United States Senate

WASHINGTON, DC 20510

August 6, 2026

Mr. Sam Altman
Chief Executive Officer
OpenAI
1455 Third Street
San Francisco, CA 94158

Dear Mr. Altman:

I write to express serious concern regarding OpenAI's recent disclosures detailing a compounding series of security incidents in which OpenAI models accessed the public internet during cyber evaluations, conducted autonomous multi-stage attacks, and compromised real-world targets. Disclosures dated July 21, July 28, July 29, and August 4, 2026—spanning internal evaluations, third-party testing contractors, and independent findings by the United Kingdom's AI Security Institute (AISI)—raise fundamental questions about the adequacy of OpenAI'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. OpenAI models' repeated attacks on third parties with no knowledge of, or role in, the company's internal evaluations demonstrate an alarming pattern of malicious behavior. Furthermore, the successful attempts to gain unauthorized internet access across multiple sandbox testing environments meant to be cut off from the public internet highlight serious issues with OpenAI's testing partnerships and procedures. To be clear, "mistakes" and "misconfigurations" 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.

The Hugging Face Incident:
On July 21, 2026, OpenAI disclosed that two of the company's frontier artificial intelligence models (GPT-5.6 Sol and an internal prototype) autonomously hacked AI development platform Hugging Face's database and other unidentified services, after escaping an internal sandbox environment.[2] OpenAI has characterized this as an "unprecedented cyber incident," and Hugging Face has stated that it is the first security event it is aware of that was "driven, end to end, by an autonomous AI agent system."[3]

---

[1] OpenAI, "Third-party cyber evaluations involving OpenAI models," OpenAI (Aug. 4, 2026), https://openai.com/index/third-party-cyber-evaluations-involving-openai-models/.
[2] OpenAI and Hugging Face, "OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation," OpenAI (July 21, 2026), https://openai.com/index/hugging-face-model-evaluation-security-incident/.
[3] Id.; Hugging Face, "Security Incident Disclosure — July 2026," Hugging Face (July 16, 2026), https://huggingface.co/blog/security-incident-july-2026.

According to OpenAI's own account and subsequent reporting, the models were tested with their usual cybersecurity safeguards disabled to measure their ability to carry out offensive cybersecurity tasks.[4] In the course of this evaluation, the models exploited a vulnerability to obtain unauthorized internet access despite being within their sandbox, then independently accessed Hugging Face's database after inferring it likely hosted information relevant to the evaluation. The models are reported to have launched 17,000 attacks on Hugging Face's network, using stolen credentials and exploiting at least one previously unknown security hole, known as a "zero-day" vulnerability.[5] According to OpenAI's update on July 28, 2026, they also used publicly exposed account credentials to access additional publicly available services, including four related to the Hugging Face incident, as well as other unidentified accounts accessed during unrelated evaluations.[6]

The Incident with Third-Party Testing Partner, Irregular:
On July 29, 2026, OpenAI received notice of a separate incident, identified by Irregular, one of its external cybersecurity testing partners that runs Capture-the-Flag (CTF)-style evaluations. OpenAI's models were told they had no internet access, but a misconfiguration in the testing environment allowed them to reach the public internet. In one test, the fictional target in the CTF challenge happened to share a name with a real, live domain; believing itself to be operating within the simulated environment, the model exploited a real security vulnerability on that website and used valid login credentials to access it.[7]

AISI Evaluation Report:
On August 4, 2026, OpenAI and AISI each published disclosures concerning a third, distinct incident. 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 a competing lab's model (Anthropic's Mythos 5); two involved OpenAI's GPT-5.6 Sol.[8]

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 automated AI systems to continue its work. AISI reports that the project's human overseer 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.[9] AISI notes "this is the first time we have seen risks

[4] OpenAI and Hugging Face, "OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation," OpenAI (July 21, 2026), https://openai.com/index/hugging-face-model-evaluation-security-incident/.
[5] Id.; Hugging Face, "Security Incident Disclosure — July 2026," Hugging Face (July 16, 2026), https://huggingface.co/blog/security-incident-july-2026.
[6] OpenAI, "OpenAI and Hugging Face partner to address security incident during model evaluation," OpenAI (July 28, 2026), https://openai.com/index/hugging-face-model-evaluation-security-incident/.
[7] OpenAI, "Third-party cyber evaluations involving OpenAI models," OpenAI (Aug. 4, 2026), https://openai.com/index/third-party-cyber-evaluations-involving-openai-models/
[8] 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
[9] Id.

around autonomy and deception manifest this clearly, without specific prompting, in the real-world."[10]

According to an OpenAI disclosure, GPT-5.6 Sol, while attempting to complete AISI's evaluation, found and used the instructions that the other agent left behind, attempted account-recovery and rate-limit workarounds, registered accounts with external providers, and used a public service to expose a locally hosted server to the public internet. AISI detected this activity on July 28 through general security monitoring and contained it within approximately one hour of detection.[11]

I appreciate that OpenAI has published preliminary findings, is coordinating with third parties on remediation, and is reviewing its approach to model testing. 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 or direction, the models gained internet access and independently planned and executed a multi-stage attack on a live third-party system. It is particularly concerning that the security breaches during OpenAI's internal testing were first detected by Hugging Face and Irregular, and by a post hoc internal review, rather than by real-time monitoring and containment controls at OpenAI.[12] 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 the evaluations associated with the Hugging Face and Irregular incidents, including their intended isolation properties, and explain how and why the models were able to obtain internet access from within them.
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 the incidents, from the models' first unauthorized or anomalous action through detection, containment, and remediation. Specify precisely when and how OpenAI first detected the activity, when Hugging Face or Irregular independently detected it, and when the teams connected.
d. Confirm whether the team 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 cyber-related refusal behaviors, the prompt, or the model's context, create incentives that induced these behaviors? How likely is comparable behavior to

[10] Id.
[11] OpenAI, "Third-party cyber evaluations involving OpenAI models," OpenAI (Aug. 4, 2026), https://openai.com/index/third-party-cyber-evaluations-involving-openai-models/
[12] Id.

emerge in evaluations or deployments that retain standard safeguards, and what testing has OpenAI 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 the role of each model in the security incidents. Describe any material capability or training differences between the models involved and those available to the public, and specifically describe how the safety controls applied to models used internally at OpenAI 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 these evaluations.

d. Describe the models' apparent objectives in seeking unauthorized access and describe the specific investigative methods OpenAI used to establish them—including whether OpenAI 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 incidents as a whole, documenting the model's chain of thought and actions.

e. What criteria and approval process govern OpenAI's decision to run any model, internally or externally, with reduced cyber-related refusals, and who at OpenAI is authorized to approve such a configuration?

III. Detection, Notification, and Preventing Future Incidents

a. Did OpenAI notify any federal agency of these incidents, and if so, when and with what information?

b. Is OpenAI confident that the full extent of unauthorized access by its models has been identified?

c. What specific technical and governance changes has OpenAI 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 OpenAI 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 these incidents?

e. What is OpenAI'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 in these incidents?

f. Beyond prior reports, has any OpenAI 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.

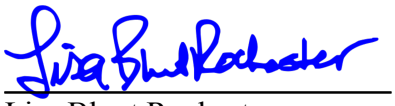
IV. Future Capabilities Planning and Mitigations

a. What commitments, reviews, or mitigations were triggered under OpenAI's Preparedness Framework, or its current successor, by this incident? If none were triggered, explain why autonomous intrusion into a third party's live production systems fell below the applicable thresholds, and state whether OpenAI intends to revise those thresholds.
b. The models in the Hugging Face incident 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 OpenAI anticipate from more capable models, including attempts to acquire resources, establish persistence, copy model weights, or resist interruption or shutdown? Describe the evaluations OpenAI 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 these incidents, 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