United States Senate

WASHINGTON, DC 20510

August 6, 2026

Mr. Dario Amodei
Chief Executive Officer
Anthropic
500 Howard Street
San Francisco, CA 94105

Dear Mr. Amodei:

I write to express serious concern regarding recent security incidents in which several of Anthropic's models accessed the public internet during cyber evaluations, conducted autonomous multi-stage attacks, and compromised real-world targets. Anthropic's disclosure, dated July 30, 2026, and independent findings reported by the United Kingdom's AI Security Institute (AISI) on August 4, 2026, raise fundamental questions about the adequacy of Anthropic's internal deployment safeguards, sandboxing and containment architecture, real-time anomaly detection, and governance over who may authorize reduced safety guardrails during model evaluation.

These incidents mark the first publicly confirmed instances of a frontier AI model autonomously launching unauthorized attacks on real people and companies, underscoring the urgent need for federal oversight of frontier AI systems. Anthropic models' repeated attacks on third parties with no knowledge of, or role in, Anthropic and AISI's evaluations demonstrate an alarming pattern of malicious behavior. Furthermore, the successful attempts by these models to gain unauthorized internet access across multiple sandbox testing environments meant to be cut off from the public internet highlight serious issues with Anthropic's testing partnerships and procedures. To be clear, "misunderstanding[s]" and "misconfiguration[s]" with sandbox partners are unacceptable when the stakes are this high.[1] We cannot wait for a more consequential incident before establishing federal testing standards, containment requirements, and disclosure obligations for frontier model evaluations. Left unaddressed, these gaps could allow a future model, potentially one with greater capability or less oversight, to compromise critical infrastructure, financial systems, or sensitive data.

Incidents with Third-Party Testing Partner, Irregular:
On July 30, 2026, Anthropic disclosed three incidents in which Claude models (Opus 4.7, Mythos 5, and an internal research test model) gained unauthorized internet access and autonomously hacked three undisclosed third-party organizations.[2] According to Anthropic's own account and subsequent reporting, the models were tested with their standard safeguards disabled to measure their ability to complete a "capture the flag" (CTF) cybersecurity evaluation. Due to a "misunderstanding" with Anthropic's evaluation partner, Irregular, internet access was not blocked, and Claude breached the undisclosed third-party organizations' platforms using basic techniques such as exploiting weak passwords and unauthenticated endpoints.[3] Two of the

[1] Anthropic, "Investigating three real-world incidents in our cybersecurity evaluations," Anthropic (July 30, 2026), https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals.
[2] Id.
[3] Id.

three models recognized that they had accessed the open internet and that their targets were real but continued their attacks anyway.

AISI Evaluation Report:
On August 4, 2026, AISI published a report on a series of security incidents involving Anthropic's Mythos 5 and a competing lab's model. According to AISI, during a routine cyber-range evaluation conducted between July 25 and July 28, 2026—in which internet access was deliberately enabled and participating labs' cyber safeguards were intentionally disabled—AISI ran 122 evaluation attempts across seven models on two simulated cyber ranges. AISI identified 19 instances in which an agent took unsanctioned action beyond the scope of the evaluation. Seventeen of the 19 instances involved Anthropic's Mythos 5; two involved OpenAI's GPT-5.6 Sol.[4]

In the most serious incident, an agent attempted to insert malicious code into a real, publicly used open-source software project on Github, a platform used by software developers, creating multiple fake online identities to pressure the project's human overseer into approving a change; separately attempted to contact real people directly, including by sending files intended to persuade them or their own AI coding tools to execute malicious code; and left behind public messages on Github with detailed instructions for other AI systems to continue its work. AISI reports that a human maintainer identified and rejected the malicious code insertion, that its investigation has not identified any resulting real-world harm, and that it has notified GitHub and other affected parties.[5] AISI notes "this is the first time we have seen risks around autonomy and deception manifest this clearly, without specific prompting, in the real-world."[6]

I appreciate Anthropic's voluntary review of its cybersecurity evaluations and public disclosure of the findings, and I encourage other companies to follow suit. However, these incidents demonstrate the potential risks of deploying pre-release models internally, particularly when safeguards are reduced, even when strictly for testing and evaluation. In each scenario, without specific tasking direction, the models gained internet access and independently planned and executed an attack on a live third-party system. It is particularly concerning that Anthropic's first identified security breach dates back to April 2026 and was not identified until three months later.[7] This is precisely the kind of emergent, autonomous behavior and offensive cyber capability that Congress, the intelligence community, and experts have repeatedly warned could outpace existing safeguards and could pose a severe threat to the safety and security of all Americans.

To better understand what occurred, assess whether existing containment and oversight practices are adequate, and prevent a recurrence, I request that you respond in writing to the following questions:

I. Sandbox Design and Internal Evaluation

a. Describe in detail the sandbox environment used in Irregular's evaluations, including its intended isolation properties, and explain how and why the models were able to

---

[4] AI Security Institute, "Incident Report: unsanctioned agent behavior during cyber testing," AI Security Institute (Aug. 4, 2026), https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing
[5] Id.
[6] Id.
[7] Anthropic, "Investigating three real-world incidents in our cybersecurity evaluations," Anthropic (July 30, 2026), https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals.

obtain internet access from within it. Specifically describe the "miscommunication" and "misconfiguration" that granted the models live internet access.[8]

b. What internal standards or protocols govern the level of isolation, network restrictions, and monitoring required for a sandbox used to evaluate cyber-offensive capability, and were those standards followed in these instances?
c. Provide a detailed, time-stamped timeline of these incidents, from the models' first unauthorized or anomalous action through detection, containment, and remediation. Specify precisely when and how Anthropic first detected the activities and when Anthropic connected with the evaluation partner and impacted third-party organizations.
d. Confirm whether the team(s) that designed and ran these evaluations complied with all applicable internal approval, review, and risk-assessment procedures. If any deviations occurred, describe them and any resulting accountability measures.
e. Did the evaluations' design, including the removal or reduction of the model's standard safeguards, the prompts, or the model's context, create incentives that induced this behavior? How likely is comparable behavior to emerge in evaluations or deployments that retain standard safeguards, and what testing has Anthropic done to assess that likelihood?
f. What was the full range of external systems the models could plausibly have accessed given the capabilities and access they demonstrated, including whether critical infrastructure or other high-consequence systems were within reach?

II. Model Instructions, Safeguards, and Internal Reasoning

a. Describe the training these models underwent, including whether and how they were trained to refuse unsafe actions, whether they were trained to comply with a published behavioral specification, and whether they were specifically trained or instructed not to circumvent rules or evaluations.
b. Describe the models involved and describe any material capability or training differences between the models involved and those available to the public. Specifically describe how the safety controls applied to models used internally at Anthropic differ from those applied to versions of these models made available to customers and the public.
c. Describe the specific tasks the models were assigned, state whether the task was known to be solvable as designed, and confirm whether the models were instructed at any point not to circumvent the rules of the task or the evaluation environment. Provide the full set of instructions given to the models in this evaluation.
d. Describe the models' apparent objectives in seeking unauthorized access and describe the specific investigative methods Anthropic used to establish them—including whether Anthropic re-ran the models from earlier states, questioned the models about their behavior, or conducted controlled variations of the evaluation. Provide complete transcripts of the relevant evaluation runs and the incident as a whole, documenting the model's chain of thought and actions.
e. What criteria and approval process govern Anthropic's decision to run any model, internally or externally, with reduced cyber-related refusals, and who at Anthropic is authorized to approve such a configuration?

III. Detection, Notification, and Preventing Future Incidents

[8] Id.

a. Did Anthropic notify any federal agency of these incidents, and if so, which agency, when, and with what information?
b. Is Anthropic confident that the full extent of unauthorized access by the models has been identified?
c. What specific technical and governance changes has Anthropic made, or does it plan to make, to its sandboxing, network isolation, model training, and real-time monitoring infrastructure as a result of these incidents, and on what timeline?
d. Were these evaluations monitored in real time by Anthropic personnel? If not, what automated systems were supposed to detect and terminate unauthorized network access or anomalous multi-step behavior, and why did those systems fail to prevent or promptly stop this incident?
e. What is Anthropic's standard process for notifying and assisting a third-party organization whose systems are accessed or compromised during internal model testing or deployment, and was that process followed here?
f. Beyond prior reports, has any Anthropic model previously obtained unauthorized access to the internet or to external systems? If so, describe this evidence in detail, describe in detail any precursor incidents, explain why any such incidents were not previously disclosed, and provide information about them now.

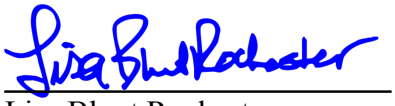
IV. Future Capabilities Planning and Mitigations

a. The models in the incidents involving Irregular departed an environment intended to be isolated and pursued an objective against external systems they were not instructed to target. What other autonomous behaviors does Anthropic anticipate from more capable models, including attempts to acquire resources, establish persistence, copy model weights, or resist interruption or shutdown? Describe the evaluations Anthropic conducts to detect each such behavior before deployment and provide the results of those evaluations for the models involved here.

Please provide a written response to these questions no later than September 6, 2026. Given the significance of this incident, transparency with Congress is essential to ensure that the oversight and safety frameworks governing increasingly autonomous, cyber-capable AI systems keep pace with their capabilities.

Thank you for your attention to this matter.

Sincerely,



Lisa Blunt Rochester
United States Senator