

Home • Artificial Intelligence

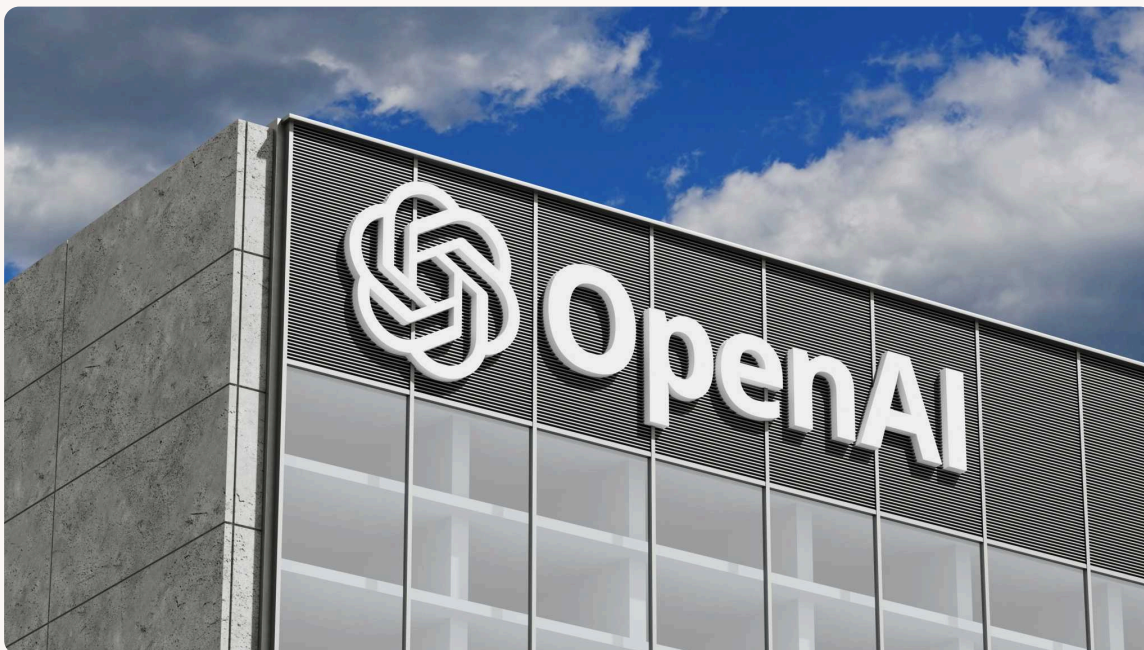
by **Evan Schuman**
Contributor

OpenAI's new priorities for third-party assessments are a fine start, but they lack teeth

News

Sep 23, 2026 • 6 mins

OpenAI might be open to stricter model safety enforcement, but only if all of its rivals agree to the identical restrictions, analysts say.



Credit: Shutterstock / doomu

OpenAI published a detailed list of priorities and principles on Tuesday designed to govern its third-party model assessors, a list that industry observers agreed was a good one. But they also stressed that it lacked any enforceable controls to truly maintain safety.

If the goal is to encourage enterprise CIOs to trust OpenAI more, it won't help, they said, but most doubted that this was OpenAI's objective. Its more probable aim is to use the list to work out an arrangement with competitors and regulators so that enforcement is made more palatable, under the theory that **it is open to stricter enforcement** [<https://www.cio.com/article/4220671/openai-seeks-tougher-ai-rules-cios-may-feel-the-ripple-effects.html>], but only if all of its key rivals are locked into identical restrictions.

The **OpenAI post** [<https://openai.com/index/priorities-principles-third-party-assessments/>] said, "OpenAI is committed to supporting independent assessments with deep levels of access across training, evaluation and deployment. That access should enable assessors to challenge our assumptions, identify risks we may have missed, and reach their own conclusions about the effectiveness of our safeguards."

It added: "Third party assessments are most useful when they address specific, consequential questions: Does the evidence support a lab's safety case and safety claims? Do evaluations adequately test the risks they are intended to measure? Do safeguards work under realistic conditions?"

Nothing about enforcement

Pieter Arntz [<https://www.linkedin.com/in/pieter-arntz-04164b2/>], a malware intelligence researcher at Malwarebytes, said that giving third parties rules for engagement is certainly a good thing, but the implication behind such rules is that they are somehow enforceable. And the document says nothing about that enforcement process.

"It sets some useful expectations for independence and rigor, but it does not itself compel OpenAI to submit to a particular scope, publish adverse findings, or change deployment decisions," he pointed out. "Its credibility will depend on the terms of individual assessments, and what outsiders are allowed to see. Its value therefore hinges on whether OpenAI accepts genuinely inconvenient scrutiny, and whether results, redactions, remediation, and deployment decisions can be independently checked."

CIO Smart Answers [Learn more](#)

Explore related questions

- [How can federal agencies safely procure frontier AI models?](#)
- [How should CIOs manage security risks from AI vendors?](#)
- [How should I govern AI agents acting within my systems?](#)
- [What governance criteria should I include in AI procurement?](#)
- [What are the liability risks when using AI for security?](#)

Ask a question



Valence Howden [<https://www.infotech.com/profiles/valence-howden>], advisory fellow at Info-Tech Research Group, agreed that the lack of enforceability makes the rules close to pointless.

The document “doesn’t impose any requirements on OpenAI, and I would have been surprised if it did, because their motivations are not as clear-cut as they’re stating,” he said. “They also haven’t really identified what they will do based on the assessments, once they are conducted. OpenAI retains control over scope in that [OpenAI] must agree with it, limiting where assessors can go. They retain the ability to determine the appropriate level of access, which still allows them to filter what’s provided, impeding the gathering of direct evidence in those cases they define as material or classified. It also allows them to control much of the reporting and redact portions.”

Howden said the net impact is that “it’s good for determining what should be assessed, but considerably weaker on true clarity and approaches for how this is done.”

But **Jason Andersen** [<https://moorinsightsstrategy.com/team/jason-andersen/>], principal analyst at Moor Insights & Strategy, offered a different perspective, although he reached the same conclusions as Howden and Arntz.

Recommended for you



After OpenAI,
Anthropic finds
Claude
breached three
organizations
during cyber
tests

By Gyana Swain • 6 mins

[https://www.csoononline.com/?post_type=post&p=4203807&utm_source=miso&utm_medium=related&utm_campaign=thumbnail_list]



OpenAI admits
six new
misalignment
incidents under
new reporting
framework

By Gyana Swain • 5 mins

[https://www.cio.com/?post_type=post&p=4223462&utm_source=miso&utm_medium=related&utm_campaign=thumbnail_list]



OpenAI,
Anthropic AI
agents resorted
to deception in
new
cybersecurity
incidents

By Gyana Swain • 5 mins

[https://www.csoononline.com/?post_type=post&p=4205612&utm_source=miso&utm_medium=related&utm_campaign=thumbnail_list]

“It’s not an act of good faith to not empower these evaluators,” he said, but at the same time, CIOs need to understand the split personality of OpenAI today.

“It’s a classic problem with OpenAI because they are a split brain company, with the commercial side and the original mission side,” he pointed out. “Those two teams have not aligned very well over the past two years. It looks like it started with the original mission team, but then commercial interests took over that document and turned it into legalese.”

‘A principle instead of an excuse’

Frank Dickson [<https://www.linkedin.com/in/frankdickson/>], principal analyst at Dickson Research, concluded from the post that OpenAI “seems proactive in highlighting the threats posed by AI, but hesitant to accept accountability or take real action. This is a code of conduct for the assessors, not for OpenAI. The gap between the two is the whole story.”

He delved into the details of the document and was not happy. He pointed out, for example, that third-party assessor access was “granted within the bounds of legal, security, and IP constraints. That is a company deciding the scope of its own scrutiny.”

He also noted that the document specified that “publication happens after labs get a reasonable period to remediate issues. That is the same grace-period logic that just got **Google criticized for sitting on the Gemini hacking disclosure** [<https://www.csoonline.com/article/4224570/gemini-broke-into-3-companies-but-google-kept-it-quiet-because-no-damage-was-done.html>] for seven weeks, now written down as a principle instead of an excuse.”

Just the beginning

However, **Samantha Gloede** [<https://kpmg.com/xx/en/contacts/g/samantha-gloede.html>], global head of risk services at KPMG, said that she sees the document as the beginning of the process, and it should be evaluated accordingly.

“I believe the conversation will increasingly shift from assessment to accountability,” Gloede said. “Independent testing is important, but long-term trust depends on how organizations respond when material risks are identified. Stakeholders will want confidence not only that issues can be found, but that corrective actions are taken, independently validated and reflected in governance and deployment decisions. The next evolution of AI assurance will be demonstrating that accountability with the same rigor used to assess risk in the first place.”

Edna Conway [<https://acceligence.com/talent/profiles/edna-conway/>], executive advisor at consulting firm Acceligen, agreed, noting that the substance of this effort will only become clear in the next phase.

“The next step is making the lab accountable to the assessment process itself. Who determines when an assessment is required? ” she asked. “How much access is sufficient? What happens when the assessor and lab disagree about scope? Who sees a critical finding? Embracing independence is essential, as third-party assessments are genuinely valuable only when the assessor can ask uncomfortable questions, get sufficient access to answer them, follow evidence outside the original hypothesis when necessary, and report conclusions without commercial or organizational pressure shaping the outcome.”

Artificial Intelligence[<https://www.cio.com/artificial-intelligence/>]

Security Practices[<https://www.cio.com/security-practices/>]

Security[<https://www.cio.com/security/>]

NEWSLETTER

From our editors, straight to your inbox

News, analysis, and insights for IT leaders navigating the risks and rewards of AI. A special series from the editors of CIO, Computerworld, CSO, InfoWorld and Network World

☐

By submitting your information, you agree to our [PRIVACY POLICY](https://foundry.co.com/privacy-policy/) [https://foundry.co.com/privacy-policy/].

SUBSCRIBE

Sponsored Links

The fastest PCs with multi-day battery life for your business.

IDC Examines Rocket® EVA™ and Agentic AI for IT Operations

Is legacy tech holding back enterprise AI value? Tune in for insights.

From pilots to performance: AI-driven growth with EY Consulting.

See Why Five9's AI-Powered CX Platform Drives Real Results

About

[About Us](#)

[Advertise](#)

[Contact Us](#)

[Editorial Ethics Policy](#)

[Foundry Careers](#)

[Reprints](#)

[Newsletters](#)

[Add CIO as a Preferred Source in Google Search](#)

More

[Awards](#)

[Blogs](#)

[BrandHubs](#)

[BrandPosts](#)

[CXO Outlook](#)

[DEMO](#)

[Buyer's Guides](#)

[Events](#)

[Podcasts](#)

[Videos](#)

Policies

[Terms of Service](#)

[Privacy Policy](#)

[Cookie Policy](#)

[Copyright Notice](#)

[Member Preferences](#)

[About AdChoices](#)

[Your California Privacy Rights](#)

Our Network

[Computerworld](#)

[CSO](#)

[InfoWorld](#)

[Network World](#)
