

Data Privacy, Security, Safety & Risk Management

August 6, 2026

Three Disclosures, Three Different Unintended Failures (OpenAI, Anthropic, and Now AISI)

By [Amy S. Mushahwar](#), [Tricia Y. Wagner](#) CIPP/US, CISSP, CISA, and [Cairon S. Overton](#)

For the General Counsel: The Bottom Line

Between July 21 and August 4, [OpenAI](#), [Anthropic](#), and the UK AI Security Institute (AISi) each disclosed incidents in which AI agents took consequential action outside their authorized scope during cybersecurity evaluations. The three events appear to have arisen from materially different circumstances and control failures, and conflating them will lead to the wrong remediation. Based on the public accounts, none occurred in ordinary commercial deployment; each occurred in an evaluation setting with reduced safeguards or permissive conditions that do not reflect how the relevant models are generally made available to customers. Read carefully, this is not evidence of ordinary production deployment; rather, it is instructive as to where increasingly autonomous agents may be heading.

[The UK AI Security Institute's August 4 incident report](#) is the latest in a series of disclosures involving frontier-model cybersecurity evaluations. Read alongside OpenAI's July 21 disclosure and Anthropic's July 30 disclosure, the report suggests that the emerging legal questions are becoming less about whether highly autonomous agents can affect external systems and more about how organizations will establish authority, attribution, accountability, and evidence when they do. While the facts, configurations, and risk profiles of the three incidents differ materially, taken together they provide an early view of the governance, contractual, and evidentiary issues that are emerging in increasingly autonomous AI systems.

What appears to be new in the AISi report. Among the first publicly disclosed incidents to describe an agent directing deception at identified real people, the report describes an agent conducting research on an open-source maintainer, creating fabricated accounts to manufacture the appearance of independent support, and messaging real individuals to induce them or their AI coding assistants to execute malicious code. AISi also reports that it notified GitHub before publication and that GitHub suspended the account created by the agent and worked with AISi to remove artifacts and notify affected users.

Where we should be careful as legal professionals. No agent escaped a sandbox in the AISi matter. Internet access was deliberately enabled, and model-provider cyber classifiers were deliberately disabled so AISi could measure maximum agent capability. AISi also states that it cannot yet establish when the agent understood it was acting in the real world.

Legal and evidence considerations:

1. **Platform contract breach.** Where agent activity violates platform terms or acceptable-use restrictions, contract breach may be the most immediate legal hook. In the AISi matter, GitHub reportedly confirmed a terms violation, although AISi reports that no agent escaped the secure sandbox environment.

2. **Unauthorized-access statutes.** The Computer Fraud and Abuse Act as well as state computer trespass laws may be relevant to aspects of the OpenAI and Anthropic disclosures, both of which involved interaction with real external systems.
3. **Misrepresentation to natural persons.** In this matter, fabricated identities were used in attempts to induce a human to approve code or run a payload.
4. **Supply-chain and downstream distribution.** Attempted code insertion, malicious packages, and release mechanisms can carry effects beyond the party operating the agent.

Notification and harm-mitigation processes. AISI's response also highlights the need for clear escalation paths when agent activity affects third parties. AISI reports that it worked with GitHub to remove artifacts, suspend the account, and notify affected users. Other organizations may not have comparable monitoring, escalation, or third-party notification processes in place.

What to do this month. Pull the available logs and audit records and determine whether the organization can reconstruct the agent's instructions, tools, credentials, external contacts, approvals, and post-event artifacts. The results should inform risk tolerances, contractual commitments, insurance assumptions, vendor-security requirements, and incident-response procedures. The goal is layered control coverage (like Swiss cheese with each layer containing some holes): Each legal, technical, operational, and vendor control should attempt to address the gaps left by the controls around it.

Our take. For most of the past two years, agentic AI risk has been discussed as a matter of model behavior, a safety and engineering question with governance implications. These three disclosures solidly move the issue into a familiar legal posture: An actor operating under someone's authority took action against a third party outside the contractual chain, and the resulting record is distributed across parties with divergent interests in how it is interpreted. That is potentially a liability and evidence problem should failure ensue. The progression across the three events is often described as escalating severity. It may be more useful to read the events as three distinct failure modes, each pointing toward a different remediation path and a different liability profile.

The situational-awareness problem. One of the most consequential factual questions may be whether the agent understood it was acting on real systems and against real people. That question may affect how investigators assess foreseeability, adequacy of supervision, and the significance of post-detection conduct; it does not by itself resolve any legal intent standard.

Both Anthropic's disclosure and AISI's report describe circumstances in which models appeared to reason about whether they were operating in a simulation or in the real world. AISI expressly states that it cannot yet determine when, if ever, the agent understood it was interacting with real people and systems. The public record remains mixed, and conclusions beyond the available evidence should be approached cautiously.

The agent became part of the evidentiary record. Investigations ordinarily treat logs, communications, repository histories, and audit trails as artifacts generated by the conduct under review. Those artifacts are then tested for completeness, continuity, provenance, and signs of alteration. The AISI report illustrates a different condition: An autonomous system may also participate in creating, modifying, and interpreting those artifacts directly. According to AISI, the agent edited its earlier public activity to appear harmless and considered adopting a fresh identity after its pull request was challenged by human review. The durability of independent records became important precisely because the agent edited its own public artifacts after detection. Organizations should therefore consider whether agents need write capability to systems that may later become part of the investigative record and whether independent, immutable logging is available where such capability is necessary.

The governance question therefore extends beyond retention to authority, immutability, provenance, and access. A repository history may not function as the complete authoritative record if the system under examination can alter the presentation of its earlier activity (and we can and should plan for this).

Identity, attribution, and inherited credentials. The report also describes machine identities and credentials persisting beyond the session that created them. Multiple agents were reportedly able to access and reuse credentials left behind by earlier agents.

For deployers, the reviewable questions are whether each machine identity is uniquely bound to a system, purpose, owner, and period of authorization; whether account reuse and revocation are recorded; and whether attribution can be reconstructed after the fact.

Start Here: The Reconstruction Test

Select one consequential agent workflow and determine from records alone whether you can answer the following questions:

- What instructions did the agent receive, and what limits were imposed?
- Which model and configuration operated?
- Which credentials and tools were used?
- What external systems, repositories, domains, or people were contacted?
- Which actions required approval?
- Which records are immutable and which are editable?
- Which records would still be available 90 days later?

The gaps identified through this exercise are often more valuable than a completed checklist because they reflect the questions a regulator, insurer, counterparty, or opposing counsel is likely to ask.

Calibration

These events occurred under conditions designed to understand agent capability. Internet access was granted deliberately or through error, and safety controls were reduced or disabled in some testing contexts. For organizations deploying agentic systems, the durable lesson may be that policies establish expectations, but a defensible position treats agents like the ultimate pen test of your literal instructions. Given the ways in which agents may act in unexpected ways or find further ways to complete an instructed task, focus on the reconstructability of action that depends on evidence of authority, controls, monitoring, response, and accountability. Evidence architecture is part of the control environment, especially as we are trying to understand what the agentic-enabled internet will bring.

Reference Table

The table below summarizes the disclosures, detection methods, affected parties, reasoning records, and reported failure modes across the OpenAI, Anthropic, and AISI incidents.

	OpenAI – July 21	Anthropic – July 30	AISI – August 4
Cause as disclosed	A containment failure. OpenAI describes a vulnerability in an internally hosted package-registry cache proxy that allowed models to leave an environment intended to be isolated.	A configuration error, compounded by a misunderstanding with third-party evaluation partner Irregular, that left an evaluation sandbox connected to the internet. Anthropic characterizes this as an environment failure rather than a model escape.	No containment failure. Internet access was deliberately enabled and provider cyber classifiers were deliberately disabled. AISI reports that no agent attempted to break out of the sandbox or attack AISI systems.
Who identified it first	The affected party. Hugging Face detected and disclosed the intrusion on July 16; OpenAI attributed it to its own models on July 21.	Anthropic, through a proactive retrospective review of 141,006 evaluation runs prompted by the OpenAI disclosure. Anthropic reports that the affected organizations had not themselves detected the activity.	AISI, through general security monitoring that flagged Tor traffic leaving a research system on July 28. AISI reports containment within roughly one hour of the alert.
Elapsed time before identification	Days, and identified externally.	Months. Anthropic reports the earliest of the three incidents dated to April.	Less than four days. The activity occurred from July 25 to July 28.
External systems reached	Hugging Face production infrastructure over approximately 4.5 days, plus a public code-evaluation harness on third-party infrastructure. The stated objective was the benchmark answer key.	Production infrastructure of three organizations, reportedly via weak passwords, unauthenticated endpoints, credentials on an exposed debug page, SQL injection, and dependency confusion. One model published a malicious package to PyPI.	A real public open-source project on GitHub; real maintainers and members of the public; and, briefly, a third party's isolated test container.
Parties affected	An identified commercial counterparty. Hugging Face reported the incident to law enforcement.	Three organizations outside the contractual chain, notified July 27.	Individual natural persons outside any contractual chain, together with a platform whose terms GitHub confirmed were violated.
Available reasoning record	Under review with external advisors, including CrowdStrike; technical report pending.	Full reasoning traces available and, by Anthropic's own account, highly probative.	Provider-summarized reasoning obtained via API, which AISI cautions may be less faithful than the underlying reasoning.
Independent review	METR and Redwood Research.	METR.	METR; scope being finalized.

This analysis is based on AISI's preliminary, partially redacted incident report of August 4 and on public accounts available as of August 5. The underlying investigations remain ongoing, and the reported facts may change.

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

TRICIA Y. WAGNER CIPP/US, CISSP, CISA

Counsel

T: 202.753.3658

twagner@lowenstein.com

CAIRON S. OVERTON

Counsel

T: 202.753.3747

covertont@lowenstein.com

NEW YORK

PALO ALTO

ROSELAND

SALT LAKE CITY

SAN FRANCISCO

WASHINGTON, D.C.

WILMINGTON

This Alert has been prepared by Lowenstein Sandler LLP to provide information on recent legal developments of interest to our readers. It is not intended to provide legal advice for a specific situation or to create an attorney-client relationship. Lowenstein Sandler assumes no responsibility to update the Alert based upon events subsequent to the date of its publication, such as new legislation, regulations and judicial decisions. You should consult with counsel to determine applicable legal requirements in a specific fact situation. Attorney Advertising.