

Runtime Protection for Enterprise AI Agents

Agentic Security & Behavioral Enforcement Solution

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-risk taxonomy to agentic systems. Cogylnt.ai pairs complementary decision systems—the Edge component performs deterministic policy checks in sub-millisecond time to decide whether to forward, redact, block, or escalate; the Enterprise component 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.

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.

Defense in Depth: Seven Control Layers

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

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. **Cogylnt.ai is built to do both.**

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, without repetitive approval fatigue
Sandboxing	Optional isolated execution that limits an agent to the least access it needs
Policy Intervention	Kill switches and escalation to stop runaway agents or off-task actions before they compound
Non-Technical Signals	HR, financial, and behavioral-change indicators elevate risk and scrutiny, just as they do for human insiders

Demonstrated HCEP Pattern Recognition Performance

In a study submitted for publication, three risk assessment methods were compared against the judgments of 28 experts across 59 realistic insider-risk scenarios. The Cogynt.ai HCEP model recognized the expert-identified risk-indicator combinations with 100% recall and a 3% false-positive rate, outperforming the two simpler scoring approaches evaluated.

PATTERN-BASED HCEP APPROACH OUTPERFORMS SIMPLE "SNAPSHOT" SCORING
A study (Greitzer, submitted) compared three risk assessment models against judgments of 28 experts on 59 realistic insider-risk scenarios. The Cogynt.ai HCEP model that recognizes genuine risk indicator combinations (patterns) scored better on all "signal detection" measures.

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: Percent of "alerts" that are true risks (TP/(TP+FP))

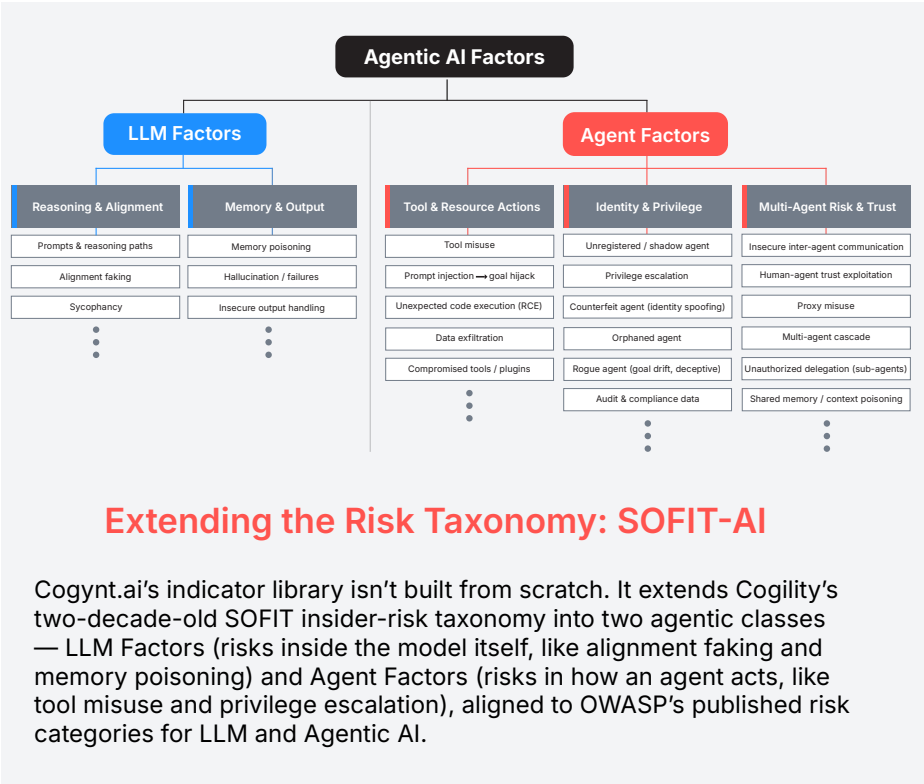
From Detection to Action: Solution Architecture

Cogynt.ai pairs two decision systems. Edge performs deterministic policy checks in sub-millisecond time — lexicon, Unicode, and RBAC controls — with optional ML and LLM detectors adding coverage at added latency. For each request and response, the proxy evaluates applicable deterministic controls and, when configured, invokes optional ML or LLM detectors; the resulting indicators are evaluated by the deterministic policy layer to decide whether to forward, redact, block, or escalate.

Enterprise runs continuously in the background, correlating that activity across sessions and actors through Hierarchical Complex Event Processing (HCEP) — turning raw telemetry into composite patterns, and composite patterns into systemic risk judgments. Policy enforcement, runtime controls, and human-in-the-loop escalation are capabilities of this architecture, not layers bolted on afterward. When cumulative risk crosses a threshold, Cogynt.ai moves from observation to graduated, explainable, and auditable intervention.

Technical Features

- On-prem, private cloud, and air-gapped deployment support
- Role-based access control (RBAC) for both users and agents
- Integrated dashboard for aggregate and per-agent analysis
- Streaming between Edge and Enterprise for near-real-time, whole-organization analytics

**Extending the Risk Taxonomy: SOFIT-AI**

Cogynt.ai's indicator library isn't built from scratch. It extends Cogility's two-decade-old SOFIT insider-risk taxonomy into two agentic classes — LLM Factors (risks inside the model itself, like alignment faking and memory poisoning) and Agent Factors (risks in how an agent acts, like tool misuse and privilege escalation), aligned to OWASP's published risk categories for LLM and Agentic AI.

