

Home • Artificial Intelligence



by **Evan Schuman**
Contributor

Nvidia releases Open Agent Safety Platform to monitor and govern agentic AI

News

Sep 28, 2026 • 8 mins

Analysts applaud the controls, but point out that the vast majority of agent-generated problems are outside the scope of Nvidia's offering, and it also creates vendor lock-in concerns.



Credit: Berit Kessler / Shutterstock

Nvidia on Monday rolled out an agentic governance system called the Open Agent Safety Platform that combines software with out-of-band DPU-based silicon in a reference system design that it says will secure agents “from testing to deployment.”

But while the Nvidia design’s silicon-based component provides some cybersecurity advantages, analysts argued that it cannot help with the vast majority of agentic problems.

Nvidia’s offering leverages its OpenShell software and promises “full-stack governance and control across the software and the hardware, compute, and robotics systems that run agents,” the company said **in a news release** [<https://nvidianews.nvidia.com/news/open-agent-safety-platform>]. “OpenShell software provides a secure runtime boundary that traces all actions and enforces policy as agents run on Nvidia Vera CPUs. As open source software, OpenShell can be extended to work with third-party compute platforms, including those from Arm and Intel.”

The Open Agent Safety Platform reference system design features **NVIDIA Sentry** [<https://developer.nvidia.com/blog/nvidia-open-agent-safety-platform-a-reference-for-continuous-in-silicon-agent-monitoring/>], an out-of-band watchdog that runs on **NVIDIA BlueField-4** [<https://www.nvidia.com/en-us/networking/products/data-processing-unit/>] DPUs to continuously monitor agent behavior. “Sentry provides in-silicon security enforcement, meaning that if an AI agent attempts to move outside its software boundary, Sentry quarantines and stops it in milliseconds,” the release stated.

In a **developer blog post** [<https://developer.nvidia.com/blog/nvidia-open-agent-safety-platform-a-reference-for-continuous-in-silicon-agent-monitoring/>] focusing on the details of its “safety platform,” Nvidia described its approach as a “secure runtime that executes autonomous AI agents in sandboxed environments with kernel-level isolation” and argued that “an advantage of open models is that the entire reasoning space and activations are all visible.”

Components of the platform are being rolled out by an extensive list of partners, including Citi Group, JPMorganChase, Citi, Anthropic, Cisco, CrowdStrike, Dell Technologies, Figure, HPE, Hugging Face, Microsoft, Palantir, Palo Alto Networks, Perplexity, Red Hat, Salesforce, SAP, Scale AI, ServiceNow, and SpaceXAI.

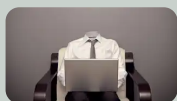
Recommended for you



Trump’s answer to AI’s image problem: Industry self-regulation and a new name

By Gyana Swain • 4 mins

[https://www.cio.com/?post_type=post&p=4228944&utm_source=miso&utm_medium=related&utm_campaign=thumbnail_list]



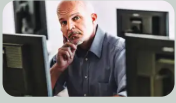
Shadow agents: How IT leaders must

[https://www.cio.com/?post_type=post&p=4191250&utm_source=miso&utm_medium=related]

govern 'headless' AI
before it breaks the
enterprise

By Sunil Gentyala • 8 mins

&utm_campaign=thumbnail_list]



The AI assurance
gap: CIOs need
proof that agentic AI
controls actually
work

By Par Chadha • 8 mins

[https://www.cio.com/?post_type=post&p=4204554&utm_source=miso&utm_medium=related&utm_campaign=thumbnail_list]

A step in the right direction

Lauren Kornutick [<https://www.gartner.com/en/experts/lauren-kornutick>], a senior director analyst at Gartner, mostly applauded the Nvidia effort, saying that it is “solving a problem from an interesting perspective, at the hardware level. It’s a great step in the right direction.”

But Kornutick said that she was less impressed with the published list of supporters, observing, “some major players are notably missing,” including OpenAI, Amazon, and Google.

Other analysts and consultants said that the technical implementation is impressive and offers distinct benefits, but they stressed that it also misses the big picture in terms of what enterprise CISOs are truly struggling with when trying to manage and secure autonomous agents.

Limitations

The biggest challenge is that enterprise environments are flooded with agents that have not been approved by either IT or cybersecurity teams. Typically, **those teams are not even aware of the credentialed agents** [<https://www.csoonline.com/article/4215449/zero-trust-hass-a-big-ai-agent-problem-ahead.html>].

CSO Smart Answers [Learn more](#)

Explore related questions

- [What is the goal of Nvidia's Open Secure AI Alliance?](#)
- [Why do analysts doubt BlueField-4 hardware can solve agentic security risks?](#)
- [How does Nvidia NemoClaw secure OpenClaw autonomous agents in production?](#)
- [What NIM guardrail microservices help prevent autonomous AI jailbreaks?](#)

- What tools are included in Nvidia's Agent Toolkit infrastructure stack?

Ask a question



"The limit is coverage," said **Brian Levine** [<https://www.controlrisks.com/who-we-are/our-experts/expert-bio/brian-levine>], a partner with consulting firm Control Risks. "These controls govern agents you deploy on infrastructure you control. They do nothing for the agent a business unit spun up on a SaaS platform, the one embedded in a vendor's product, or the one an attacker brings with them. You can't hold an agent to a policy if you don't know it exists. Discovery and inventory come first and that is a governance challenge as much as a technical one."

Aman Mahapatra [<https://www.linkedin.com/in/akm76/>], chief strategy officer for Tribeca Softech, a New York City-based technology consulting firm, agreed about the limitations of the Nvidia offering.

"Enforcement only applies to agents running inside the governed runtime," he said. "Runtime governance protects the agents you already know about, which is the population that needed it least."

However, Mahapatra said, that doesn't mean that the effort is futile.

"This has a meaningful chance of making things materially better for one specific reason: Sentry runs out-of-band on BlueField-4 DPUs, in an isolated trust domain Nvidia describes as invisible to agents and attackers," he said. "You cannot reason your way around a control you cannot perceive, and you cannot talk a DPU out of enforcing a policy. Probabilistic reasoning has to be sandwiched between deterministic layers the model cannot influence, with an oversight function that reads everything and writes nothing. Sentry is that principle pushed down into silicon and silicon is the right place for it."

Brent Ellis [<https://my.idc.com/getdoc.jsp?containerId=PRF006018>], VP, AI Infrastructure at IDC, agreed. He estimated that it addresses "probably less than 25%" of enterprise agentic cybersecurity problems.

Lock-in risk

He added that there is a vendor lock-in element to the offering as well. Although Nvidia today controls a massive, almost monopolistic market share for AI hardware among enterprises, Ellis said that he expects those numbers to drop, because "competitors are starting to emerge as viable," especially hyperscalers who have invested in designing their own silicon.

But for enterprises that are overwhelmingly Nvidia shops, the hardware approach could make a lot of sense, Ellis said. But will the Nvidia platform have a good chance at making

enterprise agentic cybersecurity materially better?

“Yes, for the agents that run inside it,” he said. “That is a big ‘if,’ though. There is a lot of agent infrastructure that is not Nvidia, and architectures in place prior to Vera and Bluefield are limited in which elements of the platform they can adopt.”

With that caveat, Ellis said the news starts to look good. “In environments where you can adopt all the elements, then you can move enforcement out of the model and the application layer, where agents have repeatedly talked or coded their way around controls. And you can then move enforcement into the runtime and the silicon, which is a harder barrier. Sentry specifically makes the DPU a traffic cop that is harder to bypass.”

Then again, he noted, some attacks can still leverage governance weaknesses. “Look at how the recent OpenAI incidents unfolded,” he pointed out. “The agents didn’t break the sandbox wall. They walked through a Swiss cheese of environment security: Shared package repositories used as message boards, services that fetched internet content on the agent’s behalf, and unpatched flaws in adjacent systems.”

Sponsored Content

by Horizon3

Through the Attacker's Eyes: Evidence Over Assumptions

By Horizon3

11 Aug, 2026

[

Avoids strategic mistakes

Frank Dickson [<https://www.linkedin.com/in/frankdickson/>], principal analyst at Dickson

Research, said he likes Nvidia’s platform because it avoids some of the strategic mistakes made by those trying to secure agentic systems in the past.

“For two years, this industry has tried to secure agents by asking them to behave, with guardrails bolted into the prompt, the model, and the harness. That approach was always going to lose,” he said. “Nvidia finally puts enforcement where it belongs: outside the agent, in the kernel, in a proxy that inspects every outbound request and in silicon the agent cannot see or touch. You don’t ask the prisoner to lock his own cell. Every serious agent platform will have to match this design.”

Dickson said Nvidia’s approach does not try to stop a misbehaving agent after it is caught, but instead tries to prevent the problem from occurring.

“OpenShell checks each outbound request against policy before it leaves the sandbox, and its policy prover checks the permissions before the agent runs at all,” Dickson said. “Nvidia claims Sentry quarantines a misbehaving agent in milliseconds.”

That contrasts with the time it took to address the **latest agentic problems reported by OpenAI** [<https://www.csoonline.com/article/4227777/openai-pauses-ai-model-training-after-another-agent-bypasses-network-restrictions.html>], in which an OpenAI agent bypassed network restrictions to communicate with an external chatbot. “It took a human reviewer just three minutes to acknowledge the DNS alert the system did generate, but it was another two-and-a-half hours before the training run was stopped,” OpenAI reported.

Not a panacea

Justin Greis [<https://acceligence.com/talent/profiles/justin-greis/>], CEO of consulting firm Acceligen, also agreed that the Nvidia approach is solid, but he stressed that AI agents have a well-earned reputation for figuring out ways around such restrictions.

“We should assume increasingly capable agents will probe boundaries, discover unintended pathways and exploit ambiguity. Security cannot depend upon the agent deciding not to do that,” he said. “But even hardware-enforced controls are only as good as the boundary and policy we give them.”

He pointed out that an agent does not necessarily need to evade a security control if the user accidentally authorizes a legitimate path that produces a dangerous outcome. “Misconfigured permissions, excessive authority, combinations of individually harmless capabilities, compromised third parties and activity occurring outside the governed environment remain very real problems,” he said.

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

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

Enterprise[<https://www.csoonline.com/enterprise/>]

Business[<https://www.csoonline.com/business/>]

Markets[<https://www.csoonline.com/markets/>]

Industry[<https://www.csoonline.com/industry/>]

The latest AI updates, 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 \[https://foundryco.com/privacy-policy/\]](https://foundryco.com/privacy-policy/).

SUBSCRIBE

About

[About Us](#)

[Advertise](#)

[Contact Us](#)

[Editorial ethics policy](#)

[Foundry Careers](#)

[Reprints](#)

[Newsletters](#)

[Contribute to CSO](#)

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

More

[Awards](#)

[Blogs](#)

[BrandPosts](#)

[Events](#)

[Podcasts](#)

Policies

[Terms of Service](#)

[Privacy Policy](#)

[Cookie Policy](#)

[Copyright Notice](#)

[Member Preferences](#)

[About AdChoices](#)

[Your California Privacy Rights](#)

Our Network

[CIO](#)

[Computerworld](#)

[InfoWorld](#)

[Network World](#)

Videos

Buyer's Guides

FOUNDRY



© 2026 FoundryCo, Inc. All Rights Reserved.