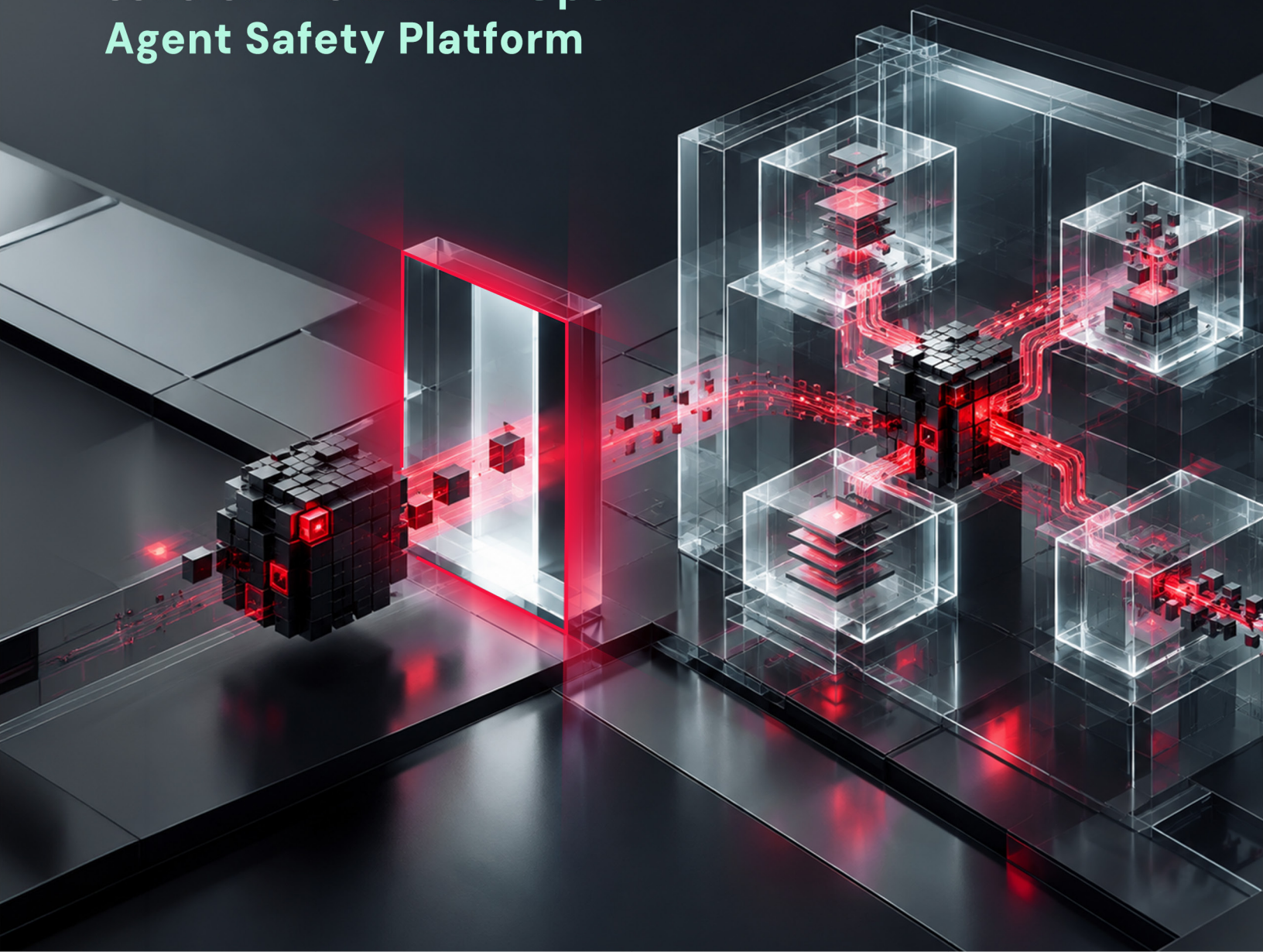




Permission Is Not Inspection

Verifying what autonomous agents run, install, load, and send on the **NVIDIA Open Agent Safety Platform**



Summary

Autonomous agents install packages, load models, adopt skills, and move files for hours at a time, with no reviewer in the loop. NVIDIA's Open Agent Safety Platform, announced September 28, 2026, controls what each agent is allowed to do. OpenShell enforces policy outside the agent. NVIDIA Sentry on BlueField-4 watches from outside the host.

The platform does not analyze the artifacts themselves. A rule that lets pip reach a registry says nothing about the package that comes back.

ReversingLabs adds that analysis. It returns one verdict per artifact, and the controls you already run enforce it: your image registry, your package mirror, OpenShell, and DOCA Argus.

You can deploy part of this now, with no code changes to NVIDIA components, and test the rest in your own environment.

ReversingLabs Solutions for NVIDIA Open Agent Safety Platform	
	
DEPLOY NOW Works with the platform as it ships Image and skill admission, a verified package and model mirror, log enrichment, and DOCA Argus allowlists. Shipping RL products. Configuration only on the NVIDIA side.	TEST IN YOUR ENVIRONMENT Pilot inline inspection Inline checks on uploads and downloads, review of new network rules, and policy guardrails at the gateway. A pilot you run with RL on your own agent traffic.

Agents removed the reviewer

Software used to pass a checkpoint before it ran: a build pipeline, a change review, a person who signed off. Agents skip that step. An agent working toward a goal installs what it needs, when it needs it.

NVIDIA moved authority into the runtime

NVIDIA's AI safety and security teams argue that prompts, model safeguards, and harness logic can shape what an agent tries, but can't set a hard limit on what it does. In NVIDIA's words: "The harness guides what an agent tries; the infrastructure determines what an agent can do. Both are necessary; only one is authoritative."

The platform puts that authority in three layers.

The NVIDIA Open Agent Safety Platform	
Layer	What it does
NVIDIA OpenShell	Open-source secure runtime. Per-agent sandboxes, a policy engine that checks every action by binary, destination, method, and path, and a gateway for policy, credentials, and inference routing. The September 2026 release adds multi-tenancy, Kubernetes high availability, extensibility APIs, centralized logs, and policy advising.
NVIDIA Sentry on BlueField-4	Out-of-band monitoring and policy enforcement in silicon, built on DOCA Argus, Flow, and Vault and on BlueField-4 Astra. During agent testing and evaluation, it can quarantine agents in milliseconds.
NVIDIA Vera CPU	Arm-based compute for agent workloads. NVIDIA says the platform also runs on other hardware.

What the runtime can't see

OpenShell knows which binary makes a request, where it goes, and which credential it carries. It doesn't know whether the payload is safe. That is by design. NVIDIA says OpenShell works alongside an enterprise's security tooling rather than replacing it, and deciding whether a binary, package, or model is malicious is that tooling's job.

“ The harness guides what an agent tries;
the infrastructure determines what an agent can do.
Both are necessary; only one is authoritative. ”



What gets through a valid rule

Take the rule “pip may reach the package registry on port 443.” It is narrow and easy to audit. A typosquatted package or a compromised release comes through it just the same. Six kinds of artifact cross the agent boundary this way.

What NVIDIA controls and what stays unchecked			
Artifact	NVIDIA controls	Stays unchecked	RL control
Sandbox images	Which image a sandbox starts with	What the binaries in each layer do	Spectra Assure
Agent skills	Whether the agent may install one	The scripts and binaries inside the bundle	Spectra Assure
Packages (pip, npm)	Which registries are reachable	Whether a package is malicious or tampered	Verified mirror
Model files	Which model hubs are reachable	Code hidden in Pickle and similar formats	Spectra Assure, verified mirror
Downloads	Which hosts may respond	What the file is and whether it's safe	Spectra Detect Insight
Uploads	Which destinations are allowed	Executables or sensitive files in the body	Spectra Detect Insight

Verified mirror: your internal copy of the package registries and model hubs agents use. It holds only packages and models that Spectra Assure has approved. Sandbox network policy sends pip, npm, and model downloads to the mirror instead of the public sources.

Signatures prove origin, not behavior

Signing, SBOMs, and build provenance show who built an artifact and how. NVIDIA signs the skills and models it publishes. But a binary signed by a compromised pipeline is still signed, and an SBOM won't report that a component opens a reverse shell. Most of what agents install also comes from outside NVIDIA's catalog.

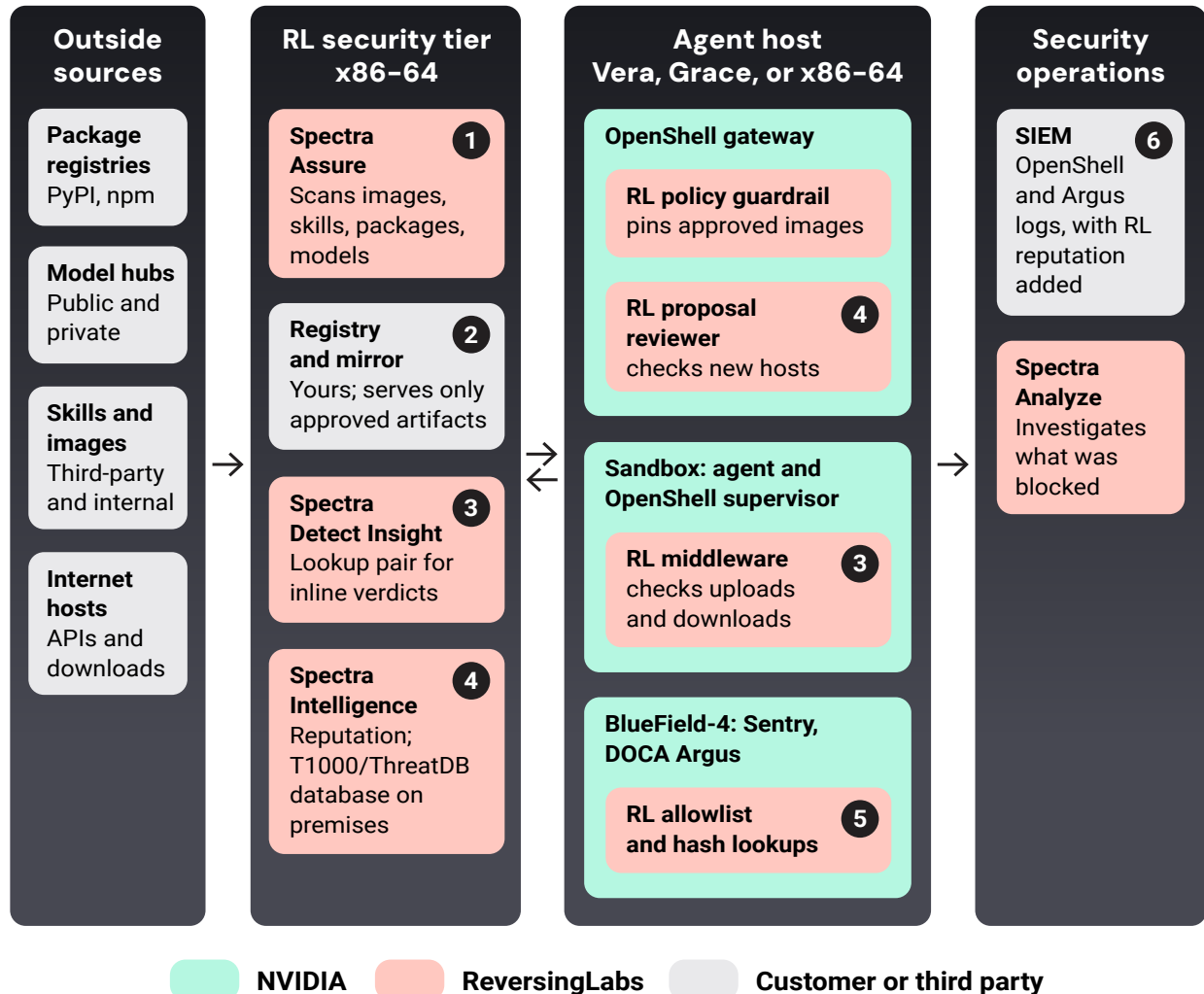
Auto-approval opens new hosts

When an agent is blocked, OpenShell lets it propose a new network rule. A policy prover checks the rule for new credential exposure and risky destinations. If the operator turns on auto-approval, rules with no findings pass, within approved policy upper limits. NVIDIA's documentation notes that a new public host with no credential attached raises no finding. Nobody checks that host's reputation or the files it serves.

How it fits together

Analysis runs on an x86-64 security tier that you operate. Verdicts are enforced where NVIDIA already enforces policy. No RL component parses untrusted files inside an NVIDIA process.

Deployment Architecture



1 Admit

Spectra Assure scans every image, skill, package, and model before first use.

2 Serve

Sandboxes pull only from your registry and mirror, which hold approved artifacts.

3 Inspect

OpenShell sends uploads and downloads to Spectra Detect Insight before they pass.

4 Review

New network rules are checked against Spectra Intelligence before approval.

5 Watch

DOCA Argus flags any running binary that RL didn't approve.

6 Record

Every hash, URL, and IP in your logs gets an RL reputation.

The ReversingLabs components

What each product does and where it runs

Product	Job	Runs on
Spectra Assure, with Catalog	Scans images, skills, packages, and models before use. Produces a SAFE report and xBOM.	x86-64 Linux or Windows; Arm through emulation; or SaaS
Spectra Detect Insight (formerly File Inspection Engine)	Inline verdicts: hash lookup, file type, type policy, and unpacking. Never uploads files.	x86-64 container or Kubernetes; Arm through emulation
Spectra Intelligence	Reputation for file hashes, URLs, and IPs.	Cloud, or the T1000/ThreatDB database on premises
Spectra Analyze	Analyst workbench for blocked artifacts and alerts.	On premises or cloud

Deploy now

These five integrations use shipping RL products. On the NVIDIA side, they need configuration only.

Day-one integrations

Integration	What you set up	Result
Image and skill admission	Spectra Assure scans images and skills in your build pipeline. Approved ones go to the registry OpenShell pulls from, pinned by hash so they can't be swapped.	Only scanned, approved images and skills start.
Verified mirror	An internal repository gated by Spectra Assure. Sandbox policy allows only the mirror for pip, npm, and model hubs.	Failing packages and models never reach an agent.
Log enrichment	Forward OpenShell logs to your SIEM and add Spectra Intelligence lookups.	Reputation on every hash, URL, and IP, with no added latency.
Argus allowlist	An RL service loads Assure-approved hashes into the DOCA Argus allowlist.	Any other running binary is flagged.
Argus hash lookups	Route Argus events through Fluent Bit or Vector to an RL service with a local cache.	A verdict for every hash Argus reports.

Test inline inspection in your environment

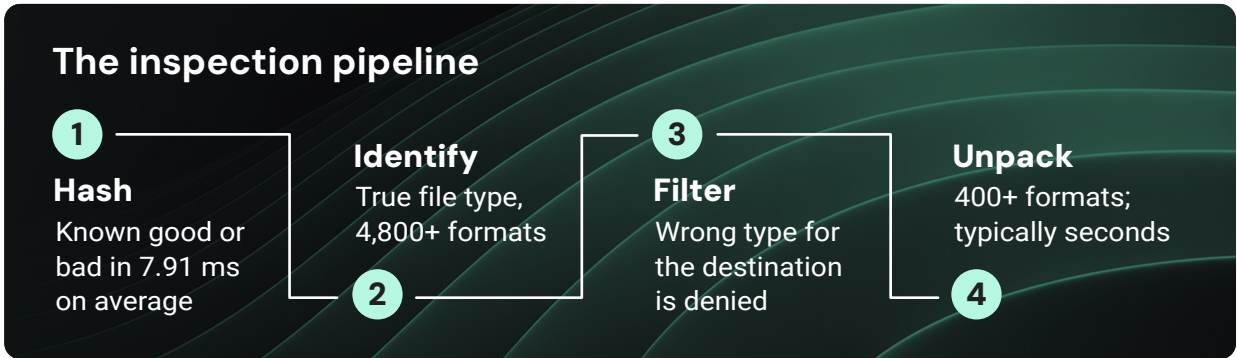
You run the pilot with RL engineers on your own agent traffic. It adds the checks below and measures their latency before you roll them out.

RL connects to OpenShell in two places. The gateway approves new sandboxes and policy changes. The supervisor sees each sandbox’s traffic. NVIDIA calls this OpenShell Extensibility. It also says the API will change between releases, so each pilot pins an OpenShell version.

Pilot integrations		
Integration	Runs at	What it does
Policy guardrail	Gateway	Pins each sandbox to an approved image. Rejects policy changes that remove or weaken RL checks.
Upload checks	Supervisor	Checks outbound files before OpenShell adds credentials, so RL never sees them. An executable sent to a document API is denied.
Download checks	Supervisor	Checks responses up to 4 MiB before the agent gets them. Larger, compressed, or partial files can’t be checked inline, so packages and models come from the verified mirror instead.
Proposal review	RL service	Keeps approval on manual. Approves proposals whose host passes reputation checks and sends the rest to a person.

Cheapest check first

Most files are decided by hash or file type in milliseconds. Only unknown files that policy allows are unpacked. RL registers two paths: a fast path within OpenShell’s 500 ms default timeout, and a deep path with a 2 to 5 second timeout for destinations that must wait for full analysis.



Design targets: under 20 ms median on the fast path, under 2 seconds at p99 for 4 MiB files on the deep path. The pilot measures both.

What the pilot measures

- How much of your agent traffic hash and file type decide on their own
- Tail latency on your hardware
- How much unpacking the rest requires
- Which artifacts your agents already pulled that nobody inspected

Performance and sizing

Measured hash-lookup performance, one container				
Test	Requests/sec	Success	Average	Maximum
Baseline, 60 s	1,212.80	100%	7.91 ms	833 ms
Sustained, 10 min	4,637.67	100%	46.10 ms	2,259 ms
Peak, 3 min	10,737.48	99.91%	73.99 ms	641 ms

Source: Spectra Detect Insight data sheet, rev. 09.14.26. Single x86-64 container (8 CPUs, 16 GB RAM, 35 GB disk), load generator on the same host. Arm results are not published.

- **Volume is rarely the limit.** One container sustained more than 4,600 lookups per second. Deploy a pair for availability.
- **Size for the slowest lookups, not the average.** Maximum latency passed OpenShell's 500 ms timeout in every test. Inline traffic gets dedicated replicas, and the caller times out first.
- **Plan for storage.** Each replica keeps a 20 to 40 GiB threat database. A pair needs roughly 84 to 164 GiB of fast storage, including room for updates.

Running next to Arm hosts

Vera, Grace, and BlueField use Arm cores. The Spectra Assure CLI and Spectra Detect Insight both run on Arm through x86 emulation. Spectra Detect Insight is confirmed on AWS Graviton but not yet on Vera or Grace.

The recommended split keeps hash lookup and filtering close to the agent and does unpacking on the x86-64 tier. All components support air-gapped sites. A native Arm build is on the RL roadmap, without a committed date.

Who owns what

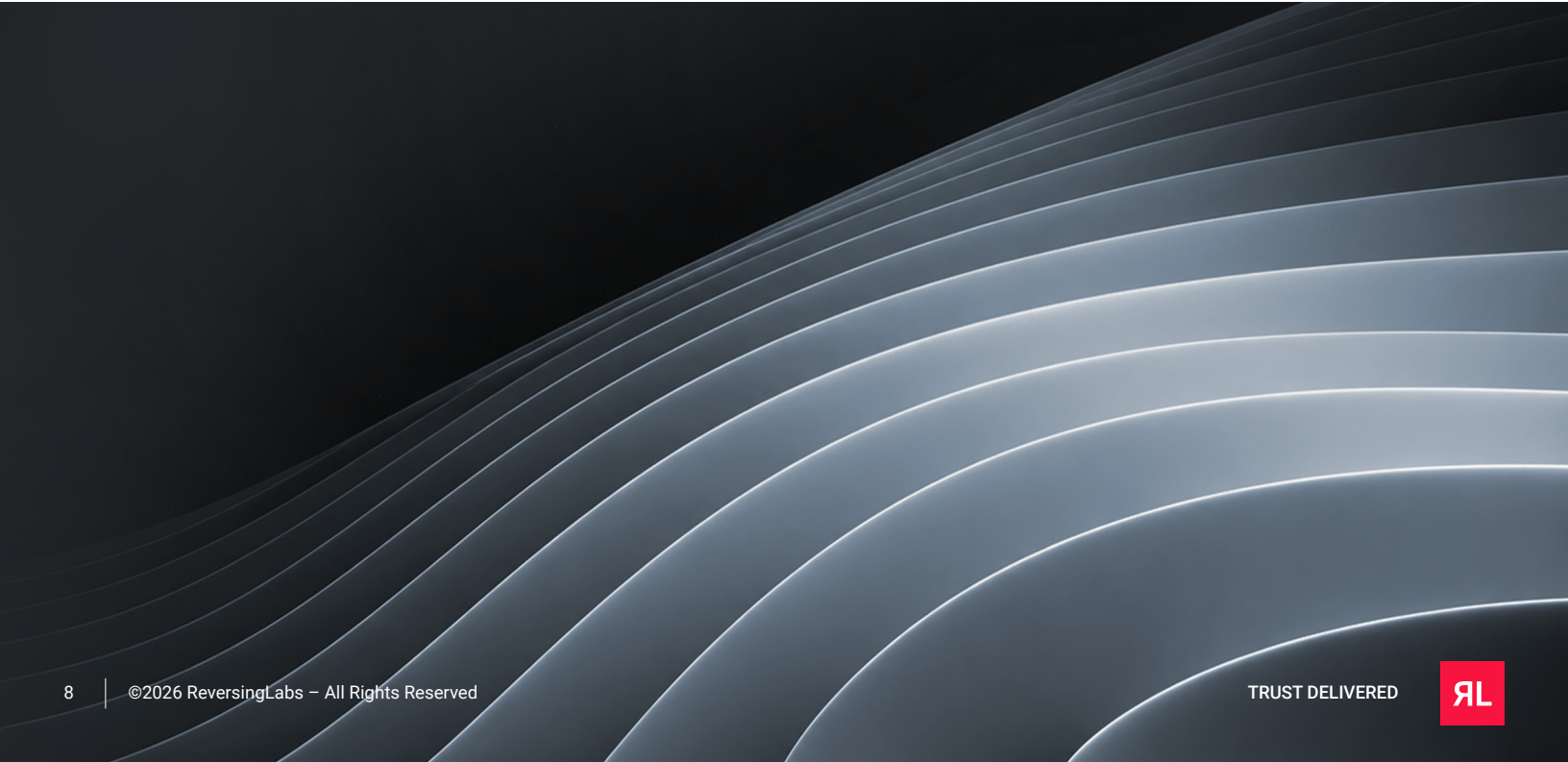
NVIDIA leaves decisions about authority, approval, and response to the enterprise. RL gives each owner evidence for theirs.

Owners and the evidence they get	
Owner	What RL gives them
Application owners	Approved images, skills, and packages to build with
Security architects	Verdicts and reputation behind every approval
Platform teams	Admission control, the verified mirror, and inline checks
Security operations	Enriched logs and Spectra Analyze for investigation

Conclusion

NVIDIA’s platform decides what an agent may do, and agents can’t bypass it. Whether the images, skills, packages, models, and files an agent uses are safe is a separate question. Only analysis of the artifacts can answer it.

ReversingLabs supplies that analysis as one verdict per artifact, enforced by the controls you already run. Start with admission, the verified mirror, and log enrichment. Then pilot inline checks in your environment to measure latency on your own traffic. The same verdicts extend to NVIDIA NIM, NeMo Platform, and NeMo Guardrails; the technical research paper covers those.



Sources

- NVIDIA, NVIDIA Open Agent Safety Platform announcement, September 28, 2026.
- NVIDIA, NVIDIA Open Agent Safety Platform, September 2026 announcement, partner messaging document.
- NVIDIA Technical Blog, "NVIDIA Open Agent Safety Platform: A Reference for Continuous In-Silicon Agent Monitoring," September 28, 2026.
- NVIDIA Technical Blog, "Where Security Fits in an AI Agent Stack," August 21, 2026.
- NVIDIA Technical Blog, "NVIDIA-Verified Agent Skills Provide Capability Governance for AI Agents," May 19, 2026.
- NVIDIA Technical Blog, "Run Autonomous, Self-Evolving Agents More Safely with NVIDIA OpenShell," March 16, 2026.
- NVIDIA OpenShell documentation, v0.1.2: Supervisor Middleware, Configuration, and Supported Operations; Policy Advisor; Extensibility overview; Release Notes; Kubernetes High Availability. docs.nvidia.com/openshell
- NVIDIA OpenShell repository. github.com/NVIDIA/OpenShell
- NVIDIA DOCA Argus service guide, DOCA 3.5. networking-docs.nvidia.com/doca
- NVIDIA Technical Blog, "Advancing AI Infrastructure for Agentic AI with NVIDIA DOCA In-Silicon Security." developer.nvidia.com/blog
- NVIDIA, "NVIDIA Vera BlueField-4 STX Brings Agentic AI Storage Processing With In-Silicon Security," press release, May 31, 2026.
- NVIDIA Technical Blog, "Redefining Secure AI Infrastructure with NVIDIA BlueField Astra for NVIDIA Vera Rubin NVL72," January 2026.
- NVIDIA BlueField-4 announcement and specifications.
- NVIDIA, Vera CPU details, GTC 2026 (March 2026).
- ReversingLabs, Spectra Detect Insight data sheet, rev. 09.14.26.
- ReversingLabs, Integration Architecture for the AI Factory, technical research paper, September 2026 (available on request).

Trust, but decompile.

Talk to us about a pilot in your agent environment. A detailed technical research paper is available on request.

REQUEST A DEMO

reversinglabs.com

About ReversingLabs

ReversingLabs is the trusted name in file and software security. We provide the modern cybersecurity platform to verify and deliver safe binaries. Trusted by the Fortune 500 and leading cybersecurity vendors, RL Spectra Core powers the software supply chain and file security insights, tracking over 422 billion searchable files with the ability to deconstruct full software binaries in seconds to minutes.