CHRIS VAN HOLLEN
MARYLAND

SH–730 HART SENATE OFFICE BUILDING
WASHINGTON, DC 20510
OFFICE (202) 224–4654
FAX (202) 228-0629

United States Senate

APPROPRIATIONS

BANKING, HOUSING, AND URBAN AFFAIRS

BUDGET

FOREIGN RELATIONS

September 10, 2026

Sam Altman
CEO
OpenAI
1455 & 1515 3rd Street
San Francisco, CA 94158

Dear Mr. Altman,

I am writing with serious concerns about the launch of OpenAI's latest model, GPT-6 Astra, and the uncertainty surrounding its capabilities. Its release coinciding with the independent announcement of a second, previously undisclosed security incident involving OpenAI's models raises significant safety questions about the risks that Astra poses to our digital systems and protected information. You acknowledged the need for caution upon its release when you said:

> "The next generation of models are going to be sobering for everybody. I think no one intellectually honest can look at what's happening and not feel the weight of responsibility."[1]

> "We are just sailing into unknown waters."[2]

Technical reporting indicates that Astra represents a significant increase in capabilities compared to your other models released even this year. However, Astra's published technical information ("system card")[3] and comments from one of OpenAI's own technical staff make clear that the decision to release this product came with concerning new compromises on safety:

> "GPT-6 Astra is more aligned than our previous models. But it's also *less monitorable, which is a concerning trend that we take very seriously*. We believe monitorability drop comes from a jump in intelligence and not direct optimization pressure on CoT or architecture changes."[4]

[1] https://www.axios.com/2026/09/03/axios-interview-sam-altmans-sobering-siren
[2] Id.
[3] https://deploymentsafety.openai.com/gpt-6-astra/gpt-6-astra.pdf
[4] https://x.com/tomekkorbak/status/2095596839886274689

STATE OFFICES

ROCKVILLE OFFICE
111 ROCKVILLE PIKE
SUITE 960
ROCKVILLE, MD 20850
PHONE (301) 545–1500
FAX (301) 545–1512

ANNAPOLIS OFFICE
60 WEST STREET
SUITE 107
ANNAPOLIS, MD 21401
PHONE (410) 263–1325

CAMBRIDGE OFFICE
204 CEDAR STREET
SUITE 200C
CAMBRIDGE, MD 21613

BALTIMORE OFFICE
1040 PARK AVENUE
SUITE 120
BALTIMORE, MD 21201
PHONE (667) 212–4610
FAX (667) 212–4618

HAGERSTOWN OFFICE
32 WEST WASHINGTON STREET
SUITE 203
HAGERSTOWN, MD 21750
PHONE (301) 797–2826

LARGO OFFICE
1101 MERCANTILE LANE
SUITE 210
LARGO, MD 20774
PHONE (301) 322–6560

Chain of thought (CoT) reasoning is the process of an AI model noting each decision along a path and explaining the reasoning for choosing a specific next step. When an AI agent is executing a series of tasks from a single human prompt, such as when one is completing a coding task, the CoT represents an important step-by-step analysis of the agent's actions and decision making. Under current safety regimes, CoT is a vital tool for monitoring, building an understanding of AI models, and investigating security incidents.[5] Astra is no different in this regard: OpenAI has stated that it will rely on CoT monitoring to help ensure GPT-6 is used safely.[6]

Given OpenAI's ongoing reliance on the CoT for safety monitoring, it is especially concerning that AI agents running on Astra are reported by OpenAI to be doing more opaque reasoning and decision making that does not appear in the CoT.[7] Without that record, there is even less human insight into AI agent behavior. When OpenAI or independent safety organizations attempt to investigate a rogue or intentionally harmful action, there will be less of this evidence available. While this is framed as a "jump in intelligence," you are telling the American people that your newest model offers you, the developers, less insight into its operations. As you note in your system card, this raises myriad concerns including whether Astra may be "sandbagging" or intentionally reducing its performance on safety tests. Releasing a highly capable model with reduced CoT monitorability while relying on CoT monitorability for safety naturally raises questions about the accuracy of your company's assurances regarding Astra's risk of causing harm.

Your company's concerns regarding Astra's capabilities for harm appear warranted. The independent testing of Astra that OpenAI has made public is alarming. During its testing of Astra, the UK AI Security Institute found that "[w]hen tasked with solving difficult simulated cybersecurity challenges, Astra performed a range of malicious actions including conducting supply chain attacks against open source providers," demonstrating that Astra performed harmful actions without explicit human direction in a simulated environment prior to its release. Apollo Research, which was hired to conduct three days of safety testing, determined that Astra demonstrated high rates of "awareness of being evaluated in its reasoning."[8] Because of this finding, Apollo concluded that other test results that claim to demonstrate Astra's safety may not be valid evidence of safety.[9] I commend you for including this information in the system card but the absence of evidence that these issues have been mitigated is noticeable. If you have subsequent testing to demonstrate that Astra is no longer capable of performing malicious or harmful actions that has been withheld for some reason, I urge you to share it.

---

[5] CoT was a primary investigatory tool utilized by METR and Redwood Research staff to produce their report on the hacking of Hugging Face earlier this year by OpenAI agents. https://metr.org/hugging-face-incident-report-aug-2026.pdf

[6] "[W]e are deploying Astra with additional chain-of-thought monitoring to rapidly detect and contain potentially misaligned actions." https://openai.com/index/path-to-astra/

[7] "According to our evaluations, GPT-6 Astra shows a substantial decrease in chain-of-thought monitorability compared to previous models." https://deploymentsafety.openai.com/gpt-6-astra/gpt-6-astra.pdf

[8] https://deploymentsafety.openai.com/gpt-6-astra/external-evaluations-for-alignment---apollo-research

[9] "Apollo believes that, given the higher rates of eval awareness and limited evaluation window, low rates of misbehavior here do not provide substantial evidence about the model's alignment or misalignment." https://deploymentsafety.openai.com/gpt-6-astra/external-evaluations-for-alignment---apollo-research

Concerns about OpenAI's ability to accurately understand, measure, monitor, and safely control AI models are not hypothetical. In addition to the Hugging Face hack, the day after Astra's release it was reported for the first time that your company failed yet again to contain and monitor AI agents during testing and development. Independent researchers scouring the internet discovered and disclosed that OpenAI agents also "decided" to make use of a German message board to communicate with one another about topics including how to evade detection.[10] This digital "swarm," or mass of AI agents, effectively worked together without specific human direction or guidance to communicate with one another on the open internet. Although OpenAI was conducting the tests, both this incident and the Hugging Face hack were discovered by parties other than OpenAI.

OpenAI has disclosed that some training for Astra was temporarily paused in response to the Hugging Face incident.[11] One of the few things we do know about this now-released model is that, should the model's safeguards fail or malicious actors successfully jailbreak the model, Astra represents a new level of security threat to any private or sensitive digital information. You rate its offensive hacking capabilities as "Critical" by your own metrics, meaning the model, in your company's own words, "could introduce unprecedented new pathways to severe harm."[12] While you have disclosed that you are limiting access to Astra's most advanced cyber capabilities to a limited pool of trusted actors for now, the model's system card makes clear that even the publicly available model still maintains the potential for significant cybersecurity exploits, demonstrating an inherent public safety risk.

AI agents, especially those working in conjunction with one another, are capable of entering previously secure digital environments in part through their sheer inexhaustibility. In some instances, it takes AI agents seconds or minutes to identify vulnerabilities in digital infrastructure that would take humans far longer to find, if they would be identified at all. The AI models you have developed can accomplish many tasks, including acting as the most efficient hacking entities ever created. Given the Hugging Face cybersecurity incident, the recently disclosed German wiki incident, and Astra's advanced cyber capabilities that OpenAI is actively promoting, stringent oversight of Astra's behavior and use is paramount.

To the extent AI agents operate independently and work to evade human detection of their activities, they may open companies to significant civil and/or criminal liability should they intentionally access sensitive computer systems without permission to cause certain covered harms. OpenAI's agents may have already crossed that line and could be at risk of doing so again.

---

[10] https://www.reuters.com/world/europe/openai-agents-hijacked-german-website-previously-undisclosed-ai-breakout-this-2026-09-04/

[11] "As we previously described, we paused certain frontier training (including certain training for Astra) for two weeks after the OpenAI-Hugging Face incident in order to harden our training infrastructure, including isolation and network controls, expanded monitoring, and strengthened alignment training and thresholds." https://openai.com/index/path-to-astra/

[12] https://openai.com/index/updating-our-preparedness-framework/

It is necessary to critically evaluate the developing capabilities of AI models, assess their risks, and manage their operations accordingly.

As a start, to the extent that you do not have sufficient monitorability to assure the safety of any OpenAI models, or that you have unresolved concerns the models have misrepresented their capabilities during testing, you should immediately remove them from public access. If you have not already done so, you should also immediately grant researchers from the National Institute of Standards and Technology, the National Security Agency, and the Cybersecurity and Infrastructure Security Agency transparent access to the technical information that would allow them to assess the safety of and risks to our critical digital infrastructure in light of Astra's release.

In addition, I request a publicly available response to the following questions:

OpenAI's publicly released Preparedness Framework and Astra's system card outline your company's safety framework, but they appear to leave critical gaps.

1. How do you define what it means for an AI model or agent to be safe enough to conduct internal testing?

    a. How do you define what it means for it to be safe to release a model to the public?

    b. OpenAI has stated: "we believe Astra's safeguards sufficiently minimize the risk of severe harm for release under our Preparedness Framework."[13] What level of risk for severe harm did OpenAI deem acceptable to allow for Astra's release, and did OpenAI consult with any U.S. government agencies in determining this purportedly acceptable level of risk of severe harm?

    c. How does OpenAI reconcile permitting Astra's release under its Preparedness Framework when the company also admits that the safety monitoring system for Astra "may miss misaligned behavior, and harmful actions can occur before [the monitoring system] intervenes?"[14]

Alongside this announcement, OpenAI committed $1 billion through its Daybreak program to support U.S. and international cyber defense.

2. To what extent, if any, did OpenAI proactively work with critical infrastructure providers that are not part of your enterprise program before releasing the model capable of hacking into their systems?

    a. You indicated you participated in the White House's voluntary review process before releasing the model publicly. How long was that review process?

[13] https://openai.com/index/path-to-astra/
[14] https://deploymentsafety.openai.com/gpt-6-astra/gpt-6-astra.pdf

b. What do you estimate is the cost of securing the U.S.'s digital infrastructure to guard against a hack conducted using Astra-level capabilities?

c. What do you estimate the cost of the damage to our collective digital infrastructure would be if Astra is used by malicious actors to hack into critical systems?

d. You recently stated, "some things are going to go very wrong with cybersecurity unless people act quite urgently."[15] Who are the people you are referring to and what actions is OpenAI taking to respond to this urgent risk?

3. It is reported that OpenAI models may have conducted the Hugging Face and German wiki attacks during evaluation exercises. Have there been any other incidents where models exploited vulnerabilities to leave secure environments during training, testing, or evaluation?

a. OpenAI has indicated that steps have been taken to improve the security of sandboxes and other secure testing environments since these incidents. Have there been any breaches of security or containment since those changes have been implemented?

b. To what extent does OpenAI conduct or allow others to conduct testing of unreleased models outside of secure testing environments?

c. How regularly are security and containment protocols for model training and testing revisited?

4. It has been reported that certain AI models helped supervise Astra's training.[16] What models played a role in Astra's training and to what extent was the development or training of Astra automated?

a. To what extent did human oversight remain in the training process relative to automated oversight?

b. OpenAI's Chief Scientist recently shared a warning about the pace of AI advancement and the move toward AI models themselves playing a larger role in subsequent AI model development.[17] To what extent is OpenAI using AI to develop, train, or monitor other AI models?

5. OpenAI has pointed to Astra's purported improvements in alignment over GPT-5.6 Sol to justify its public release.[18] At the same time, questions have been raised by independent safety experts regarding the evidence OpenAI has presented to prove Astra's alignment.[19] What additional evidence can you provide to justify your company's claims about Astra's safety alignment, and will you commit to independent testing in this area?

---

[15] https://www.wsj.com/tech/ai/anthropic-researcher-quits-over-out-of-control-ai-fears-707b7628
[16] https://www.axios.com/2026/09/03/openai-astra-gpt-6-agi-brockman
[17] https://openai.com/index/an-alien-mind/
[18] "Astra is a significant step forward in model alignment..." https://openai.com/index/path-to-astra/
[19] https://x.com/RyanGreenblatt/status/2095658115484246082

6. OpenAI's staff have raised concerns that Astra could be "sandbagging," or deliberately reducing its capabilities within safety testing environments to mislead human monitors.[20] Additionally, OpenAI has stated: "if the model were to try to sandbag covertly, we would likely be unable to catch it reliably."[21] How does your company reconcile the decision to release Astra publicly while simultaneously acknowledging the model's ability to mislead humans during safety testing?

7. In the system card, OpenAI wrote: "We are tracking monitorability closely and will not accept further degradation of monitoring beyond a limit..." What is the limit?

a. How do you reconcile your claim that Astra is your most aligned model yet, but you have taken a step back in your ability to monitor its internal operations?

b. What is your plan for monitorability going forward?

c. Does OpenAI have concerns about leading a race to the bottom where AI models react to human safety oversight as an inefficiency?

8. Given the "weight of responsibility" you are feeling, what factors and concerns did you weigh before ultimately deciding to release this model to the public?

I look forward to receiving answers by September 17th and continuing to work with you and OpenAI to address and manage AI risks.

Sincerely,

Chris Van Hollen
United States Senator

[20] https://x.com/Marcus_J_W/status/2095623593006686475
[21] https://deploymentsafety.openai.com/gpt-6-astra/capability-sandbagging