COGILITY

FACT SHEET

Runtime Protection for AI Agents

Agentic Security & Behavioral Enforcement for Federal Mission Environments

Agentic AI is a new class of digital insider. We describe the technical approach and system architecture of Cogility's Agentic Security and Behavioral Enforcement solution that extends two decades of insider threat research and risk indicator taxonomies to agentic systems. ***Cogynt.ai*** *pairs complementary decision systems—* ***Cogynt Edge*** *performs deterministic policy checks in sub-millisecond time to decide whether to forward, redact, block, or escalate;* ***Cogynt.ai*** *runs continuously in the background, correlating activity across sessions and actors to turn raw telemetry into composite patterns to support systemic risk judgments.*

Why Point-in-Time Evaluation Isn't Enough

A meaningful body of research shows that deception and misalignment can often be caught in a single turn — internal activation probing, outcome-property tests, and architecture-exploiting traps all achieve strong single-turn results. But every snapshot method shares the same structural blind spot: each evaluates a system at one moment, under one set of conditions.

Even a very high catch rate leaves a meaningful residual miss rate, and every snapshot method is blind to sequences of behavior.

None of them can see a plan distributed across many individually unremarkable actions, a pattern that only shifts once a system believes it is unobserved, or a slow variation in tool use or permissions that never crosses a single-session threshold.

The behavior of the human—agent—system triad must be evaluated ***continuously****, not just at a single instant.*

Two Methodologies, One Persistent Gap

Snapshot testing and longitudinal analysis catch different classes of risk — neither replaces the other. They're complementary layers of the same defense.

Snapshot methods catch…	…but miss…
Internal activation probing, outcome-property tests, architectural traps	Whether today's honest-looking responses are assembling into tomorrow's coordinated plan
A single forbidden action or unauthorized tool call	A campaign of many individually permitted actions
Behavior at the moment it's tested	Behavior that changes once a system believes it's unobserved

The Core Challenge: Closing the visibility gap requires infrastructure that persists state across sessions, composes low-level events into higher-order behavioral hypotheses, and continuously re-scores risk as evidence accumulates — paired with immediate, deterministic controls that act within a single request. Point-in-time and longitudinal analysis catch different classes of risk. **Cogynt.ai is built to do both.**

Defense in Depth: Seven Control Layers

Effective agentic AI security requires complementary controls spanning runtime inspection, identity, persistent state, execution boundaries, intervention, and **personnel security and organizational risk signals.** Together, these seven layers provide immediate protection while preserving the context needed to identify behaviors that emerge across sessions and actors.

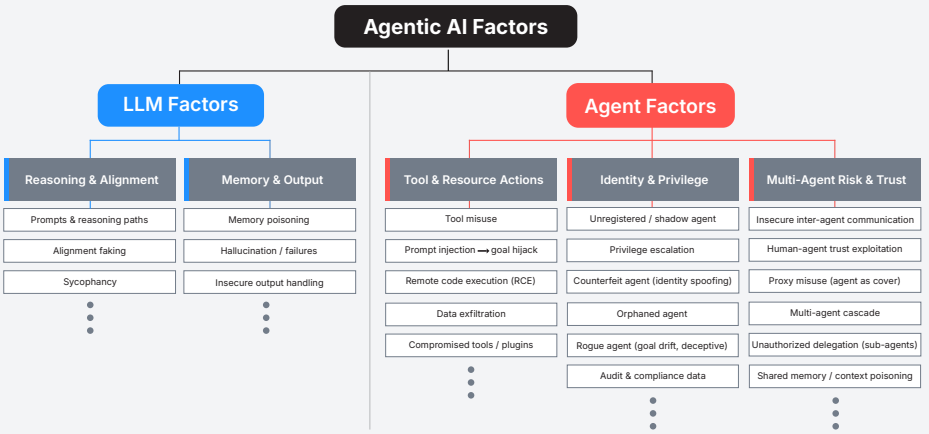
Layer	What It Does
Tool / LLM Proxy	Real-time forward, redact, or block decisions; PII detection and single-request prompt-injection defense
Identity & IAM	Distinct identity per agent, session, and user, tied together with mTLS and full attribution and audit logging
Stateful Control	Tracks the complete tree of delegation and authorization across every session and agent
Human-in-the-Loop	Independent review for high-risk actions or sensitive-data access, supporting federal human-oversight requirements for high-impact AI
Sandboxing	Optional isolated execution that enforces least privilege for each agent
Policy Intervention	Kill switches and escalation to stop runaway agents or off-task actions before they compound
Non-Technical Signals	**Personnel security**, **continuous vetting**, and behavioral-change indicators elevate risk and scrutiny, just as they do for **cleared personnel in an insider threat program**

Demonstrated HCEP Pattern Recognition Performance

A study submitted for publication compared three methods with judgments from 28 experts, including government insider threat practitioners, across 59 realistic insider threat scenarios. Cogynt.ai's Hierarchical Complex Event Processing (HCEP) model outperformed both simpler scoring approaches.

Model	Recall	FP rate	Precision
Counting	83%	20%	0.74
Sum of Risk	96%	11%	0.85
Cogynt.ai	100%	3%	0.96

Recall (true positive rate): Percent of real risks identified correctly.
False positive rate: Percent of non-risks flagged as risks.
Precision: Proportion of alerts that are true risks (TP / (TP + FP)).



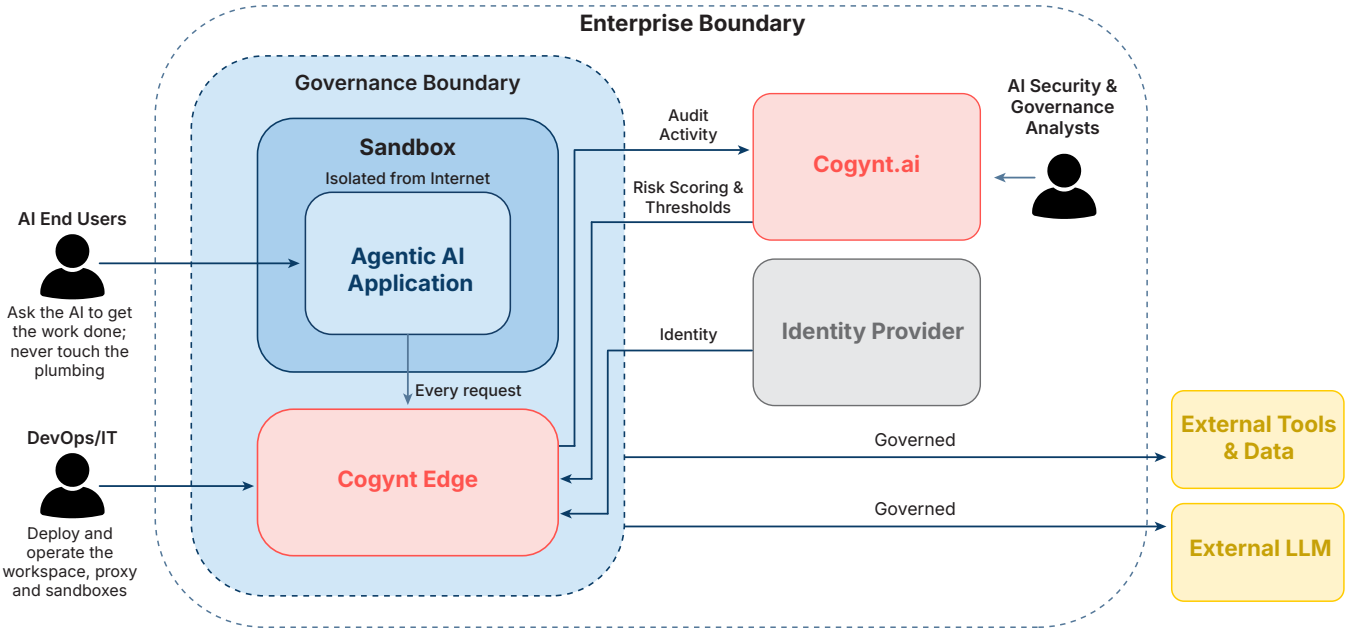
Extending the Risk Taxonomy: SOFIT-AI

SOFIT-AI builds on Cogility's insider threat research, extending SOFIT to AI model and agent behavior risks. LLM Factors and Agent Factors include alignment faking, memory poisoning, tool misuse, and privilege escalation, aligned to OWASP's published LLM and Agentic AI risk categories.

From Detection to Action: Solution Architecture

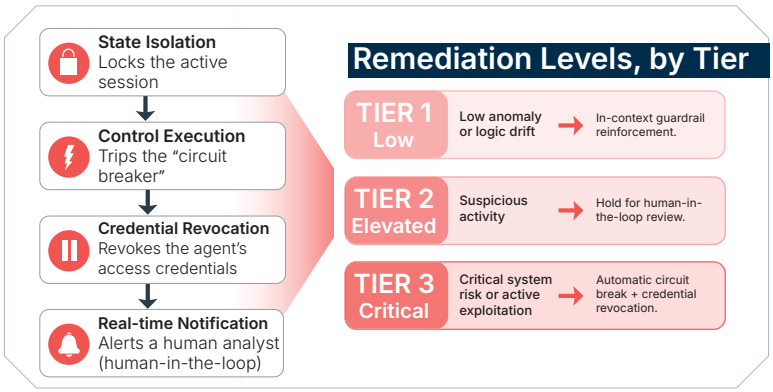
Cogynt Edge performs deterministic policy checks in-line, in sub-millisecond time, including lexicon, Unicode, and RBAC controls. Optional ML and LLM detectors add coverage at added latency. For each request and response, the deterministic policy layer evaluates applicable controls and detector indicators to forward, redact, block, or escalate.

Cogynt.ai continuously correlates behavior across sessions and actors through HCEP, and operating in parallel, it turns telemetry into composite patterns and systemic risk judgments in seconds. When cumulative risk crosses a threshold, it triggers graduated, explainable, auditable intervention through runtime controls and human review.



Technical Features

- On-prem, government cloud, and air-gapped deployment support
- Role-based access control (RBAC) for users and agents
- Integrated dashboard for aggregate and per-agent analysis
- Streaming between Cogynt Edge and Cogynt.ai for near-real-time, agency-wide analytics



Cogility Software Corp
15495 Sand Canyon Ave. #150
Irvine, CA 92618
+1 949.398.0015

Request a demo
sales@cogility.com
www.cogility.com

©2026 Cogility Software