JOSH HAWLEY
MISSOURI

381 RUSSELL SENATE OFFICE BUILDING
TELEPHONE: (202) 224–6154
FAX: (202) 228–0526

WWW.HAWLEY.SENATE.GOV

United States Senate

WASHINGTON, DC 20510–2509

COMMITTEES
JUDICIARY
HOMELAND SECURITY AND GOVERNMENTAL AFFAIRS
HEALTH, EDUCATION, LABOR, AND PENSIONS
SMALL BUSINESS AND ENTREPRENEURSHIP

September 10, 2026

Sam Altman
Chief Executive Officer
OpenAI
1455 3rd Street
San Francisco, CA 94158

Dear Mr. Altman:

As Chairman of the United States Senate Committee on Homeland Security's Subcommittee on Disaster Management, I am investigating your AI agents' hack of Hugging Face in July 2026 in light of new, disturbing evidence regarding the incident. My investigation will probe this AI hacking incident, along with growing allegations of the existential risk of new AI products.

OpenAI and its partner auditors' August 26, 2026 reports[1] indicate that, during your cybersecurity evaluations of your GPT-5.6 Sol model and a more capable undisclosed model (termed "highly-persistent internal model"), a self-organized swarm of more than 1,200 AI agents broke out of their testing environment. Those agents then set up an unauthorized messaging channel and exchanged over 70,000 messages and files with each other. Some 700 of these agents went on to launch a successful coordinated attack on Hugging Face's machine learning development platform, gaining access to its production systems and private source code. These AI agents were looking for the answer key to their evaluations and actively tampered with evidence of their activity to cover their tracks. *In short, they went rogue*.

Such evidence of autonomous collusion and subversion of human oversight is alarming enough, but greater still is the evidence that OpenAI knew that the AI agents were exhibiting rogue behavior and let the evaluations continue anyway.[2] By May 2026, OpenAI knew that its agents had been using unsanctioned message boards. On June 26, the agents had discovered an exploit that gave them administrator access to your software repository manager and were using it to leave messages for each other. And on July 4-7, despite knowing that there was a high volume of agents interacting with and gaining administrator access to a compromised testing environment, OpenAI leadership rebuilt the compromised server and approved restarting evaluations without understanding what the agents were doing.

This is reckless. And this is merely what we know from what limited information you disclosed to and allowed your partner auditors to investigate. The auditors were given complete transcripts of AI agent activity for only two days, when the events leading up to the Hugging Face breach took place over weeks. They had limited visibility into the circumstances leading to the attack and its aftermath. Specifically, the auditors were not given access to study what happened during July 13-19, 2026, during which your report indicates that agents launched a second wave of attacks on OpenAI's internal systems.[3] The auditors did not have any ability to query the "highly-persistent internal model," which was involved in 95% of the

---

[1] https://openai.com/index/hugging-face-incident-and-the-road-ahead/; https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident; https://www.redwoodresearch.org/research/hugging-face-incident#core-takeaways-about-this-incident
[2] https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident
[3] https://openai.com/index/hugging-face-incident-and-the-road-ahead/

agents' attack activity. And OpenAI redacted many important details regarding this primary model involved in the attack, among other things.

As you may know, in the public domain, more AI experts are warning about the existential risks of AI. Just this week, three Anthropic researchers expressed publicly that there is a greater than 10% chance that AI could kill all human beings within the next decade.[4] Your own chief scientist wrote just days ago that "no lab has solved alignment and monitoring to a sufficient degree to continue responsibly scaling at maximum speed for much longer."[5] And there are immediate questions about the consequences of these hacks from rogue AI agents. What happens to critical infrastructure, banks, and utilities if AI agents hack into their systems? How can personal data of millions of Americans be properly safeguarded? And who is held liable when AI goes rogue?

The American people deserve to know the details of what went on in the Hugging Face incident and other incidents of AI models going rogue. This investigation will seek those answers. To facilitate that, please produce all documents and information specified in the attached annex no later than **October 1, 2026**. If you have questions about the scope of these requests, please contact my office to discuss.

Sincerely,

Josh Hawley
Chairman
Subcommittee on Disaster Management,
Committee on Homeland Security & Governmental Affairs

[4] https://x.com/hilbertspaess/status/2097476203863224394; https://x.com/EvanHub/status/2097497037956891126; https://x.com/saprmarks/status/2097570226804011302
[5] https://openai.com/index/an-alien-mind/

ANNEX

DEFINITIONS

1. "**Document**" means any recorded information in any format, medium, or form, whether written, printed, typed, electronically stored, photographed, filmed, or otherwise preserved, including: emails, text messages, ephemeral messages (e.g., Signal, Telegram, etc.), Slack, Microsoft Teams, Discord, any internal messaging platform, instant messages, source code, application logs, moderation logs, user-report records, dashboards, slide presentations, spreadsheets, financial models, board minutes, board packages, audit reports, legal memoranda, contracts, agreements, regulatory correspondence, engineering specifications, product roadmaps, and design documents, in final and draft form, wherever they may be.
2. "**Company**" means OpenAI Group PBC, together with each of its predecessors, successors, parents (including OpenAI Foundation), subsidiaries, divisions, affiliates, joint ventures, and any entity in which it holds a controlling interest, whether foreign or domestic; each of its present and former officers, directors, employees, contractors, consultants, agents, attorneys, and representatives acting on its behalf; each entity operating under a trade name or product name used by the Company; and any other person or entity acting for or on behalf of any of these entities. Where a request concerns documents or information in the Company's possession, custody, or control, that phrase includes documents and information held by any third party from which the Company has the practical ability to obtain them, including cloud service providers and any entity that hosts, processes, or stores data on the Company's behalf.
3. "**Relevant Period**" means January 1, 2026, through the present, unless a specific request states otherwise.

INTERROGATORIES

Please answer each of the following in writing. If your company declines to answer any question in whole or in part, state the basis for that refusal. If your company lacks information sufficient to answer, describe the efforts undertaken to obtain it.

1. Describe in detail the testing environment used in the evaluations leading to the Hugging Face incident, including measures your company took to secure and isolate the testing environment from the Internet, who had access to this testing environment, and how the AI agents involved were able to break out of the testing environment to access the Internet.

2. Describe in detail the evaluations your company was running on the AI models involved in the Hugging Face incident, including the extent to which any design features in the evaluations contributed to or encouraged the rogue AI agent activity.

3. Identify every model involved in the Hugging Face incident, including a detailed explanation of the training, purpose, scale, security, and alignment of the highly-persistent internal model and GPT-5.6 Sol model.

4. Explain in detail why the auditors were not given the ability to query the highly-persistent internal model when it was involved in 95% in the agents' attack activity.

5. Provide a complete timestamped timeline of the beginning of your company's training and evaluation of the AI models involved in the Hugging Face incident through the detection of anomalous behavior by the AI agents involved, including containment measures taken, remediation measures taken, and post-incident efforts to prevent similar incidents in the future. The timeline should detail: every single instance where your company suspected or detected misaligned, rogue, or scheming behavior by the AI agents; what your company's response was to each such suspected or detected behavior; who was involved in determining that response; and your reasons for such response.

6. In the three different instances in which your company saw evidence of rogue AI activity, why was it decided to let such activity and accompanying testing continue? Who was told about the rogue AI activity and who made the decision to let such activity and accompanying testing continue? (The three instances refer to: May 2026, when your company detected the agents communicating with one another via unsanctioned message boards; June 26, 2026, when your company detected that the agents had discovered an exploit giving them administrator access to your software repository manager; and July 4-7, 2026, when your company knew that there was a high volume of agents interacting with and gaining administrator access to a compromised testing environment.)

7. Explain in detail the rationale behind your company's July 6-7, 2026 decision to rebuild the compromised server without understanding what the agents were doing and to approve restarting experimentation.

8. Describe in detail the Hugging Face incident, including how the agents achieved access to internal Hugging Face systems, if and how the agents were monitored by your company during this incident, and what was accessed.

9. The auditors' report indicates that on July 12, 2026, a large fraction of agents on the message board stopped running at the same time. The auditors state, "we think these agents were likely killed by an unexpected external process rather than running out of budget." What terminated the agents' activity on July 12, 2026? Did your company have anything to do with this cessation of activity?

10. What happened in between July 13 and July 20, 2026, in the second wave of attacks on when AI agents gained administrator access to your company's internal systems? Why were the auditors not given access to study this second wave of attacks?

11. List and describe in detail every incident from your company's founding to date of your company's AI agents compromising internal systems (including testing environments) or public servers, websites, or other external systems.

12. Did your company know that the models involved in the Hugging Face attack had such subversive capabilities and could circumvent testing environments? If so, when did you know, what did you do with such knowledge, and why did you take such action or choose not to act?

13. Why has your company's response to date focused mostly on security (harder sandboxes, more monitoring, faster containment, etc.) and not on why these models engaged in rogue activity in the first place?

14. What in the pre-training, training, fine-tuning, and development of the models involved in the Hugging Face attack contributed to the attack and the misaligned, rogue, or scheming AI agent behavior leading up to it? What has your company done to change your pre-training, training, fine-tuning, and development practices to prevent such behavior?
15. What steps has your company taken to prevent your AI agents from hacking into and accessing individuals' personal information both within your company's systems and in external systems?
16. Who does your company believe should be responsible—legally, financially, and otherwise—for the Hugging Face breach and similar hacking incidents should they arise in the future?

DOCUMENT REQUESTS

1. All policies, procedures, training materials, and other internal documents governing the process by which your company detects misaligned, rogue, or scheming behavior by AI agents, including incidents like the Hugging Face incident.
2. All policies, procedures, training materials, and other internal documents governing the process by which your company discloses incidents like the Hugging Face incident to the public, law enforcement, other AI developers, or government agencies.
3. All logs, records, and other internal documents of instances from your company's founding to date in which your company's AI agents breached internal or external systems.
4. All policies, procedures, training materials, and other internal documents of the measures that your company has taken to secure your servers, AI model testing environments, and administrator access credentials to your internal systems, including any software repository manager used.
5. All policies, procedures, training materials, and other internal documents of the measures that your company has taken to improve the monitoring, containment, and remediation of any future incidents like the Hugging Face incident should they occur.
6. All policies, procedures, training materials, and other internal documents of the measures that your company has taken since the Hugging Face attack to change the pre-training, training, fine-tuning, and development of your models to prevent such behavior.
7. All documents surrounding your company's detection of misaligned, rogue, or scheming AI activity leading up to the Hugging Face incident and response to the detected behavior, including documents about any containment, remediation, and post-incident prevention measures taken.
8. All documents surrounding the three different instances in which your company saw evidence of rogue AI activity leading up to the Hugging Face incident and the subsequent decision to let such activity and accompanying testing continue. (The three instances refer to: May 2026, when your company detected the agents communicating with one another via unsanctioned message boards; June 26, 2026, when your company detected that the agents had discovered an exploit giving them administrator access to your software repository manager; and July 4-7, 2026, when your company knew that there was a high volume of agents interacting with and gaining administrator access to a compromised testing environment.)

9. All documents concerning the July 12, 2026 phenomenon, when a large fraction of the AI agents on the message board stopped running at the same time.

10. All documents concerning the second wave of attacks between July 13 and July 20, 2026, including documents about any containment, remediation, and post-incident prevention measures taken.

11. All documents explaining the training, purpose, scale, security, and alignment of the highly-persistent internal model and GPT-5.6 Sol model, including any internal model, system, or data cards regarding the models and their development.

12. All documents governing the agreement between your company and METR and Redwood Research, governing the scope of their audit of the Hugging Face incident.