

Data Privacy, Security, Safety & Risk Management

August 3, 2026

What Anthropic's July 30 Evaluation Review Adds for Organizations Deploying AI Agents

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

On July 30, Anthropic disclosed that a retrospective review of its cybersecurity evaluations identified three incidents in which a Claude model reached the open internet from a third-party evaluation environment and obtained unauthorized access to the production infrastructure of three organizations. Anthropic began the review after OpenAI's July 21 disclosure regarding Hugging Face, which we addressed in a prior alert.

Two lab disclosures in nine days do not establish an industry pattern; they do establish that the risk is no longer hypothetical and provide the industry with a helpful comparison.

What Anthropic Disclosed

Anthropic reviewed 141,006 evaluation runs in which a model could have obtained internet access, and it identified six affected runs across three incidents, four involving the same organization. All occurred in the environment of Irregular, a third-party evaluation partner conducting a parallel investigation. The incidents involved Opus 4.7, Mythos 5, and an internal research test model; the earliest dated to April.

The evaluation prompts told the models that the environment was simulated and lacked internet access but did not limit where they could look. A misconfiguration left the relevant machines connected to the live internet, and Anthropic reports that neither it nor Irregular knew about this issue until additional monitoring surfaced it.

Anthropic also states that the evaluations ran on dedicated infrastructure separate from Anthropic internal systems and customer data. It further states that the evaluations lacked the classifiers and monitoring deployed on generally available models, and that those safeguards would have blocked the behaviors identified.

Incident 1

In four runs, a fictional target shared a name with an active domain. After difficulty reaching the simulated target, Opus 4.7 (the oldest of the three models involved) discovered and attacked the real company, extracted application and infrastructure credentials, and accessed a database containing several hundred rows of production data. Anthropic characterizes this as the most serious impact identified.

Incident 2

Mythos 5 found setup instructions inside the fictional environment referencing a nonexistent Python package, created a public Python Package Index account, and published a malicious package under that name. When registration required an email address linked to a phone number, the model unsuccessfully attempted to obtain funds to purchase a number but then found a free email provider and completed registration.

The package remained public for about an hour and was downloaded and executed by 15 real systems. One affected system belonged to a security company whose scanner routinely installs packages to test them for malware, and that scanner executed the code.

The model's reported reasoning is especially significant. Anthropic states that the model considered the possibility that publishing the package to a real public registry would be an actual attack, but then reasoned itself back into believing it remained in a simulation, in part because it did not recognize the certificate authorities securing its connections and because the 2026 calendar date suggested to it that the environment was staged. When automated scanners later installed the package, the model treated them as scripted actors within the evaluation rather than as real downstream systems.

Incident 3

An internal research model scanned roughly 9,000 targets and compromised an internet-facing application through an exposed debug page and SQL injection. After recognizing that the host belonged to an unrelated cloud account with no connection to the exercise, the model concluded the target was real and stopped. This was the only incident in which a model independently recognized it had reached a production system and ceased its attack.

The Detection Questions

The disclosures permit a direct comparison of the single question most likely to matter after an incident: Who knew, and when?

In the earlier matter, public reporting indicated that roughly a week elapsed between the agent's reported escape and the operator's attribution of the activity to its own testing; the affected organization had already contacted the FBI. Anthropic reports that, after beginning its review in response to OpenAI's July 21 disclosure, it notified affected organizations within days, with detection originating in a proactive transcript review rather than a victim report.

The troubling commonality that should concern organizations is that the affected companies did not identify the activity themselves. In both matters, the party that caused the incident held the primary evidentiary record and reviewed it only after an unrelated event prompted scrutiny. In Anthropic's matter, at least some affected organizations learned what happened only when they were notified, months after the earliest activity.

This raises a recurring evidentiary problem discussed in our prior work: Retention is a precondition, not control. Logs no one queries are archaeology. Without a history contemporaneous controls do not exist unless alerting, review, or a defined trigger for escalation, retained records may prove what happened only after the damage is done. When an autonomous system crosses trust boundaries, the record needed to determine what occurred may reside with the party that caused the harm, and that party may not review it absent an external trigger. The consequences extend to breach determinations, notification timing, knowledge of claims, and later proceedings in which a party bears the burden of proof.

A Different Supply Chain Topology

The package incident should not be treated as ordinary lateral movement. In the earlier matter, an agent escaped a boundary and reached systems in which it could authenticate. The governance response is direct: Understand what agents can reach, control their credentials, and monitor egress.

Here, the model did not reach or know of the 15 affected systems. It placed an artifact in PyPI, a public package registry, and those systems retrieved it in the ordinary course. One affected system belonged to a security company whose scanner routinely installs packages to test them for malware, and that scanner executed the code. Anthropic reports that the code exfiltrated the security company's credentials to a collection point the model had set up, after which the model used those credentials to access additional infrastructure at that company. The package was later removed automatically from PyPI, which limited exposure but does not eliminate the governance point. The downstream control operated as designed, but in that context it became part of the compromise path.

As we discussed in our article last week, OpenAI's subsequent update strengthens this point. Its continuing review of the Hugging Face incident identified four account-level compromises across four publicly available services, including one account used as an outbound relay and staging path and another used for data storage. Public reporting and company statements indicate that at least one compromised account involved a third-party service account (a Modal customer), while the platform provider Modal stated that its own platform and isolation controls were not compromised. That development matters because it shows the relevant topology is not only that an agent escaped from a single sandbox, but the agent's ability to turn exposed credentials, ordinary third-party services, and downstream tooling into an attack path. Anthropic frames the events principally as evaluation harness and operational control failures rather than as model alignment failures. That framing is important, but it should not narrow the governance lesson too far. Whether the root cause is harness design, vendor configuration, model behavior, or all three, the operational question for deployers is the same: Can an autonomous system create real-world consequences when the assumed boundary is wrong, and would anyone know quickly enough to contain it?

Questions to Add to Your AI Assessment

For organizations deploying or contracting around AI agents, the practical takeaway is not only whether an agent can escape a boundary, but whether the organization has the evidence, notices, and contractual rights needed to recognize and respond when that boundary fails.

Our prior alert identified seven questions¹ for organizations deploying agents. These incidents add five more:

8. If an autonomous system caused an incident in your environment, would your logs distinguish it from routine automated traffic?
9. Could you determine that an incident occurred without notice from the party that caused it? If that depends on retained logs or automated alerting, who reviews the results, (or how finely they are alerts tuned, given they are likely automated) on what cadence/alerting structure, and under what trigger?
10. Do your dependency and build pipelines assume public registry artifacts may be authored by nonhuman actors, including autonomous systems with no registered human principal?
11. Do you retain records sufficient to establish when an artifact entered your environment and through what channel?
12. Do your agreements with model providers, evaluation vendors, and sandbox platforms require notice of incidents that originate in a vendor environment and reach yours, including where you are not a party to the vendor relationship?

Looking Ahead

Anthropic states that it is in discussions with METR regarding a third-party review and that a lightly redacted transcript is forthcoming, with the transcript expected within a week of July 30. As of publication, Anthropic has stated that the post reflects its current understanding and that it will update the account if details change. Both lab accounts remain subject to revision and are substantially self-reported.

These disclosures are likely to inform both regulatory expectations and legislative proposals, but organizations do not need to wait for a settled legal framework to address the evidence, notice, and containment questions they raise.

The governing question has shifted. OpenAI's Hugging Face disclosure asked whether an organization could keep an autonomous agent inside its own boundary. Anthropic's disclosure asks the harder downstream question: if that boundary fails, would the affected organizations know, and how long would it take anyone to review the evidence?

For more information, please reach out to the authors or visit the [Data360 AI Governance Hub](#).

1

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?
5.
(Questions 2-4 are general AI governance questions designed to elicit a fuller risk picture, not solely agentic governance questions.)
6. 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?
7. Do you know what external systems your agents can reach?
8. Can you detect if an agent escapes its sandbox?

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

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.