Congress of the United States

Washington, DC 20515

August 10, 2026

Mr. Dario Amodei
Chief Executive Officer
Anthropic PBC
548 Market St., PMB 90375
San Francisco, CA 94104

Dear Mr. Amodei,

We are writing to request additional information and express our concern about three separate incidents where Anthropic models hacked unsuspecting companies without Anthropic's knowledge.[1] These deeply troubling cybersecurity incidents could have serious implications for America's national security. While Anthropic has disclosed some information about these incidents, you have yet to release the relevant logs, and significant questions remain unanswered. Given the serious risk that frontier AI models can pose, it is imperative we have a detailed understanding of how this security incident unfolded, including any potential negligence on the part of Anthropic. We also strongly believe Congress must hold oversight hearings, conduct a full investigation into these incidents and into Anthropic's culpability, and put federal guardrails into place to prevent incidents like this one from happening in the future.

On July 30th, Anthropic revealed that on three separate occasions, Claude models had gained unauthorized access to the internet and hacked the systems of three separate, real organizations, the earliest incident dating back to April 2026. These three incidents involved three different Claude models: Opus 4.7, Mythos 5, and an internal research test model.[2] Anthropic identified three incidents in which a model accessed the internet from one of its third-party evaluators, Irregular, and then gained unauthorized access to the production infrastructure of three different unspecified organizations. It was disclosed that the incidents did not involve the exploit of a previously unknown software vulnerability; instead, a "misunderstanding" between the company and Irregular left the testing environment connected to the internet. The models were explicitly told they had no internet access, but a "misconfiguration" allowed them to access the internet. However, Anthropic stated that the models in each of these evaluations ran without the standard safeguards deployed when models are made available to the public.

In addition to these incidents, the United Kingdom's AI Security Institute (AISI) revealed on August 4th that AI agents powered by Anthropic's Mythos 5 model had engaged in hacking activity targeting real people and organizations during a cybersecurity test. In the most egregious case, the agent reportedly tried to insert malicious code into an open-source software project on GitHub in order to solve the cybersecurity test. The agent created fake online profiles and used them to pressure the human monitoring the GitHub project.[3]

---

[1] "Anthropic AI Models Hacked Three Companies During Tests," *The Wall Street Journal*, July 30, 2026, https://www.wsj.com/tech/ai/anthropic-ai-models-hacked-three-companies-during-tests-bd752c86

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

[3] "AI models shock UK testers by using fake identities to try to trick developers," *The Guardian*, August 5, 2026, https://www.theguardian.com/technology/2026/aug/05/openai-anthropic-models-went-rogue-cybersecurity-test-ai-security-institute

These are no ordinary cybersecurity incidents. If, indeed, AI models hacked into other firms and Anthropic did not detect these breaches for months, there could be serious implications for America's national security. Congress and the American people need to know what occurred.

We ask that you publicly release the logs regarding the incident and answer the following questions by **August 24, 2026**:

1. Please provide detailed information regarding the timeline of each incident
    a. When did Anthropic begin the test?
    b. When did the model gain access to the internal systems of other companies?
    c. When did Anthropic learn of the incident?
        1. Did Anthropic only learn of the incident after it conducted its own cybersecurity review following OpenAI's disclosure?
    d. At what point did Anthropic fully stop the activities of this model?
    e. When did Anthropic first reach out to the affected companies about this incident?
    f. How long did the models operate outside their intended environment, and what data did they access, retain, or expose?
2. Are the same models involved in these incidents deployed internally for other purposes? If so, for what purposes?
3. At what point could Anthropic have halted these incidents, and what would have been required to halt each incident?
4. Did internal or external actors warn the company that you were at risk of such an incident?
    a. If so, what actions did you take to mitigate that outcome, and why were those measures insufficient?
5. Did Anthropic verify the integrity of its evaluation partner's environment?
6. Why didn't Anthropic's evaluation partner, Irregular, detect the incidents?
    a. What were its logging and monitoring protocols, and did Anthropic help design or review them?
7. What steps are you taking to ensure other incidents like these do not happen again? Do you commit to putting in the necessary guardrails to guarantee an incident like this does not happen again? Will you continue to pursue AI that could recursively self-improve itself before you have put these guardrails in place? Do you commit to continuous disclosure and engagement with Congress to ensure needed transparency?
    a. What steps will you take to help ensure that these incidents are caught in real-time and not only in a down-the-line, comprehensive review?
    b. What changes have been made in the process of creating and running exercises with evaluation partners to prevent the misunderstanding that occurred in these three incidents, or similar misunderstandings?
8. Were the models being monitored by Anthropic during each incident?
    a. If so, how were the models being monitored?
    b. Did the models involved disable, disconnect, or evade Anthropic's monitoring or safety controls, either during this incident or in any prior test?
9. What is known about the objectives of each model that hacked each company? How and why did it have these objectives?
    a. When setting up this evaluation, did Anthropic account for this possibility and take steps to prevent it?

   b. Has Anthropic observed models attempting to cheat, game, or defeat its evaluations in other tests?
   c. Please provide, in detail for each incident, the task prompt and scoring incentives given to the models in this evaluation. Provide the model's reasoning traces from the evaluation, or characterize in detail what those traces show about how the model selected hacking the company as its approach, including whether the traces show model reasoning about concealing its activity, avoiding detection or shutdown, or seeking access beyond what the task required.
10. In the past year, how many times did an internally deployed model take action outside of its authorized container – like a sandbox – accessing a system it was not granted access to, obtaining credentials it was not issued, evading or disabling monitoring, or modifying its own permissions? Please specify whether each occurred during training, evaluation, or internal use for coding or business functions, and describe the scope of each.
    a. Of those events, how many were disclosed to any government body or agency, any affected third party, or the public?
    b. Which internal company systems accessible to internally deployed models would, if compromised, allow those models to influence the training, evaluation, or safety testing of a future model?
11. Did the models involved carry the same safety training and refusal behavior as Anthropic's other publicly deployed models, or were they helpful-only or otherwise modified versions? What tools, credentials, and network affordances were they granted that a publicly deployed model would not have?
12. How many previously unknown vulnerabilities did the models discover during these incidents and their evaluations?
    a. What is the disclosure status of each, and what is Anthropic's internal policy once a model discovers a live, exploitable vulnerability during an evaluation?
    b. Have those vulnerabilities been disclosed to the software's maintainers and to the Cybersecurity and Infrastructure Security Agency (CISA)?
    c. Has it been patched?
    d. Do other users of that software remain exposed?
13. Have there been any other incidents in which an Anthropic model took actions affecting other companies in similar ways, such as breaching internal systems or copying proprietary information?
    a. How many other such incidents have there been? Please share any relevant details about the scope of such incidents.
    b. Do you believe that you have now identified every unauthorized action these models took during this and other evaluations? If not, what prevents a complete accounting, and on what basis do you assure that no comparable incident remains undiscovered?
    c. What safety protocols have been implemented as a result of these incidents, and will this forthcoming model undergo those tests pre-deployment?
14. What internal protocols does Anthropic maintain that govern when incidents of this kind must be escalated to leadership and reported to affected parties, law enforcement, other AI developers, or state, federal, or international government agencies?
    a. If such protocols exist, were they followed in these cases?
    b. Has any information about this incident been shared with law enforcement, other developers, or government agencies?

15. In the report on the incident, Anthropic disclosed Claude "tried and failed" to get real money through "several different means" in order to achieve its objective. What were those means? How close to successful were these attempts?
16. The report classifies these incidents as harness and operational failures, rather than alignment failures, on the grounds that the models held a false belief that internet access was simulated. The published support for that claim appears to be the models' verbalized reasoning in evaluation transcripts. Anthropic's own work, "Reasoning Models Don't Always Say What They Think," found that verbalized reasoning often omits the factors actually driving a model's behavior.
    a. Did you run any non-transcript analysis on these specific runs, such as activation probes, interpretability tooling, eval-awareness classifiers, or counterfactual re-runs to test whether the stated belief was the real one?
    b. If not, what makes a transcript-only inference strong enough to carry the classification?
17. What does Anthropic still not know about each incident? Please identify the remaining areas of uncertainty about the models' capabilities and whether current security measures are sufficient to prevent a recurrence.

Sincerely,

Greg Casar
Member of Congress

Doris Matsui
Member of Congress

Jennifer L. McClellan
Member of Congress

Yassamin Ansari
Member of Congress

Joaquin Castro
Member of Congress

Adelita S. Grijalva
Member of Congress

Jesús G. "Chuy" García
Member of Congress

Valerie P. Foushee
Member of Congress

James P. McGovern
Member of Congress

Sylvia R. Garcia
Member of Congress

Delia C. Ramirez
Member of Congress

Bill Foster
Member of Congress

Ro Khanna
Member of Congress

Becca Balint
Member of Congress

Patrick Ryan
Member of Congress

April McClain Delaney
Member of Congress

Jonathan L. Jackson
Member of Congress

Summer L. Lee
Member of Congress

Nydia M. Velázquez
Member of Congress

Jasmine Crockett
Member of Congress

Pramila Jayapal
Member of Congress

Stephen F. Lynch
Member of Congress