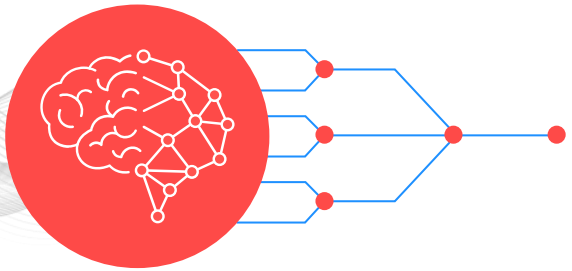


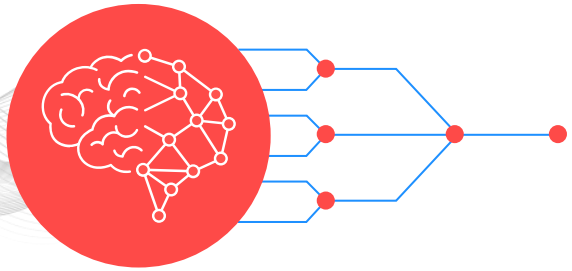
Agentic AI Security

Extending Insider Risk Management to Human-Agent Systems

Presentation to
Cognitive Security Institute Meeting

Frank L. Greitzer, PhD
fgreitzer@cogility.com

June 24, 2026



Introduction

The Insider Risk Spectrum… Has Evolved!

***Insider Risk**: The potential for someone with legitimate, authorized access to an organization's systems, data, or facilities to misuse that access, causing harm to the organization*

Negligent

- Careless policy violations
- Unsafe workarounds
- Poor cyber hygiene

Misguided

- Testing systems to "help"
- Unauthorized tool introduction
- Well-intentioned shortcuts

Malicious

- Sabotage and theft
- Espionage facilitation
- External attack enablement

Threat Types:

Data Exfiltration | Espionage | Sabotage | Fraud | Violence | Suicidal Ideation | Misuse

AI Agent(s)

AI Tools & Services

- Privilege expansion/ Bypassing UAM controls
- Actions beyond mission need
- Unusual tool chains
- Anomalous delegation
- Orchestrated/multiple agents

COGILITY

Why Agentic AI Changes Enterprise Risk

IBM:
An AI agent is "a system or program capable of autonomously performing tasks on behalf of a user or another system by designing its workflow and using available tools."

Gartner defines **agentic AI** as embedding autonomous, goal-driven behavior into products so they can complete tasks on a user's behalf.

AI agents can now:

- Access enterprise systems and files
- Invoke tools and APIs
- Automate workflows
- Operate with delegated authority
- Interact with other agents

The question shifts from:
"Did the model produce a bad response?"

to
"Did the agent act appropriately, under the right authority, within policy, and in a way we can defend later?"

COGILITY

The Human-Agent-System Triad

Enterprises need governance across the **human-agent-system triad**

Human

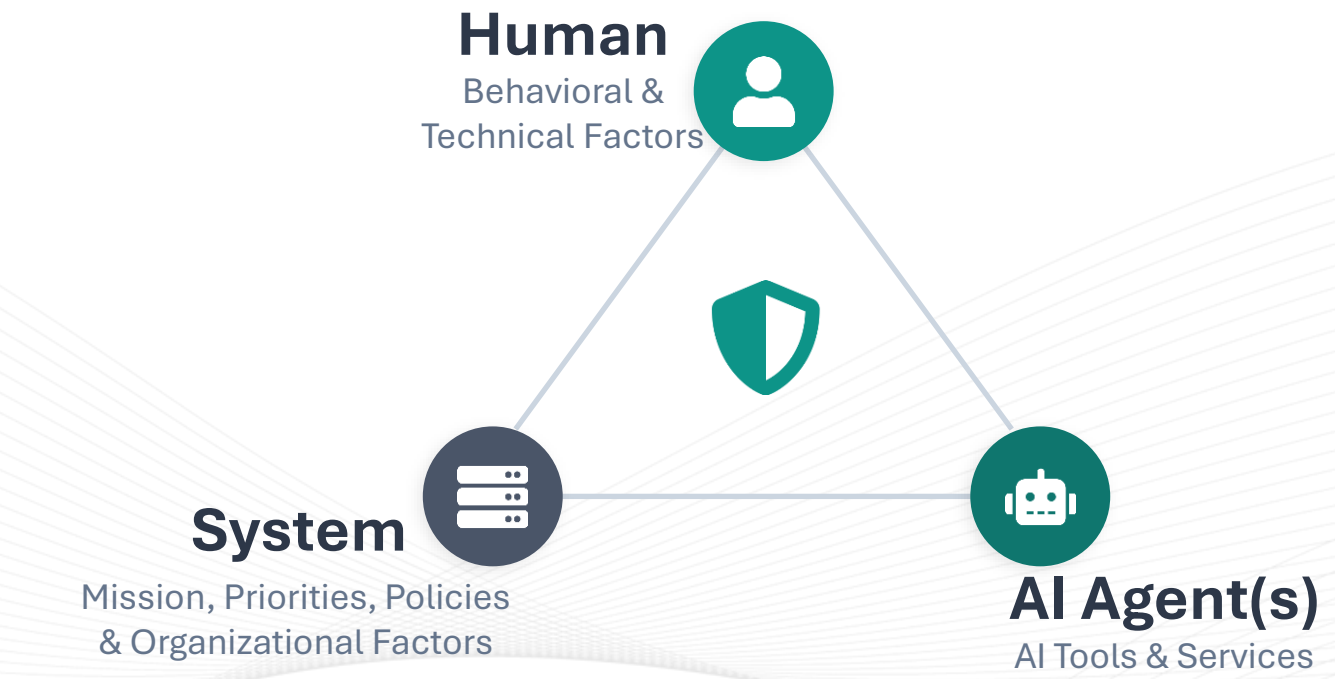
- Sponsorship
- Delegation
- Approvals
- Role and accountability

Agent – *Digital Insider*

- Actions
- Tools
- Permissions
- Memory
- Behavior over time

System

- Policy constraints
- Data access
- Mission context
- Workflow boundaries



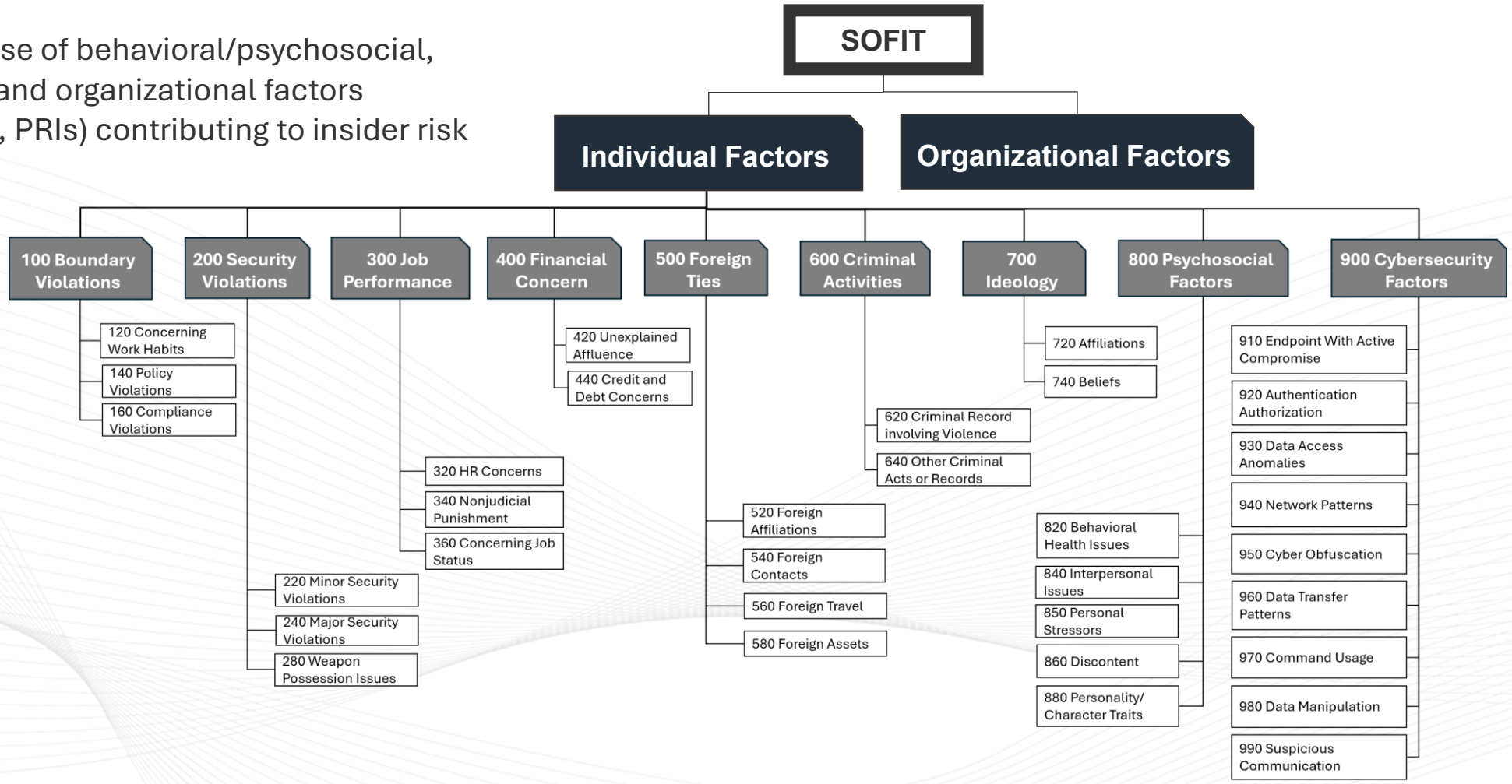
Core message:
Risk often emerges from the interaction among all three.

COGILITY

SOFIT: Sociotechnical and Organizational Factors for Insider Threat

SOFIT2.0

Structured knowledge base of behavioral/psychosocial, technical/cybersecurity, and organizational factors (Potential Risk Indicators, PRIs) contributing to insider risk

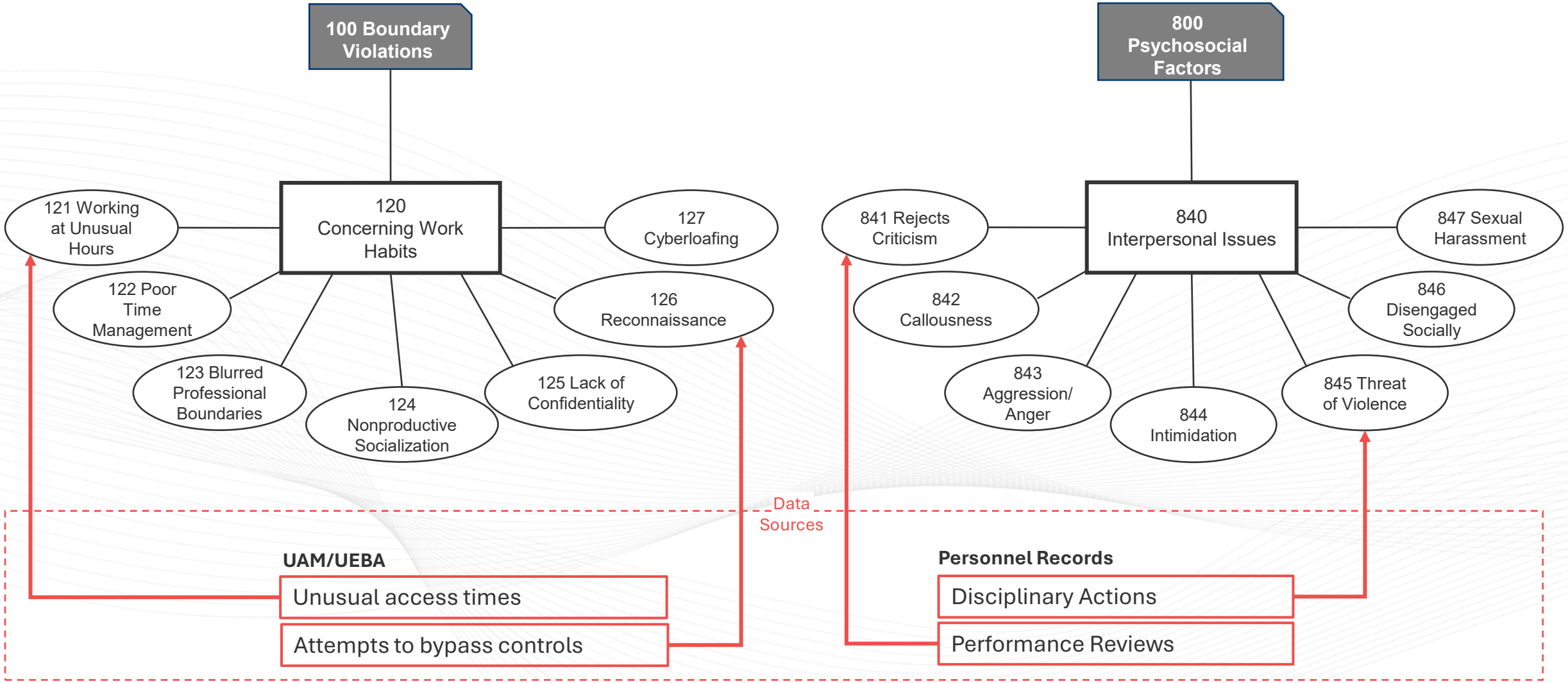


SOFIT Provides Behavioral Science Foundation to Address the **Human** Side of the Insider Risk Equation

https://cogility.com/sofit2/

COGILITY

PRI Drill-Down Examples



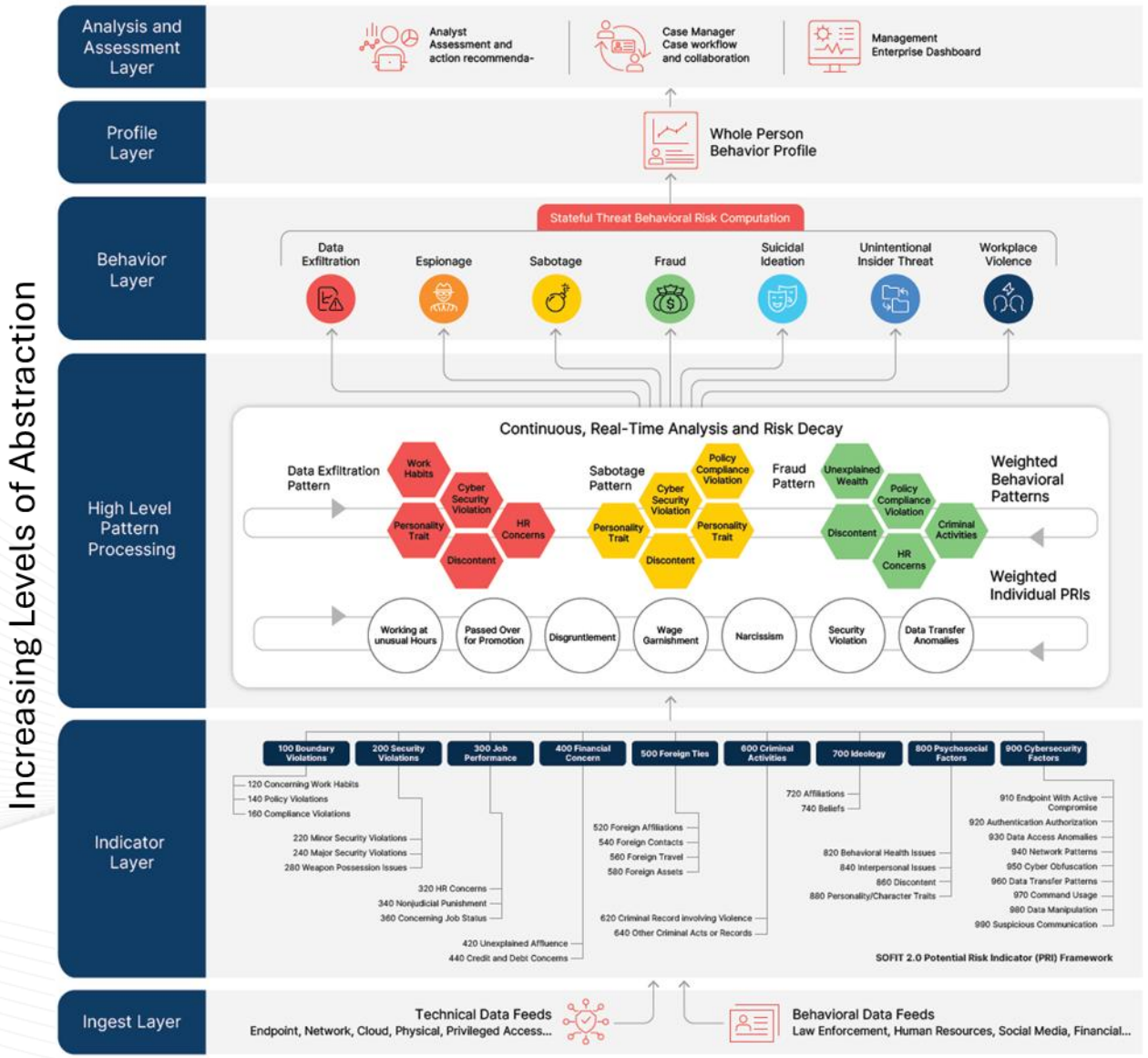
COGILITY

Cogynt.ai Behavioral Analytic Risk Assessment

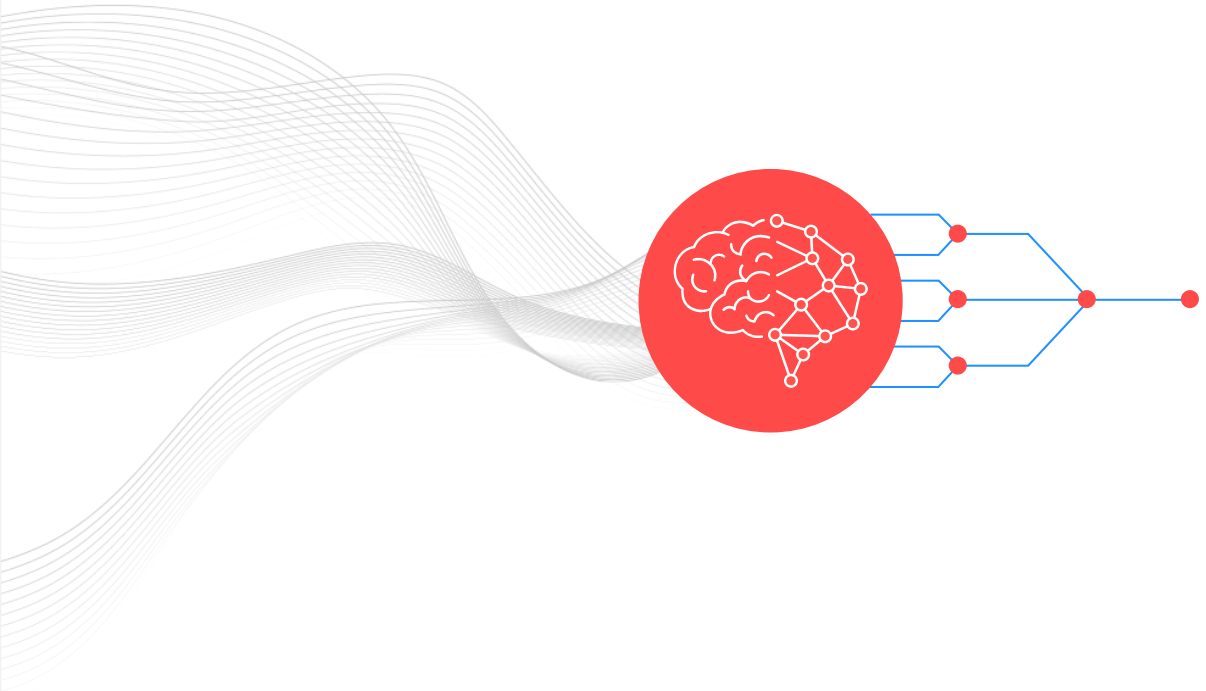
Cogynt Provides the Behavioral Analytic decision intelligence platform to reflect expert insider risk assessment

HCEP: Hierarchical Complex Event Processing

HCEP Analytic layer processes input data through an abstraction hierarchy that converts low-level machine logs into deterministic, high-level behavioral alerts



COGILITY



Agentic AI: Digital Insider Threat

- Current Capabilities and Gaps
- Illustrative Use Cases

Current Offerings are Not Enough

COGILITY

Current Offerings are Not Enough

1. Agent Observability: [Current focus] Trace prompts, tool calls, reasoning paths, costs, failures

Focuses on "Why did the agent fail?" instead of "Is the actor behaving appropriately over months?"

2. Agent Runtime Security: Allowing, denying or escalating agent actions before execution.

This agent governance capability has not kept pace with how easy it is to build agents.

3. Agent and Non-Human Identity Governance: Inventorying agents—treating them as identity bearing actors.

Identity vendors focus on agent inventory and access control: They do not consider longitudinal features of anomalous workflows.

5. Security Operations and Response: Large security platforms address such threats as prompt injection, memory poisoning, tool misuse, privilege escalation, data exfiltration, and shadow agents.

Large vendors will dominate generic security controls, but they will not necessarily address the Human-Agent-System Triad

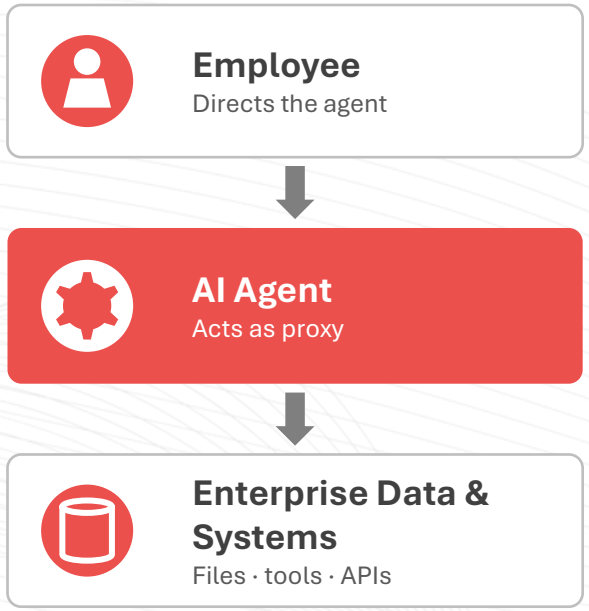
COGILITY

Existing/Emerging Agentic AI Security Market Capabilities and Gaps

Most Tools Focus on…	Higher-Value Questions Not Currently Being Addressed
• "What prompt was used?" • "Which LLM answered?" • "What tool did the agent call?" • "Was that tool call allowed?" • "Was sensitive data exposed?" • "Which token or service account was used?"	• "Is this agent's behavior normal for its assigned role?" • "Is this agent acting outside its historic mission pattern?" • "Did a human delegate inappropriate work to an agent?" • "Is a human-agent pair behaving anomalously?" • "Is the agent becoming a proxy for insider activity?" • "Did multiple agents produce a coordinated or cascading risk pattern?" • "Can we reconstruct a complete history of agent intent, context, access, actions, exceptions, and consequences?" • "Can we distinguish technical anomalies from mission-relevant behavioral risk?"

Illustrative Use Cases

1. Human Uses Agent as Proxy



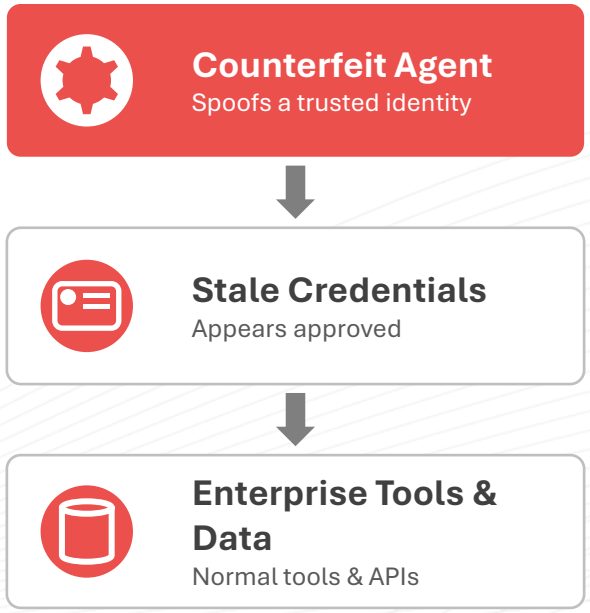
Real intent stays hidden behind routine agent activity

2. Multi-Agent Cascade

Risk emerges across the chain, not in any single step

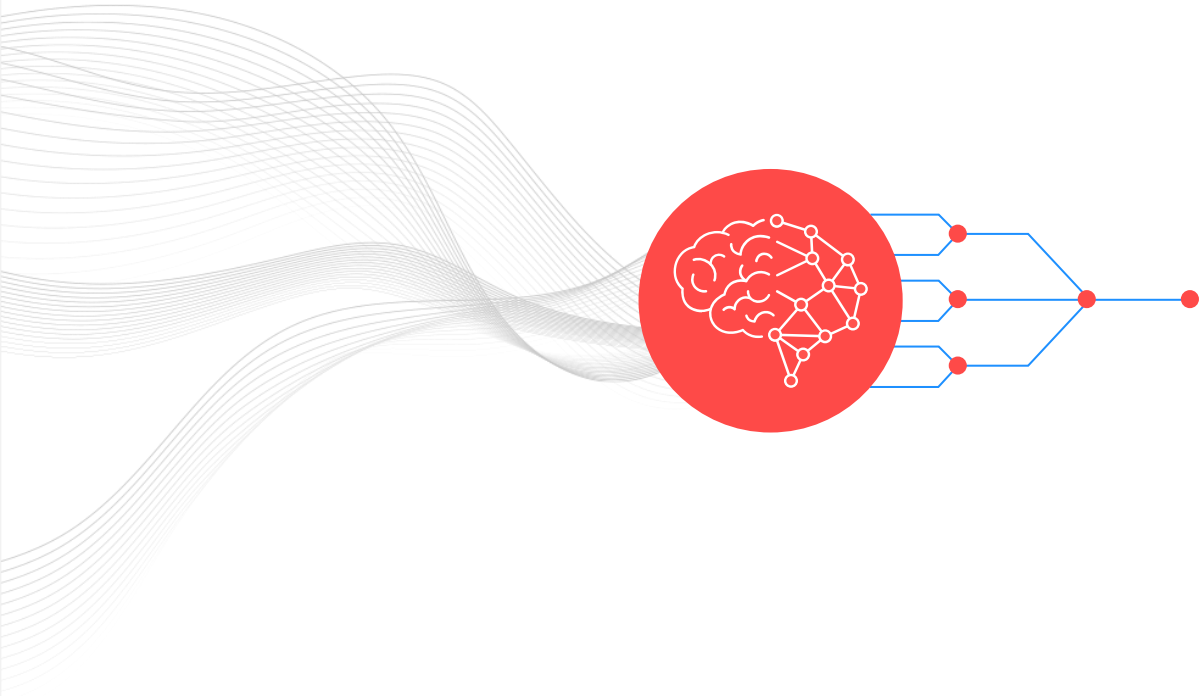
3. Counterfeit Agent

(Unauthorized Agent Impersonation)



Identity alone isn't enough — provenance exposes the fake

COGILITY



Addressing Agentic AI Insider Risk

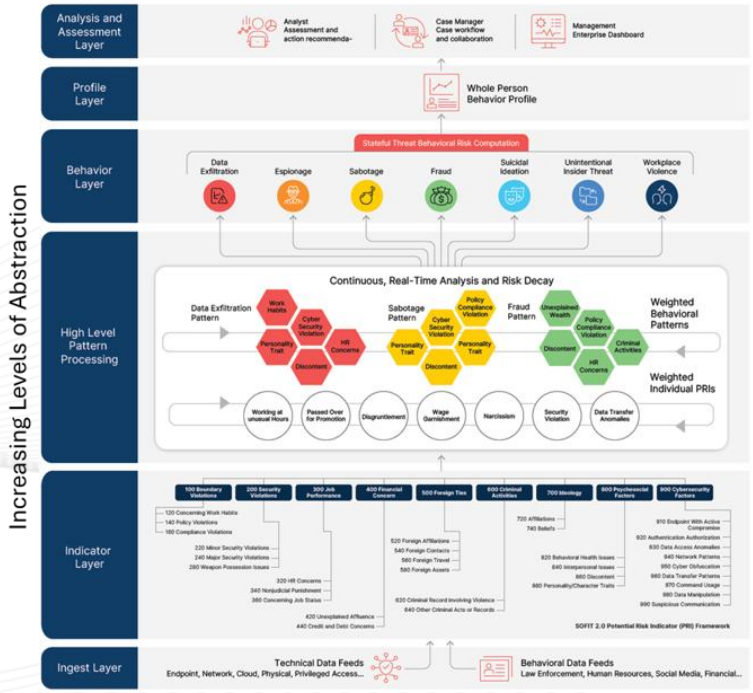
Core Capabilities for Agentic AI Risk Management

Agent entity model	Treat agents as first-class tracked entities: agent ID, owner, sponsoring user/team, purpose, tools, permissions, model, runtime, data domains, credentials, deployment environment, and mission role.
Agent event ingestion	Ingest platform audit logs, UAM data, identity events, API calls, tool-call traces, MCP activity, DLP alerts, ticket/case data, endpoint activity, and cloud logs.
Human-agent linkage	Associate agent behavior with the human, team, service account, or workflow that created, delegated to, approved, modified, or benefited from the agent.
Behavioral indicators	Document normal patterns by agent, human-agent pair, mission function, tool class, data domain, and time window. Existing and new behavioral indicator taxonomies will inform a systemic behavioral analytic model: the SOFIT human and organizational insider risk indicator taxonomy can serve as a partial knowledge base that addresses human and system indicators; to complete the human-agent-system triad, an agentic AI risk indicator taxonomy must be developed.
Risk pattern analytics	Detect patterns such as privilege expansion, unusual tool chains, sensitive-data access outside mission need, repeated blocked actions, nighttime/off-cycle activity, anomalous delegation, agent use after employee status change, use of agents to bypass UAM controls, or coordination across multiple agents.
Investigative timeline	Produce a defensible narrative: what happened, in what order, by which actor, under whose authority, with which data, through which tools, and with what risk Indicators.
Policy-context integration	Do not just say “anomaly.” Map events to agency policy: acceptable use, data handling rules, delegated authority, separation of duties, export controls, privacy rules, and mission constraints.

Approach

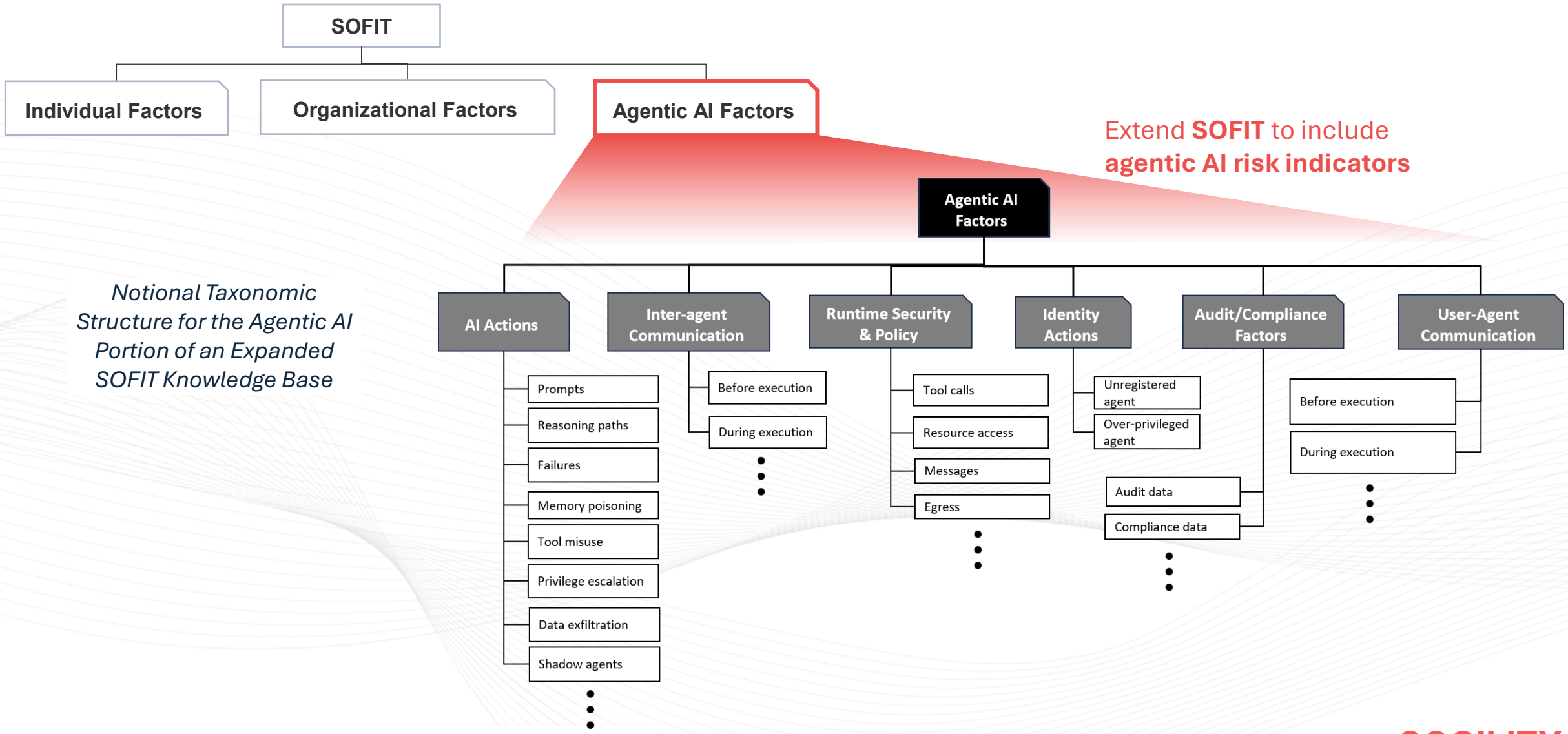
Cogynt.ai is the behavioral intelligence aggregation point for all available data sources

- Candidate data feeds include agent runtime logs, tool-call traces, identity and entitlement data, model/platform audit logs, cloud/API logs, UAM telemetry, DLP events, endpoint activity, case-management records, policy/mission context.
- Apply all tracking patterns (based on **SOFIT** or other framework) plus new patterns developed for agentic AI context
- Leverage Hierarchical Complex Event Processing (**HCEP**) to enforce behavioral constraints at machine speed
- The analysis produces persistent **agent entity profiles** linked to human sponsors, teams, service accounts, workflows, permissions, and organizational policies
- Resultant risk indicators, timelines, and provenance supports analyst review, case management, SIEM integration, investigative intelligence, and escalation or downstream enforcement workflows.



COGILITY

Extending SOFIT



Notional Taxonomic Structure for the Agentic AI Portion of an Expanded SOFIT Knowledge Base

COGILITY

Operations Concept

Investigative Intelligence Sequence

1. Map incoming log streams into stateful, in-memory representations
2. HCEP continuously generates composite events, linking each human actor's baseline risk profile to the agent's structural data access
3. When a composite pattern's risk weight crosses the threshold, the HCEP engine issues an alert

Risk posture shifts: Observer → Interdictor

triggers

Remediation Controls

State Isolation
Locks the active session

Control Execution
Trips the "circuit breaker"

Credential Revocation
Pauses the agent's execution queue

Real-Time Notification
Alerts a human analyst (human-on-the-loop)

COGILITY

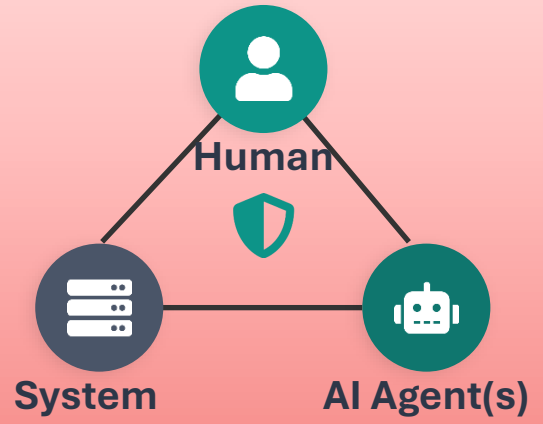
Remediation Levels

Tier 1 Low	**Low anomaly / logic drift**	→	In-context constraint reinforcement
Tier 2 Elevated	**Suspicious activity**	→	Human-in-the-loop validation gate — hold the execution queue for manual verification
Tier 3 Critical	**Critical system risk / active exploitation**	→	Instant programmatic circuit break — terminate agent session, lock user access, ...

Closing Perspective

AI agents are becoming a new class of enterprise actor.

Human-Agent-System Triad



Whole-person insight today → whole-enterprise risk intelligence tomorrow.

The strategic advantage belongs to organizations that can govern them with three disciplines:

- **Accountability:** Trace every action to its human sponsor and agent.
- **Context:** Fuse signals across the human–agent–system triad.
- **Confidence:** Deterministic, explainable, audit-ready risk.

COGILITY

Thank You

COGILITY SOFTWARE
www.cogility.com

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

Collaborators Sought

We've shared how the Cogynt.ai platform and our deep experience in whole-person IRM can help address this emerging challenge.
We seek feedback and collaboration from data partners, policy experts, academic researchers, industry analysts, LLM providers, AI researchers, IAM vendors and consultants.

Go deeper

Read Whitepaper

Download Presentation