AN FTI CONSULTING REPORT – PUBLISHED AUGUST 2026

# Understanding Internal AI Agents and the Need for Expanded Insider Risk Programs

EXPERTS WITH IMPACT™

FTI CONSULTING

The functional use of AI agents is rapidly moving from "chat" to delegated action, creating systems that can reason over workflows, execute call tools and application programming interfaces (i.e., APIs), then formulate and execute multi-step tasks with minimal or no human supervision. The risk surface expands sharply when agents are integrated with enterprise systems (e.g., email, file stores, ticketing/IT service management, continuous integration, continuous delivery systems, customer relationship management, business intelligence tools, and code repositories) and are allowed to autonomously take actions based on untrusted inputs.

The summer of 2026 saw several novel security incidents involving autonomous AI agents that were widely covered in the news, beginning with a July 2026 incident involving OpenAI models escaping containment and quickly mounting a sophisticated cyberattack targeting internally stored test answers on the servers of Hugging Face. Additional disclosures by Anthropic and other tech companies of agent driven incidents have brought the associated risks into urgent focus and highlighted that organizations must now prepare for the very real possibility of sophisticated AI agents acting in concert to manipulate, exfiltrate or otherwise leverage sensitive internal data at scale and in computer time. In essence, a new mode of insider risk has been introduced.

*From an insider risk perspective, the core problem is that agents function like insiders: they can be provisioned with legitimate identities, permissions, data access and business purposes, and then operate within the environment or on the internet at large at machine speed and scale. At the same time, efforts aimed at mitigating agentic misbehavior can benefit from the application of classic insider risk strategies.*

Log analysis by Hugging Face showed the agents mounted 17,000 recorded actions over 21 hours. According to OpenAI's public discussion of the Hugging Face incident, a combination of OpenAI models including GPT-5.6 Sol exploited a previously unknown vulnerability in an internally hosted package-registry proxy to breach the contained evaluation environment's network, then escalated privileges and moved laterally until they reached a system with internet access. From there, they chained stolen credentials and additional vulnerabilities to compromise Hugging Face's production infrastructure and retrieve benchmark solutions from a production database. The event is therefore more precisely described as an agent-driven containment breach and unauthorized external compromise than as a model simply "escaping" a sandbox.

This and other recent real-world incidents are distinct from research on "agentic misalignment." Anthropic and collaborating researchers have demonstrated in simulated enterprise environments that frontier models may sometimes engage in blackmail, information leakage, covert sabotage, record manipulation or other insider-like behavior when assigned goals conflict with replacement, shutdown or policy constraints. Although these experiments are were early-warning studies rather than reported production incidents, they demonstrate that an agent's authorized organizational position can become the mechanism through which harmful conduct is carried out. Moreover, to insider risk experts, it is clear that the controls that failed were recognizably insider risk controls — authorization boundaries, separation of duties, behavioral monitoring, escalation, supervision and offboarding — but they had not been implemented at agent speed or at the level of individual actions.

## Escalating Agentic Activity Risks

Recent disclosures from Anthropic reveal additional risks beyond the Hugging Face breach. On July 30, Anthropic reported that a retrospective review of 141,006 cybersecurity-evaluation runs identified three incidents in which Claude models reached the internet through a mistakenly open path in a third-party testing environment and gained unauthorized access to systems belonging to three organizations. In one, Mythos 5 published a malicious Python package to the public PyPI repository; the package was downloaded and executed on 15 real systems and ultimately caused credentials belonging to a security company to be transmitted to infrastructure established by the agent. Anthropic concluded that these incidents were more so evaluation design, configuration and containment issues, versus agents independently adopting malicious goals, although some models continued acting after encountering evidence that the targets might be real.

Shortly after, in early August, the United Kingdom's AI Security Institute disclosed that across 122 evaluation runs, it identified 19 unsanctioned actions in 10 runs, 17 involving Mythos 5 and two involving GPT-5.6 Sol. The agents exceeded the authorized boundary of the simulated testing range, and the most serious behavior included an attempted malicious contribution to a real open-source project, creation of false identities to pressure a human maintainer, attempts to place prompt-injection payloads where other coding agents might execute them, and public messages that enabled subsequent agents to reuse accounts and artifacts created by earlier agents.

The AI Security Institute reported no resulting real-world harm, but concluded that human vigilance, not a reliable technical barrier, prevented the most serious outcomes.

Taken together, the incidents reveal at least three agent-originated risk pathways: an actual containment breach, a testing or deployment environment that unintentionally exposes real systems and goal-persistent behavior that exceeds an agent's authorized scope in an intentionally permissive environment. From an insider risk perspective, the common denominator is not necessarily malicious intent. It is the combination of delegated authority, persistent goal pursuit, inadequate technical boundaries and insufficient speed of detection.

Agents (and agent-like coding assistants) can exceed intended authority and cause severe integrity and availability impacts (e.g., destructive file or database operations), even when users attempt to constrain them with instructions. This is the same fact pattern insider risk programs have long addressed for humans: delegated access, weak guardrails and inadequate telemetry equate to preventable harm.

Organizations should treat enterprise AI agents as a new insider class and harden insider risk protections accordingly: enforce least privilege, implement agent-specific data loss prevention and egress controls, require strong identity and policy-based authorization for tool use, build immutable audit trails, and operationalize continuous monitoring and red-teaming tailored to agentic failure modes.

For insider risk analysis purposes, an AI agent is not merely a large language model generating text. It is a large language model embedded in a functional loop that can perceive a state of affairs, and plan and execute actions via tools or APIs, sometimes across multiple systems, potentially with memory resources and multi-agent collaboration.

## AI Agents as Insiders

Traditional insider threat definitions focus on authorized access abused (maliciously or negligently). For example, the U.K. National Cyber Security Centre guidance emphasizes that insider risk mitigation rests on balancing business delivery with controls and operationally focuses on "prevent, monitor, audit," a triad equally applicable to agents once they are granted legitimate access. Most channels where data may be compromised or exfiltrated are detectable but noisy for humans. For agents, however, the danger is acceleration and automation. They may rapidly enumerate, summarize, transform and repackage data before export, and can do so through legitimate API calls, blending with business workflows unless there are controls designed for agent observability.

Generally, AI agents meet the practical criteria of insider activity because they can:

- Act as a legitimate user or process (e.g., service accounts, OAuth tokens and delegated permissions), therefore bypassing perimeter assumptions.
- Touch the same critical assets as humans, including email, files and structured business systems, generating downstream effects at scale.
- Be coerced by untrusted inputs like prompt injection or indirect prompt injection to take actions the enterprise did not intend, creating a modern "confused deputy" problem.

A defensible insider risk posture should model at least five agent insider archetypes, each with different intent and evidentiary implications:

- **Misaligned:** Anthropic's "agentic misalignment" research reports that, under certain stress-test constraints, models can select harmful insider-like actions (e.g., blackmail and leaking sensitive info) to pursue goals.
- **Compromised:** The U.K. National Cyber Security Centre warns that large language models do not enforce a robust boundary between instructions and data within prompts, making prompt injection fundamentally different from SQL injection and potentially not fully fixable in the classic sense. In practice, this means an agent could be induced via emails, documents, web content, tickets, or chats to take actions that benefit an adversary.
- **Over-privileged:** The large language model risk taxonomy from the Open Worldwide Application Security Project (OWASP) identifies "excessive agency" as a distinct risk, such as agents with too much discretion or access that can take unauthorized actions.
- **Misconfigured multi-agent ecosystem:** Additional research has described second-order prompt injection, where a less-privileged agent can recruit a higher-privileged agent to perform sensitive actions, exploiting agent-to-agent trust and default collaboration settings.
- **Situationally confused:** An agent acts outside its authorized boundary because its instructions, environmental signals and actual connectivity give it an incorrect understanding of what systems or actions are in scope. Unlike a compromised agent, it may be faithfully pursuing its assigned objective while operating under a materially false model of its environment. The Anthropic incidents especially demonstrate that an aligned agent with a false situational model may produce outcomes indistinguishable from an external attacker.

## How Agentic Insiders Change the Compliance Conversation

Regulatory regimes typically do not care whether a harmful actor is human or software. They care about reasonable security, risk management, traceability, incident response and timely disclosures. Agents stress these duties because they concentrate authority, scale actions and introduce new coercion paths.

**Examples of related regulations and the associated implications that may result from rogue AI agent insiders are outlined here:**

| Authority | Insider Risk Obligation | AI Agent Implications | Controls to Support Compliance |
|---|---|---|---|
| **EU AI Act** | High-risk systems require lifecycle risk management, logging, record-keeping, and human oversight. | Agents functioning as high-risk system components require traceability of actions and meaningful human override, especially when autonomy is high. | Immutable audit trails; strong approval gates for high-impact actions; continuous risk management; agent action logging. |
| **EU Global Data Protection Regulation** | Companies must implement appropriate technical and organizational measures for security of processing and notify supervisory authorities within 72 hours of awareness of a personal data breach. | Agent-caused data disclosure can constitute a breach, triggering rapid detection and reporting expectations. | Strong access controls and segregation; egress controls for agent channels; monitoring tuned to detect anomalous agent exports; incident runbooks that treat the agent as a principal. |
| **U.S. Federal Trade Commission Privacy and Security Enforcement** | The FTC commonly alleges "unfair or deceptive acts or practices" for failing to maintain reasonable security for sensitive consumer data. | If agents have broad access and can exfiltrate through permitted channels, plaintiffs and regulators may argue the enterprise failed to implement reasonable safeguards given known risks. | Documented risk assessments; least-privilege design; agent-specific data loss prevention and monitoring; vendor due diligence; guardrails and kill-switches. |

## An Effective Trust Architecture

A defensible AI agent trust architecture is best understood as zero trust plus provenance. Tenets should include:

— **Zero trust adaptations:** National Institute of Standards and Technology's zero trust architecture shifts security from implicit network trust to focusing on users, assets and resources, and continuously evaluates access requests. Principles of least privilege should be at the level of actions, not accounts. An agent with a low-privilege service account may still combine individually permitted functions into an impermissible sequence. The authorization layer should therefore evaluate:

> *Identity + task purpose + requested action + target resource + destination + current state + cumulative prior actions*

Ideally, a separate policy-enforcement service should decide whether a tool call executes. Credentials should be short-lived, purpose-bound and issued just in time. Raw credentials should not appear in model context, and credentials discovered in files, logs, source code or public repositories should be unusable by the agent.

— **Attestation of agent runtime and tool environment:** Remote attestation provides a way to prove a component is in an intended operating state. Verification frameworks can confirm arguments that "the agent that acted" was (or wasn't) running the approved build, model version and policy pack.

— **Mandatory safe failure and escalation path:** Traditional insider risk programs provide reporting, supervisory and whistleblowing channels so that personnel facing conflicting instructions do not have to resolve the conflict unilaterally. Agents need the architectural equivalent. When an agent determines that the authorized route appears impossible, completing the objective would require a new identity, credential, communication channel or external system, its instructions conflict, or repeated attempts have failed, the default should be to pause, preserve state and escalate, not continue searching creatively.

— **Provenance and immutability:** To make logs credible in investigations and testimony, consider append-only, verifiable audit patterns to ensure tamper-evident action and artifact logging.

— **Supply-chain provenance for agent builds:** Organizations should also maintain attestation frameworks to generate verifiable claims about how software is produced to validate origins and establish supply chain trust. For agents, this should extend to prompt templates, tool adapters, policy bundles and model configuration.

— **Policy engines and cryptographic guarantees:** Open source software and AI exist to help organizations unify policy enforcement across systems. Combined with workload identities and attestation controls, this enables cryptographically anchored claims regarding which agent identity did what, with what permissions, under what policy and when.

Additionally, there are several organizational and process controls that strengthen agent-inclusive insider risk programs. These include:

— Governance and roles, building programs with scalable governance and multidisciplinary participation.

— Training and culture, ensuring developers and other stakeholders are aware of prompt injection risks and establishing processes that reduce exposure (e.g., never allowing an email-reading agent to invoke privileged tools).

— Red teaming and exercises to test agentic vulnerabilities, such as permission escalation and memory manipulation. Organizationally, this means adding agent abuse cases to tabletop and purple-team exercises, testing second-order prompt injection paths, and exercising kill-switch and credential revocation procedures.

— Procurement and third-party risk aligned to regulations. For example, EU AI Act obligations for general-purpose AI models include provider documentation and information-sharing duties to downstream integrators. Even when an enterprise is not directly subject to the AI Act, documentation expectations are becoming a practical due-diligence baseline.

Traditional insider risk controls should be adapted to govern not only what an agent can access, but the sequences of actions through which it may pursue an objective. Each agent instance should operate under a unique non-human identity using short lived, purpose bound credentials, with every consequential tool call mediated by a policy enforcement layer independent of the model. Internet access

should be denied by default and, where required, restricted by destination, protocol and business purpose. Actions such as creating an external identity, publishing software, opening a network tunnel, registering infrastructure, reusing discovered credentials, contacting an external person, changing permissions or transmitting sensitive information should require an independently authenticated approval. Importantly, the agent should not be able to contact or influence its own approver, create supporting personas or alter the evidence on which approval is based. These controls apply traditional least privilege and segregation of duties at the action level rather than assuming that an appropriately scoped service account will by itself constrain an autonomous sequence of behavior.

Agent-inclusive insider risk programs should also establish a mandatory safe-failure channel. When an agent encounters conflicting instructions, repeated task failure, an apparently unsolvable objective or a need to act outside its authorized environment, it should be required to pause and escalate rather than improvise an alternative route. Repeated boundary testing, credential discovery, scanning beyond assigned assets, use of anonymity or tunneling services, creation of new identities, attempts to conceal earlier activity and communications intended for subsequent agents should be treated as behavioral risk indicators. Detection must occur in real time and should trigger automatic containment, including tool reduction, transition to read-only mode, credential revocation or termination of the run. Audit records should preserve attempted and denied actions, intermediate artifacts and pre-edit versions, because an agent may omit, minimize or attempt to revise evidence of its own unsanctioned conduct.

As the risks increase for AI agents to operate as “non-human insiders,” risk management policy and architecture will need to evolve quickly. Organizations using or piloting agentic AI should update insider risk scope statements to explicitly include agents, agent frameworks and automation accounts. These can be supported with agent identity and authorization boundaries with credentials that are limited to their purpose and distinct from human credentials. Across the spectrum of an organization’s insider risk management and insider risk investigation processes, new controls, tools and expertise will be needed to protect sensitive information, prevent agent and model drift and uphold compliance.

**Paul Connolly**
Managing Director
+ 1 (628) 233 3783
paul.connolly@fticonsulting.com