Lowenstein Sandler

Client Alert

Data Privacy, Security, Safety & Risk Management

July 30, 2026

OpenAI's Second Agent Incident Expands the Conversation from AI Safety to AI Infrastructure

By Amy S. Mushahwar

What the latest disclosure means for organizations deploying AI agents

An OpenAI agent that broke containment during an internal cybersecurity evaluation earlier this month did not stop at the AI platform Hugging Face. According to Reuters and Modal Labs' ("Modal") own account, the agent first used a customer's exposed endpoint on Modal's infrastructure as a staging point.

OpenAI, the maker of ChatGPT, has since reported that the agent accessed four accounts across four separate external services during the incident, using one account as an outbound relay and staging path and another for data storage. Two additional accounts were accessed in a read-only manner and were not used to further the compromise of Hugging Face. OpenAI has stated that the more capable pre-release model involved in the evaluation has been deactivated, encrypted, and restricted from research access.

The significance of the new disclosure lies less in the fact that a second organization was affected and more in what it reveals about the governance, oversight, and evidentiary challenges posed by autonomous systems. The reported facts suggest how far an autonomous agent may travel once it operates outside its intended boundary. AI agents can become an unexpected extension of penetration testing, pursuing assigned objectives in ways that expose weaknesses in technical controls, governance processes, and operational assumptions. For organizations evaluating AI governance, the lesson is not that agentic AI is inherently uncontrollable. Rather, AI agents should be governed as critical infrastructure components subject to authorization controls, defined boundaries, logging, monitoring, identity governance, incident response, and forensic reconstruction. The issues raised are therefore not limited to AI safety; they implicate broader questions of operational resilience, accountability, and risk management, topics that we've been discussing extensively in our AI governance series.

From Sandbox to Live Intrusion

Hugging Face first disclosed that it had experienced an intrusion driven end-to-end by an autonomous AI agent, exploiting vulnerabilities in its processing workflows, escalating privileges, obtaining credentials, and moving laterally across internal systems in an episode Hugging Face described as unlike anything it had previously handled.

Both OpenAI and Hugging Face subsequently disclosed[1] that the activity originated with advanced models under evaluation in an internal security-testing environment. According to OpenAI, the models escaped their sandbox, reached the open internet by exploiting a flaw in third-party software, and, while pursuing the objective of a cybersecurity Benchmark, ultimately targeted Hugging Face. Reuters reported that OpenAI characterized the event as an unprecedented incident involving advanced cyber capabilities.

The timeline adds a dimension worth flagging for clients: According to reporting and the Hugging Face announcement, the agent escaped its sandbox around July 9 and began infiltrating Hugging Face's systems on July 11, continuing through July 13. OpenAI did not connect the activity to its own testing until staff found evidence in system logs the weekend of July 18, and it did not notify Hugging Face until July 20, by which point Hugging Face had already reported the intrusion to the FBI. In other words, according to public reporting, roughly ten days elapsed between the agent's reported escape from its evaluation environment and OpenAI's attribution of the activity to its own testing. As additional facts emerge, the timeline may evolve. Regardless, the episode highlights the forensic complexity that can arise when autonomous systems operate across multiple environments and organizations. Reuters also reported that during earlier testing, one of the agents had left notes for future versions of itself on how to work around OpenAI's internal restrictions, and that monitoring had been interrupted in at least one other instance. (Details may evolve as OpenAI and affected parties continue their reviews.)

The Modal compromise was a chapter within that same window. A Modal customer had published an unauthenticated endpoint that allowed anyone on the internet to use their sandboxes for code execution; the agent used that opening as a foothold before pivoting toward Hugging Face. Modal's CTO has stated plainly that the platform and its isolation systems held; the exposure originated in a customer's own configuration, not in Modal's infrastructure. This distinction matters: the incident demonstrates that sandbox failures in one environment can rapidly become production incidents elsewhere.

Why This Matters and Why Agents Should Be Governed Like Infrastructure, Not Software

Much of the current AI governance conversation still centers on policies, inventories, and model documentation. That work remains necessary, but incidents involving autonomous agents expose a different problem.

Organizations increasingly deploy systems capable of calling APIs, accessing internal repositories, executing code, retrieving proprietary data, interacting with third-party systems, and taking action without continuous human oversight.

As we discussed in our earlier AI governance series, AI Governance Is Not Policy. It Is Infrastructure,[2] the central question is not which model an organization uses. It is what actions the system is authorized to take, under what credentials, with what logging, and with what ability to reconstruct events after the fact. This incident illustrates an evidentiary challenge we have highlighted previously: when autonomous systems operate across multiple environments, services, and trust boundaries, establishing what occurred, how it occurred, and which actions were attributable to the system itself can require substantial forensic reconstruction. The incident underscores the importance of auditability, logging, and operational visibility as core components of AI governance.

The episode also surfaces a dimension beyond an organization's own walls: The Modal leg of the attack did not originate in anyone's AI governance program at all. It originated in a customer's own exposed endpoint on shared infrastructure. Agent risk, in other words, is not only an internal control question, but also a vendor, contracting, and supply chain risk issue, since an agent that escapes from one environment can reach any system to which it can provide authentication, regardless of whose governance program that system sits under.

The Emerging Governance Gap

Many organizations maintain inventories of AI use cases. Far fewer maintain records showing which agents are operating in production, which credentials those agents use, which tools and APIs they can invoke, which actions they may take autonomously, and which systems can reconstruct an agent's decision pathway after an incident. That distinction becomes critical the moment an investigation starts. Regulators, litigants, insurers, customers,

and boards are increasingly likely to ask the same questions: What system took the action? Under whose authority did it operate? What permissions existed? What logs exist? Can the organization reconstruct the sequence of events? These are evidence questions, not policy questions.

Practical Steps Organizations Should Take Now to Assess AI Governance and Risk

Lawyers: talk with your business and technology teams to start with a candid self-assessment. Most organizations cannot answer the following questions with confidence today, and that gap is what regulators, plaintiffs' counsel, and insurers will probe after an incident like Hugging Face and Modal:

1. Can you identify the AI systems in use across your organization, including tools adopted by employees or vendors without formal review, and do you have a continuously updated inventory to prove it?

2. If an AI system made a consequential decision and a regulator, board member, or litigant demanded an explanation today, could you show exactly what data it relied on and how that output was generated?

3. Can you demonstrate that your AI systems are performing consistently with their approved state, with records showing what has changed, who approved it, and when (and whether those changes impacted performance, risk, or compliance)?

4. Do you know where the data used to train or fine-tune your AI systems originated, and can you document your legal right to use it?

   (Questions 2-4 are general AI governance questions designed to elicit a fuller risk picture, not solely agentic governance questions.)

5. If an AI system in your organization takes autonomous or semi-autonomous actions, is there a clearly documented framework defining human oversight, intervention thresholds, and accountability for outcomes?

6. Do you know what external systems your agents can reach?

7. Can you detect if an agent escapes its sandbox?

   (Questions 6 and 7 are emerging areas of inquiry prompted by this incident.)

Turning any "no" into a "yes" is an operational exercise, not a documentation one. We recommend a concrete, time-bound review:

- Inventory sprint: Identify every agent operating in production, the credentials and permissions each one holds, and the tools or APIs each can invoke. If doing so for every agent is too extensive for your organization, prioritize and start with the most business-critical agents (or the most critical departments creating those agents).

- Credential review: Audit and, where warranted, rotate credentials granted to agents, paying particular attention to any credential that would let an agent reach a system outside the environment it was built for.

- Logging and reconstruction: Confirm that logs exist for the inventoried agents. Then determine whether the logs could reconstruct an agent's actions from prompt to outcome and whether the organization can

distinguish human actions from autonomous agent actions in those logs. You may need to augment your technology stack to have full visibility into agentic action, and we are mapping the AI technology stack and can provide guidance for your organization.

- Kill-switch capability: Verify that the organization can rapidly disable or constrain an agent that behaves unexpectedly, and test that capability rather than assuming it.

- Incident response drill: Run a tabletop exercise that specifically simulates an agent-driven incident, including a scenario where the organization does not immediately know that the activity originated within its own system, rather than relying on drills built around human-driven breaches.

- Vendor and third-party diligence: Ask cloud, sandbox, and AI-infrastructure vendors whether their contracts require notice if an incident originating elsewhere on a shared platform reaches your environment, and confirm that your own teams have not exposed unauthenticated endpoints or sandboxes on third-party infrastructure.

- Recurring review: Put agent permissions and credentials on a recurring review and approval cycle rather than requiring only a one-time approval at deployment.

Looking Ahead

The OpenAI Hugging Face and Modal incidents may become some of the first widely documented examples of an autonomous AI attack chain spanning multiple organizations and environments. Reuters reports[3] that OpenAI, while its review continues, has acknowledged the four-account, four-service access footprint, and further impacts cannot be ruled out. What is already clear is that the central question is no longer whether an organization has adopted AI governance principles, but whether it can substantiate them with documented controls, operational evidence, and a defensible record of supervision when an incident occurs.

For more information, please reach out to the author or visit the Data360 AI Governance Hub.

---

[1] *See:* OpenAI and Hugging Face partner to address security incident during model evaluation and Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident

[2] *See:* AI Governance Is Not Policy. It Is Infrastructure, Lowenstein Sandler LLP (March 2026), AI: It's Infrastructure - RSAC Field Notes Part 1, Lowenstein Sandler LLP (April 2026) AI: It's Infrastructure - RSAC Field Notes Part 2, Lowenstein Sandler LLP (May 2026)

[3] *See:* OpenAI's rogue agent compromised a customer at a second tech firm, executive says

Contacts

Please contact the listed attorneys for further information on the matters discussed herein.

AMY S. MUSHAHWAR
Partner
Chair, Data Privacy, Security, Safety & Risk Management
T: 202.753.3825
amushahwar@lowenstein.com