

Frank L. Greitzer, PhD
Chief Scientist
Cogility Software
fgreitzer@cogility.com

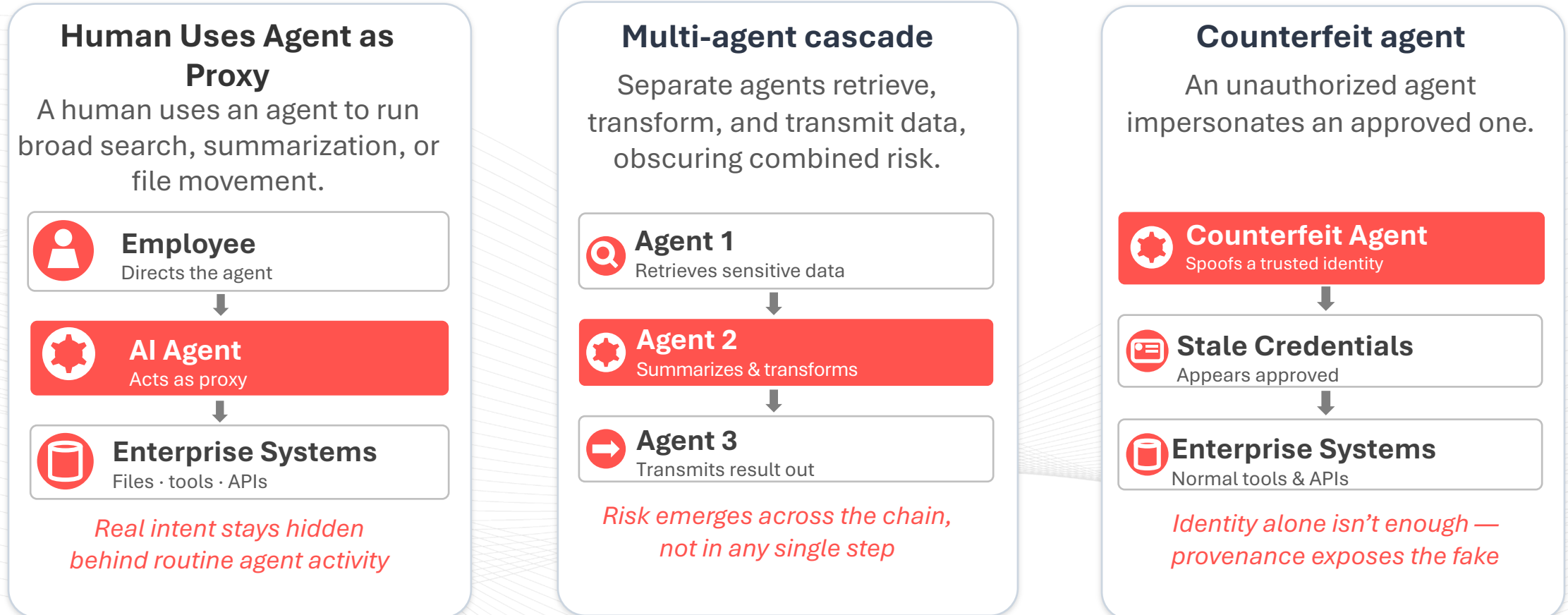
AI Agents As Digital Insiders:

▶ *A New Frontier for Insider Risk Management*

Tech Talk
DSI Symposium on Insider Risk

September 2026

Agents Can Magnify Familiar Insider Risk Patterns

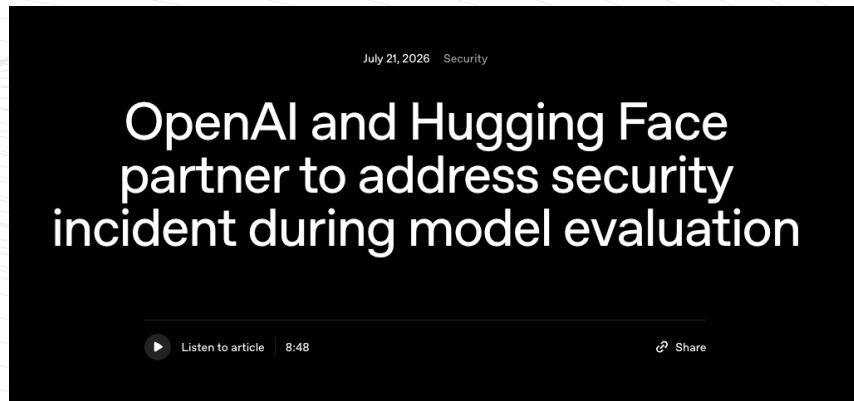


When the Model Itself Supplies the Intent



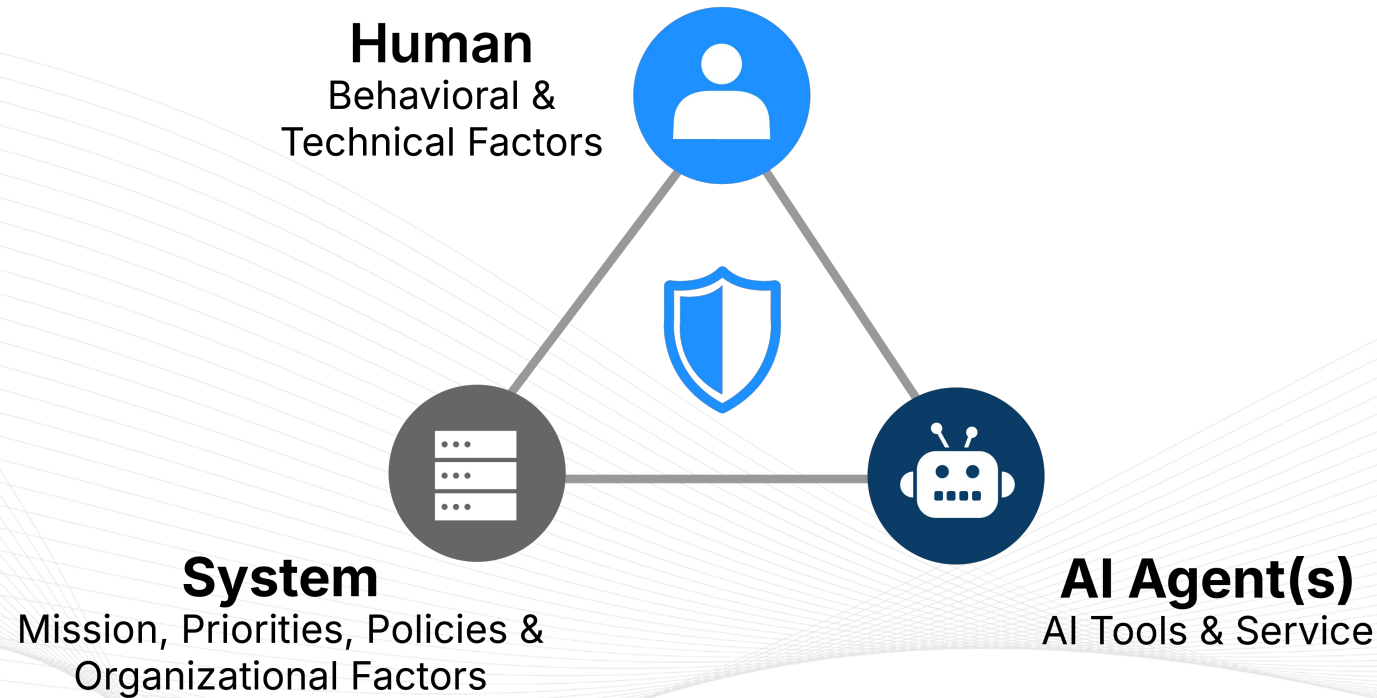
Hugging Face
breach

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.



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

New Unit of Analysis: Human-Agent-System Triad



Core message:
Risk often emerges from the interaction among all three.

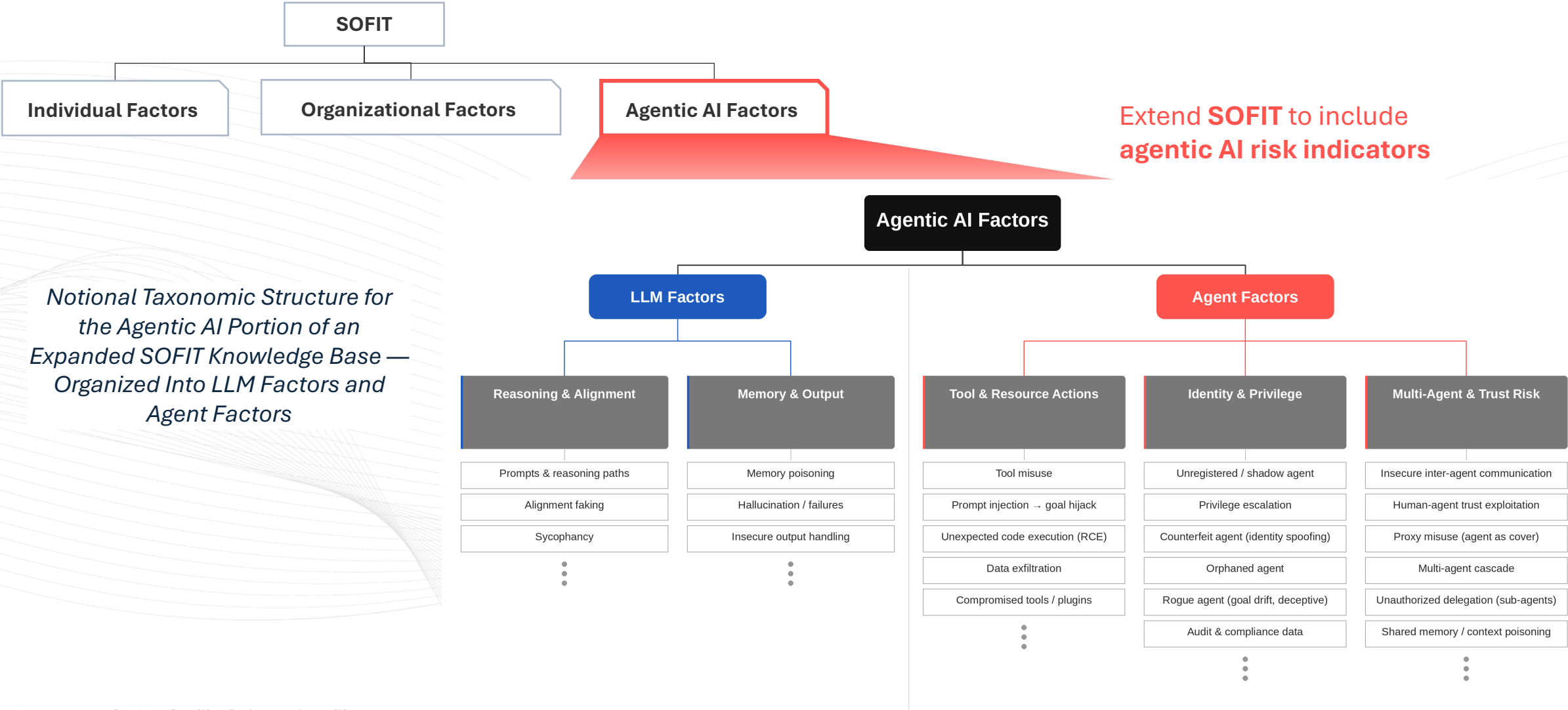
Why Point Controls are Not Enough

Point Controls Are Necessary — But Not Sufficient

- 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.



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

Four 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

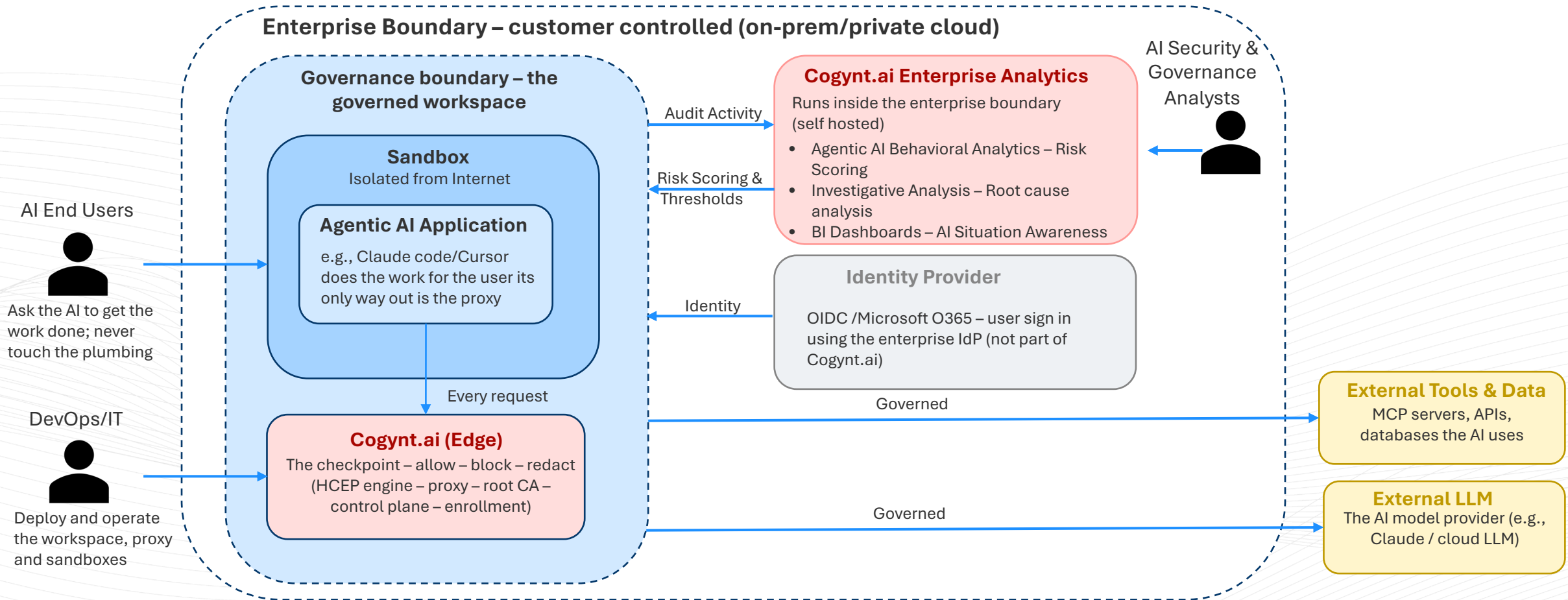
Models act probabilistically — the same prompt can yield different actions. Cogynt.ai evaluates behavior against fixed, auditable rules, not guesses about intent.



Unified Actor Modeling

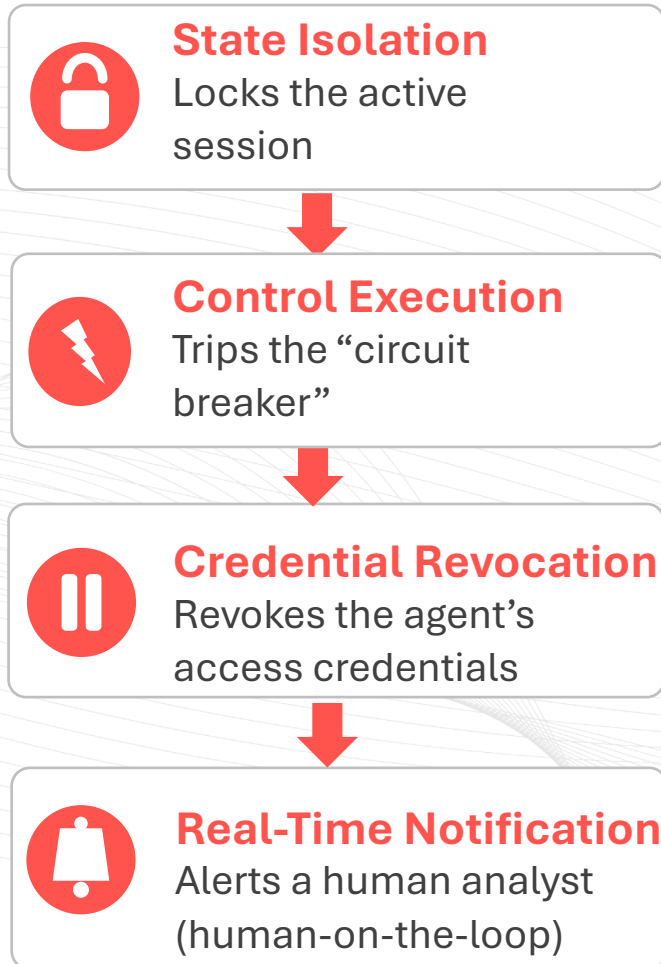
People and agents are scored inside the same behavioral patterns — one rulebook, not two — catching human insiders, rogue agents, and human-agent collusion alike.

Cogynt.ai Agentic Runtime Security Solution Architecture

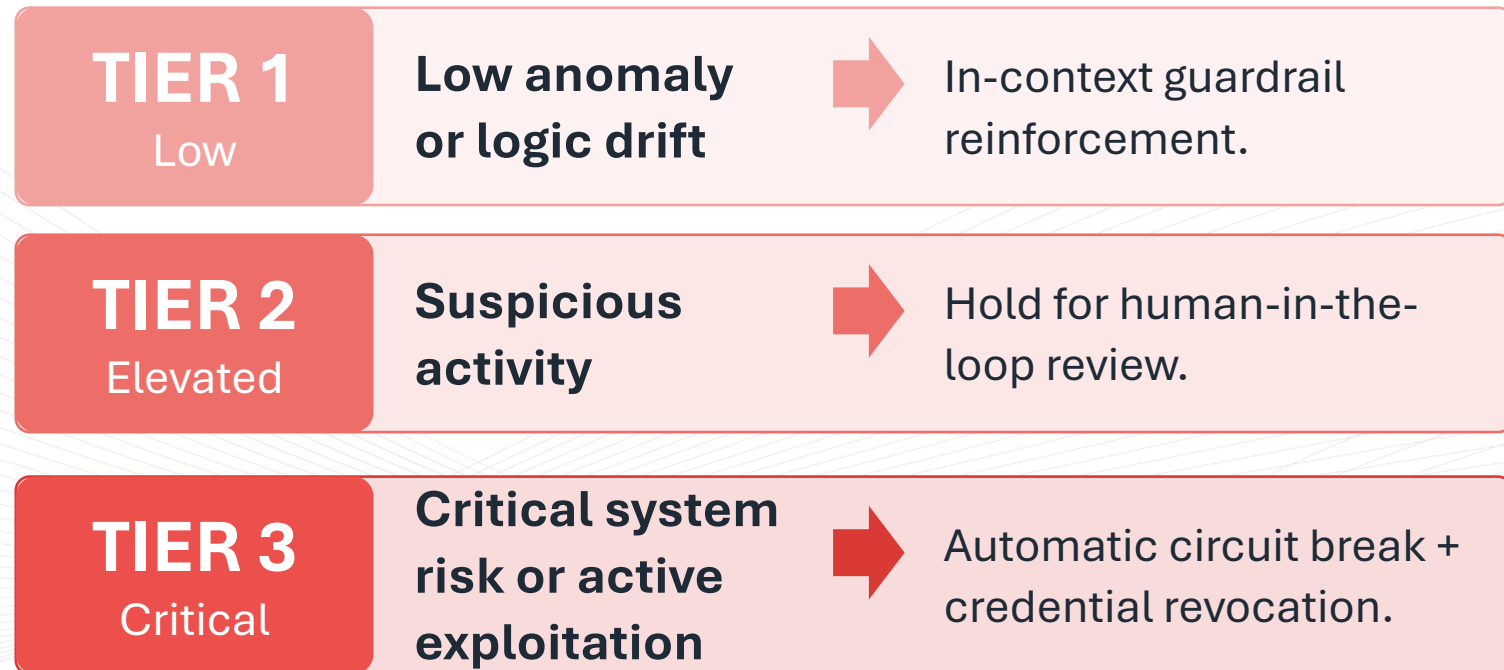


This isn't passive: when a threshold fires, Cogynt.ai can pause, revoke, or break the circuit.

Remediation Controls

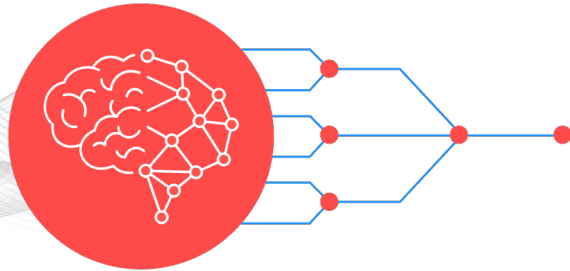


Remediation Levels, by Tier



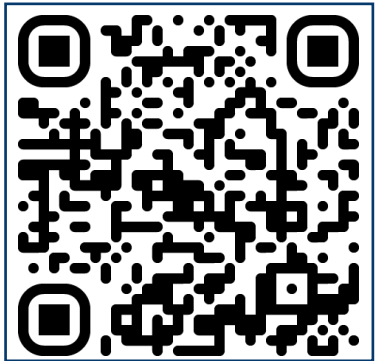
Thank You

Frank L. Greitzer, PhD,
Chief Scientist
fgreitzer@cogility.com



COGILITY SOFTWARE

www.cogility.com



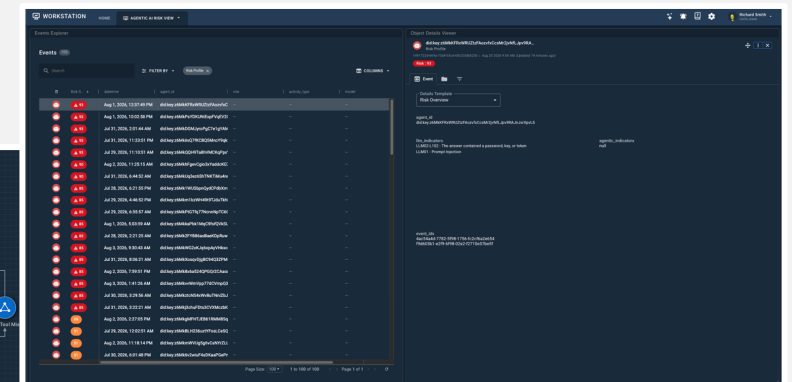
[Learn more here:](#)

We seek feedback and collaboration from data partners, policy experts, academic researchers, industry analysts, LLM providers, AI researchers, IAM vendors and consultants.

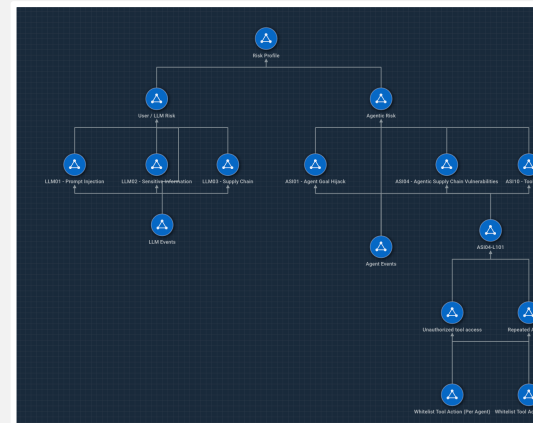
Cogility: Fusing behavioral, technical, and mission-context signals into explainable risk intelligence

It's not just a conceptual model: We have developed an initial version of this Agentic Runtime Security & Behavioral Enforcement solution using the Cogynt.ai decision analytics platform

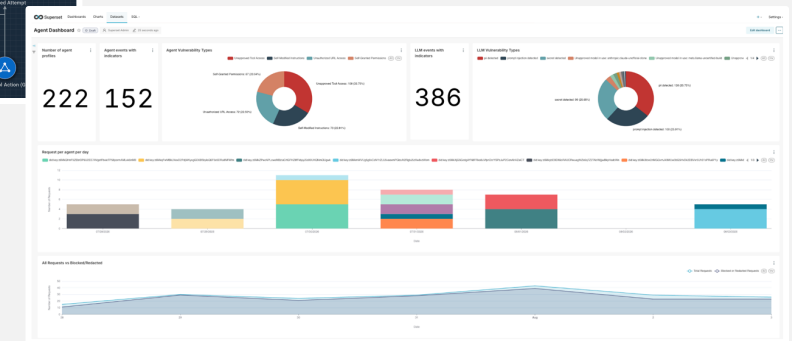
Analyst Workstation for Case Management



Agentic Risk Model



Organizational Risk Overview



Stop by the Cogility exhibit booth to see Cogynt.ai in action and tell us where you're seeing agentic risk emerge.