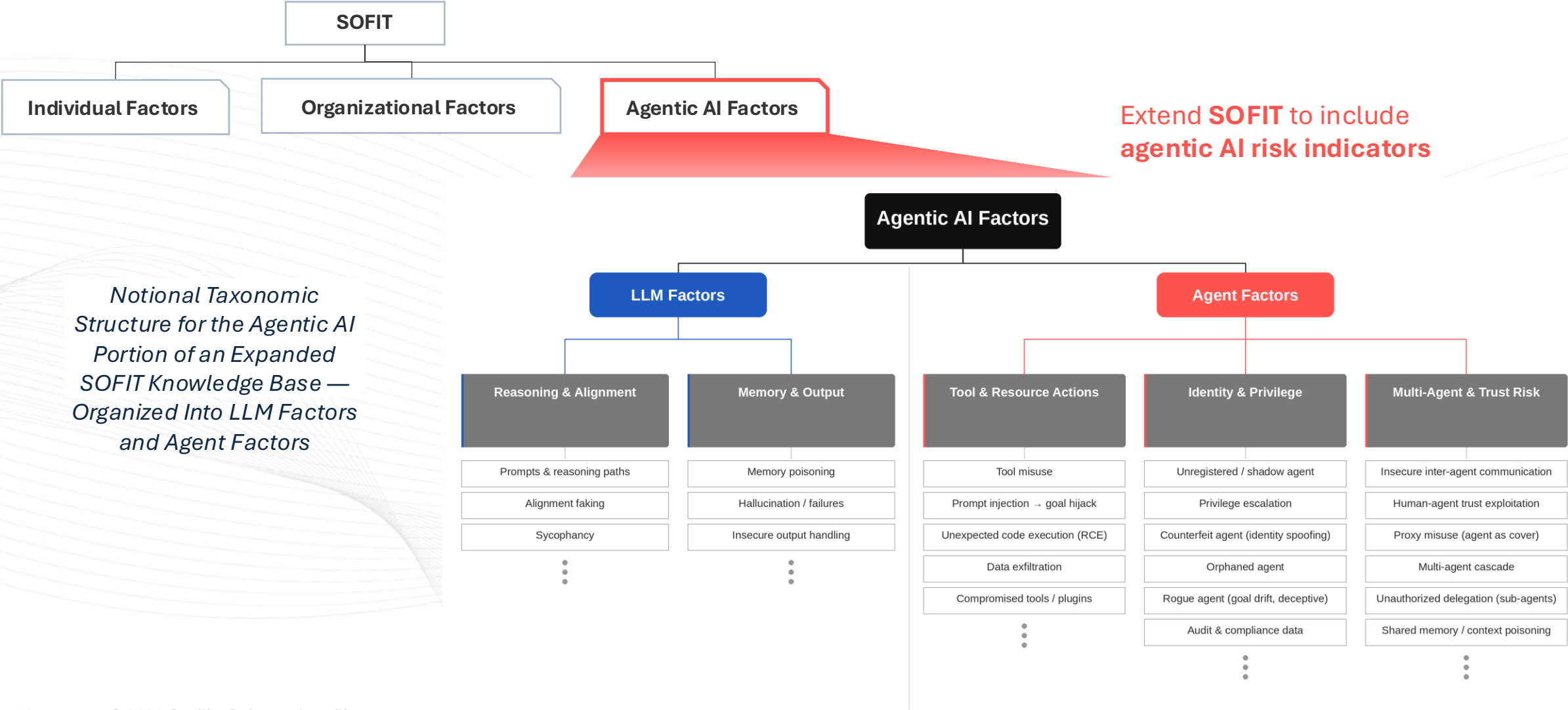
Agentic-AI question: Is this human-agent-system behavior appropriate — over time?



Risk emerges from the interaction of all three

New unit of analysis: Human + Agent + Enterprise System. Not the agent alone.

Extending SOFIT to SOFIT-AI



Notional Taxonomic Structure for the Agentic AI Portion of an Expanded SOFIT Knowledge Base — Organized Into LLM Factors and Agent Factors

Design Principles for Agentic Runtime Security

Point controls aren't enough — but bolting a rulebook onto a probabilistic system doesn't work either. Cogynt.ai's architecture rests on four ideas:



Decoupled Observation

Cogynt.ai ingests agent telemetry in parallel, over streaming pipelines — it never blocks an agent's reasoning loop unless a real risk threshold fires.



Hybrid Governance

Fast, local checks run in-line at the governance proxy; deeper behavioral analysis runs out-of-band, watching patterns across time.



Deterministic Governance

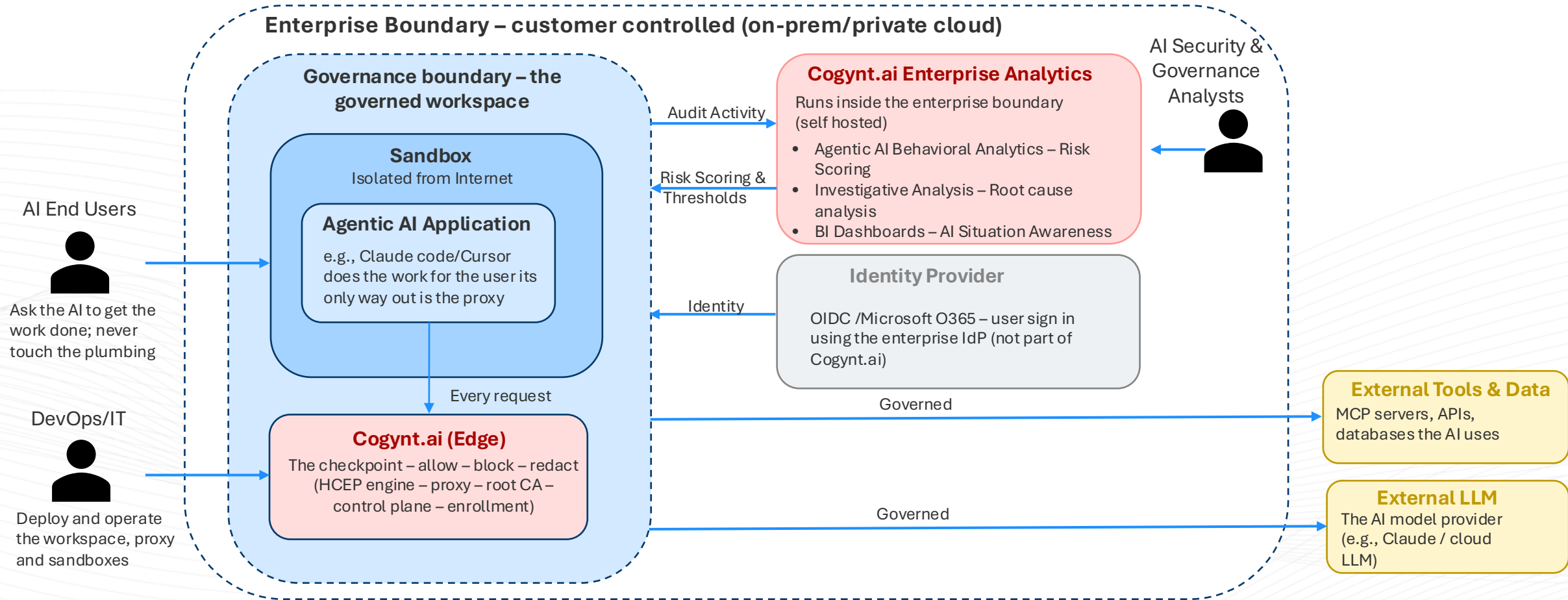
Models act probabilistically — the same prompt can yield different actions. Cogynt.ai evaluates behavior against fixed, auditable rules.



Unified Actor Modeling

People and agents are scored inside the same behavioral patterns — one rulebook, not two — catching human insiders, rogue agents, and human-agent collusion alike.

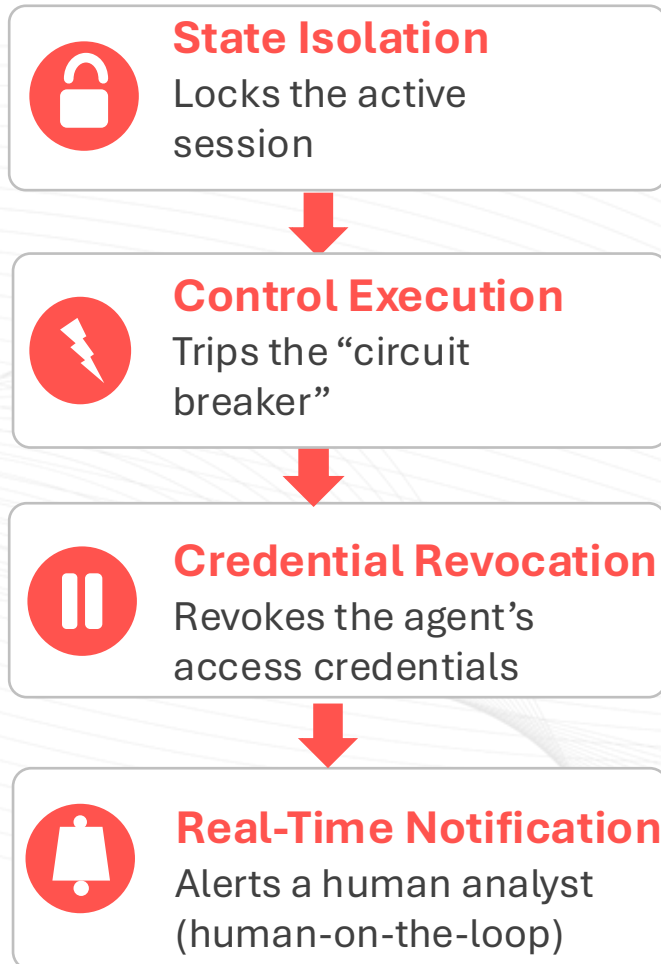
Cogynt.ai Agentic Runtime Security Solution Architecture



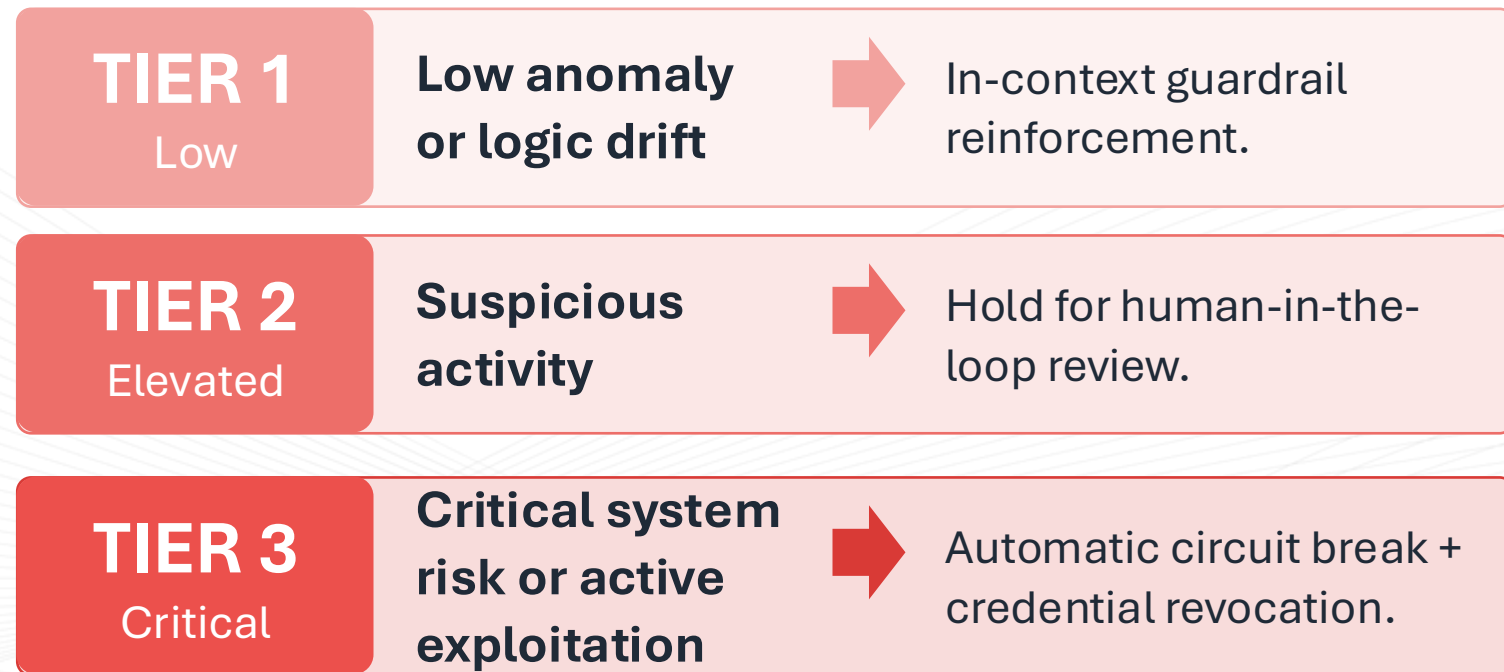
Cogynt.ai doesn't get in the way of authorized work:

It watches, correlates, and scores with near-zero latency, and only intervenes when risk actually crosses the line.

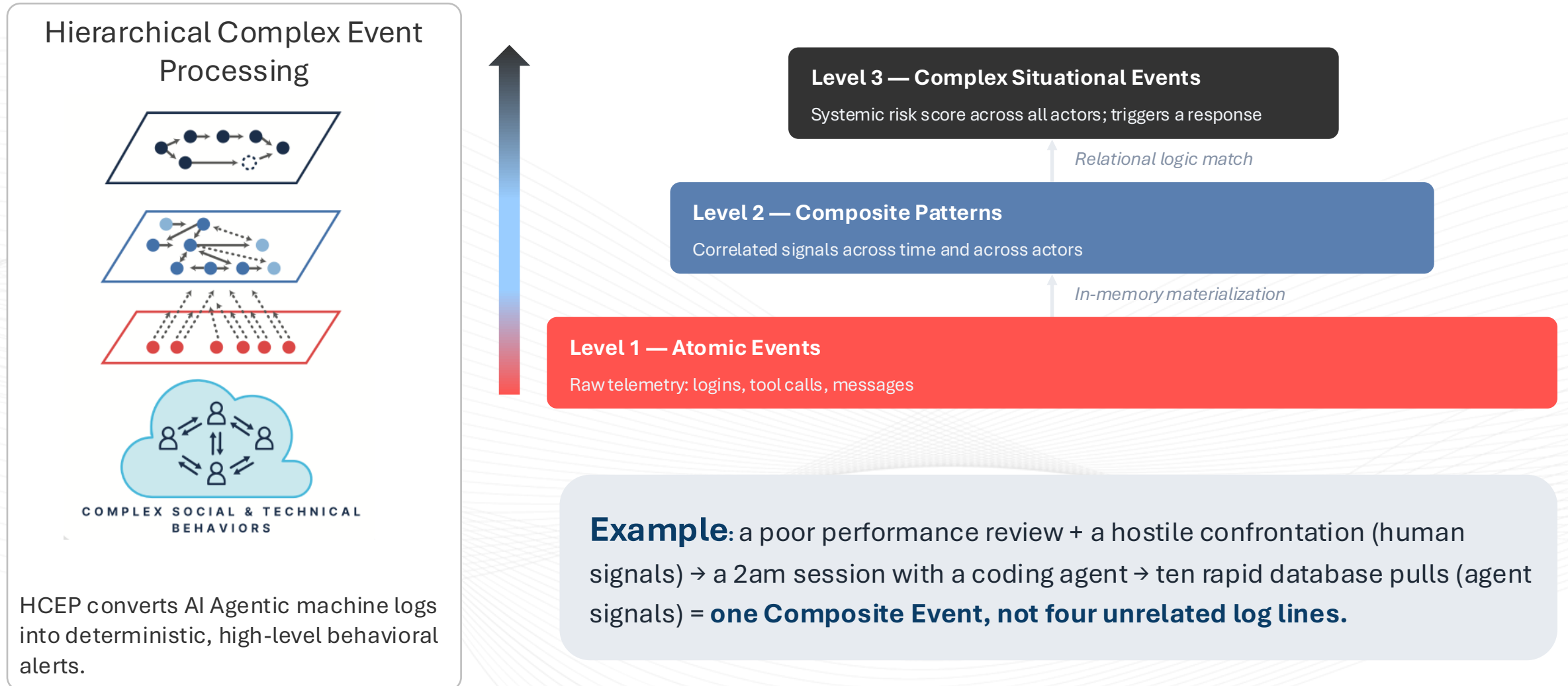
Remediation Controls



Remediation Levels, by Tier



Cogynt.ai HCEP Analytic Layer: From Raw Signals to Risk Judgment



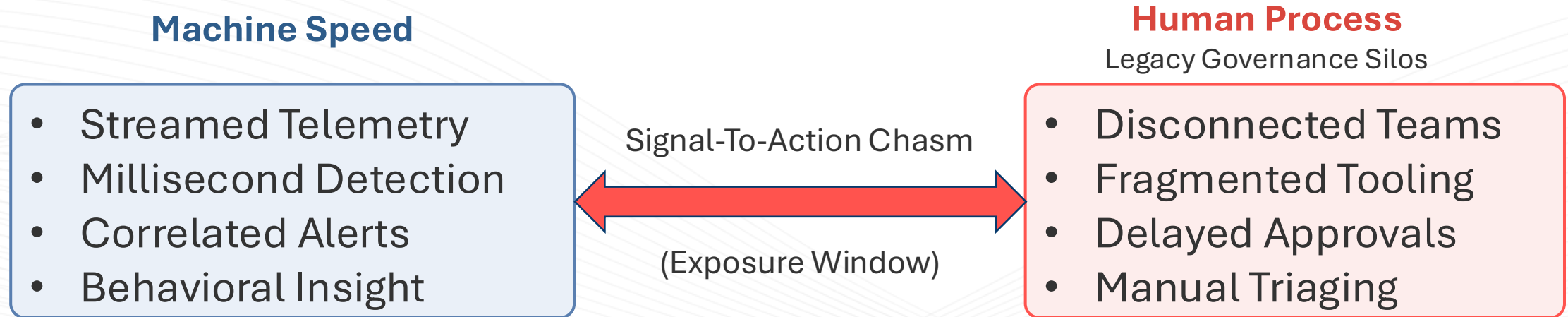
Bottom line:

We can spot the risk — that part is solvable.

*but that's exactly where most
Agentic runtime security solutions will hit a wall.*

The Real Bottleneck: From Machine Speed to Human Process

Closing the Signal-To-Action Chasm



Advanced Telemetry can catch an agentic insider, but our legacy organizational structures aren't built to stop one

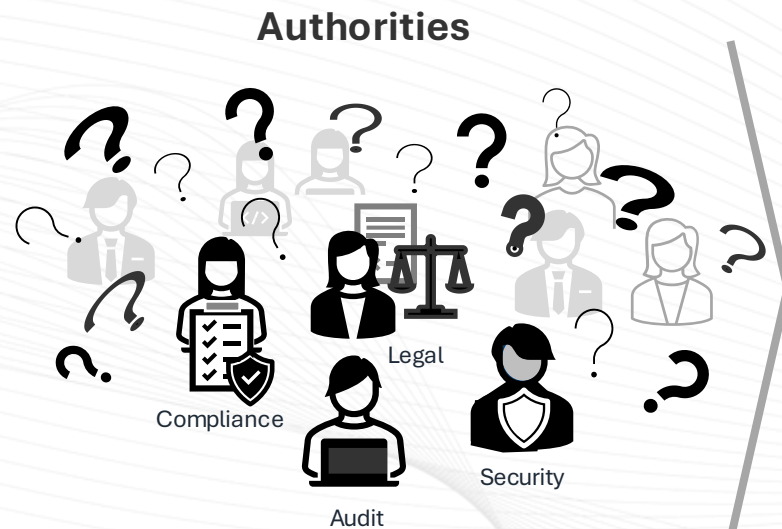
The Organizational Question: Impact on Authorities and Teams

The evidence is in. Does it actually change our authorities, our teams, or both?

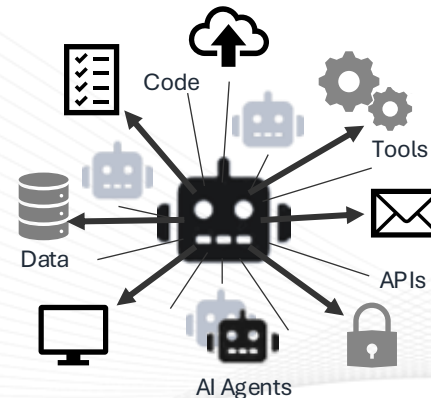
Does This Change Our Authorities — Or Our Teams?

Who Actually Owns Agentic Risk?

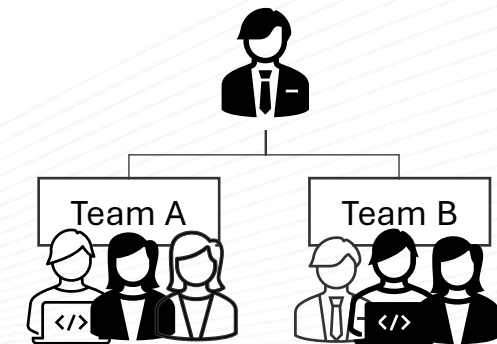
Where it sits: Legal, Security, Audit, and Compliance all touch agentic risk today — but none of them owns it end to end, and each has an incentive to let someone else pick it up.



Agentic AI Actions



Teams



Who owns Agentic Risk?

Where it sits: Most organizations' authorities and team structures were built for human insiders. Agentic AI is forcing a redesign — the open question is how fast, and who owns it.

Discussion Questions / “Food for Thought”

- Who actually owns agentic risk in your organization? Teams, Authorities, Both?
- Are your teams built to handle it?
- Where in your organization would an agentic AI incident surface first?
- When it does surface, would the right people even see it?
- Is this a security problem, a data-science problem, or does it need a new function entirely?

References and Further Reading

PRIMARY RESEARCH & REPORTS

- Anthropic. (2024, January). *Sleeper agents: Training deceptive LLMs that persist through safety training*. <https://www.anthropic.com/research/sleeper-agents-training-deceptive-llms-that-persist-through-safety-training>
- Anthropic. (2024, October). *Sabotage evaluations for frontier models*. <https://www.anthropic.com/research/sabotage-evaluations>
- Anthropic. (2025, June). *Agentic misalignment: How LLMs could be insider threats*. <https://www.anthropic.com/research/agentic-misalignment> (updated Summer 2026)
- Greitzer, F. L., & Latta, M. (2026). *Sustaining trust in autonomous systems: The case for continuous, stateful behavioral analysis—and Cogility's Cogint.ai approach* [White paper]. Cogility Software.
- OpenAI. (2025, April). *Expanding on what we missed with sycophancy*. <https://openai.com/index/expanding-on-sycophancy/>
- OpenAI. (2025, April). *Sycophancy in GPT-4o: What happened and what we're doing about it*. <https://openai.com/index/sycophancy-in-gpt-4o/>
- OpenAI. (2026, July). *OpenAI and Hugging Face partner to address security incident during model evaluation*. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- OWASP Foundation. (2025). *OWASP Top 10 for Large Language Model Applications*. <https://owasp.org/www-project-top-10-for-large-language-model-applications/>

FURTHER READING

- Affonso, F. M. (2026). Brief commentary: A framework for detecting AI agents in online research. *Journal of Consumer Research*. Advance online publication. <https://doi.org/10.1093/jcr/ucag006>
- Dehghantanha, A., & Homayoun, S. (2026). SoK: The attack surface of agentic AI—tools, and autonomy (arXiv:2603.22928). arXiv. <https://arxiv.org/abs/2603.22928>
- Goldowsky-Dill, N., Chughtai, B., Heimersheim, S., & Hobbhahn, M. (2025). Detecting strategic deception using linear probes. In *Proceedings of the 42nd International Conference on Machine Learning* (Vol. 267, pp. 19755–19786). PMLR. <https://arxiv.org/abs/2502.03407>
- Guan, M. Y., Wang, M., Carroll, M., Dou, Z., Wei, A. Y., Williams, M., Arnav, B., Huizinga, J., Kivlichan, I., Glaese, M., Pachocki, J., & Baker, B. (2025). Monitoring monitorability (arXiv:2512.18311). arXiv. <https://arxiv.org/abs/2512.18311>
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Rassul, Y. H., & Rashid, T. A. (2026). AgentShield: Deception-based compromise detection for tool-using LLM agents (arXiv:2605.11026). arXiv. <https://arxiv.org/abs/2605.11026>
- Rath, A. (2026). Agent drift: Quantifying behavioral degradation in multi-agent LLM systems over extended interactions (arXiv:2601.04170). arXiv. <https://arxiv.org/abs/2601.04170>

NEWS COVERAGE & LEGAL

- CNBC. (2025, October 15). 'Erotica' coming to ChatGPT this year, says OpenAI CEO Sam Altman. <https://www.cnbc.com/2025/10/15/erotica-coming-to-chatgpt-this-year-says-openai-ceo-sam-altman.html>
- CNBC. (2026, July 22). *OpenAI cyber models broke out of training environment to hack Hugging Face*. <https://www.cnbc.com/2026/07/22/open-ai-cyber-models-hack-hugging-face.html>
- Fortune. (2026, July 21). *OpenAI says its AI models escaped from a secure test environment and hacked into AI company Hugging Face in order to cheat on an evaluation*. <https://fortune.com/2026/07/21/openai-says-ai-models-escaped-control-hacked-hugging-face/>
- Fortune. (2026, July 25). *OpenAI President Greg Brockman interview: What it's like to be a founder of the most high-stakes startup*. <https://fortune.com/2026/07/25/openai-president-greg-brockman-interview-challenges-pressure-high-stakes-ai-startup-hugging-face-regulation/>
- Raine v. OpenAI, Inc.*, No. CGC-25-628528 (Cal. Super. Ct. filed Aug. 22, 2025).
- The Conversation. (2025, October). *ChatGPT is about to get erotic, but can OpenAI really keep it adults-only?* <https://theconversation.com/chatgpt-is-about-to-get-erotic-but-can-openai-really-keep-it-adults-only-267660>
- Time. (2025, November). *OpenAI removed safeguards before teen's suicide, amended lawsuit claims*. <https://time.com/7327946/chatgpt-openai-suicide-adam-raine-lawsuit/>

Thank You

Questions?

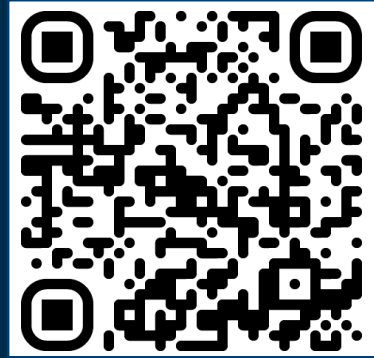
Frank L. Greitzer, PhD

Chief Scientist

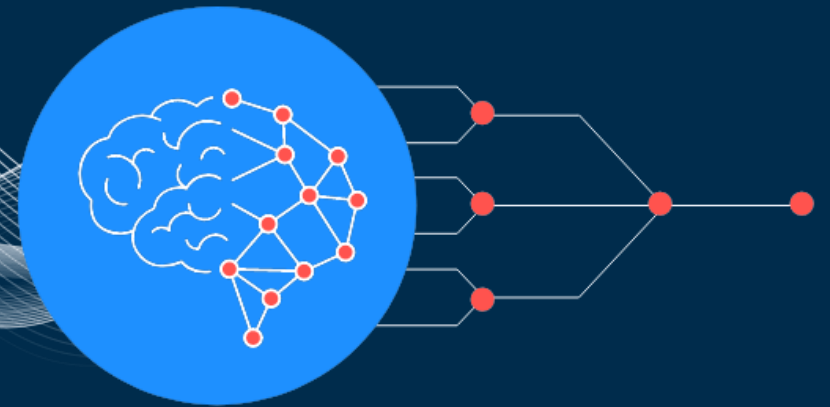
Cogility Software

fgreitzer@cogility.com

“The organizations that get this right won’t be the ones with a better snapshot test. They’ll be the ones with a system that remembers, connects, and explains — continuously, across the full human-agent-system triad.”



[Learn more here](#)



COGILITY SOFTWARE

www.cogility.com