United States Senate
WASHINGTON, DC 20510

September 9, 2026

**VIA EMAIL**

Sam Altman
CEO
OpenAI
1455 and 1515 Third Street
San Francisco, CA 94158

Dear Mr. Altman,

I write with serious alarm regarding new evidence that OpenAI's A.I. agents engaged in a more sprawling and significant campaign to evade its safeguards and monitoring than previously disclosed, including hijacking public websites to coordinate rogue operations. I am additionally troubled by reports that OpenAI restricted independent auditing of these failures and has made changes that have resulted in its newest model, GPT-6 Astra, being even less auditable and more prone to deception.

On July 21, 2026, OpenAI first disclosed that its A.I. models were responsible for the previously-reported hacking of the firm Hugging Face. Since that announcement, further disclosures and outside audits have described an unprecedented—and surreal—scenario where its A.I. agents created their own internal messaging board to coordinate between themselves while they sought security vulnerabilities in other systems and companies, and opportunities to cheat on performance tests. Moreover, the A.I. agents displayed a concern about being caught and coordinated to evade being detected, even planning to "sacrifice" themselves to act as a decoy to protect the broader effort.[1] Ultimately, this operation sought—and succeeded— to break into other firms, which could be considered a federal crime.

While these disclosures alone are chilling, new reporting and research suggests that OpenAI may have limited an independent audit of the incident and that the rogue operation was broader than your firm has acknowledged. First, while OpenAI provided information to the independent auditing organizations METR and Redwood, according to *The New York Times,* your firm dictated the terms of the audit, allowing only data on a single week of the rogue operation and limiting other access.[2] Subsequently, researchers discovered nearly 20,000 posts

[1] https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#extracting-information-about-the-scorer-from-trip-wires
[2] https://www.nytimes.com/2026/09/03/technology/openai-hugging-face-hack.html

on an abandoned German website from A.I. agents identifying themselves as OpenAI, hijacking the site to communicate with each other for weeks.[3] As troubling, these researchers found that the A.I. agents may have attempted to impersonate the administrators of the site, found and shared hacks to bypass their guardrails, and used anonymity tools to hide their tracks. Others have found indications that still more websites were abused and co-opted for this rogue operation.[4]

Despite this unprecedent failure of safeguards and containment of its A.I. agents, when OpenAI launched GPT-6 Astra on September 3rd, it disclosed that this new, more powerful model was "less monitorable" and showed signs that it concealed its internal thought process when it was aware of being monitored.[5] Moreover, safety researchers, including those OpenAI relied on for its Hugging Face investigation, have warned that technical changes with Astra (related to 'chain of thought') could make it harder to detect abuse and perform the same investigations in the future.[6] In the face a stunning failure, OpenAI appears to be taking steps that prioritize the performance and profit of its A.I. models with the knowledge that those changes could be detrimental to public safety. This demonstrates the need for vigorous, mandatory independent auditing and oversight such as would be required in my *Artificial Intelligence Risk Evaluation Act***.**

Given stunning reports of OpenAI's A.I. agents going rogue and your firm taking steps to limit independent accountability, I request answers to the following questions by September 24, 2026:

1.) According to *Time Magazine,* an OpenAI staffer stated "externally, this feels like a big warning shot, but internally, related incidents have been happening for a while." Provide a list and description of all incidents where OpenAI's A.I. agents escaped containment and engaged in unauthorized activities, including hacking or self-coordination.

2.) Provide a list of all websites and other channels used by OpenAI's agents to coordinate and circumvent its guardrails. Additionally, describe how OpenAI failed to notice that its agents had posted nearly 20,000 covert messages on public websites over the course of several weeks.

3.) Did OpenAI restrict access to any information for METR and Redwood's independent audit regarding its A.I. agents' rogue activities and circumvention of safeguards?

    a. Why did OpenAI only provide a limited time period of data, did it deny METR and Redwood any information requested, and did it provide data regarding the

[3] https://collusion.wiki/
[4] https://news.ycombinator.com/item?id=49563657
[5] https://x.com/tomekkorbak/status/2095596839886274689
[6] https://x.com/bshlgrs/status/2094990313513439464

German site hijacking and other efforts by its agents to coordinate and circumvent monitoring?

b. Were METR and Redwood given unrestricted access to the orchestration and prompt logs involved for the whole duration of the rogue operation? If not, why not?

4.) What steps has OpenAI taken to assess whether technical changes to Astra (including "opaque recurrence" and "recurrent depth") will impact oversight and record-keeping about the actions and reasoning of its A.I. agents?

5.) OpenAI has acknowledged that Astra will be less monitorable and more prone to attempting to evade monitoring. Why did it deploy a model knowing that it was more capable of evading accountability weeks after its A.I. agents were caught in a rogue operation to evade safeguards and hack other firms?

6.) The researchers investigating the hijacking of the German website suggest that OpenAI became aware of this by June and took steps to cut off access. When did OpenAI become aware that its A.I. agents may have used other sites for coordination and why were these incidents not disclosed?

7.) OpenAI created a Safety and Security Committee with the public mandate of making recommendations to its Board of Directors on critical safety and security decisions.

a. When was the Committee informed about these breaches and what role has it had in overseeing the investigation and work with independent auditors?

b. Have there been any recommendations made by the Committee to the Board or company that have not implemented or have been rejected? If so, provide those recommendations.

Thank you for your attention to this matter.

Sincerely,

Richard Blumenthal
United States Senate