STAFFINITY

Agent Safety Standard

Adoption of the NVIDIA Open Agent Safety Platform Five Principles

Version 1.0 | September 28, 2026 | Internal standard and external positioning

1. Purpose

Staffinity builds and operates autonomous AI agents that work inside client business systems — reading and writing SAP data, moving documents, and communicating with people. That authority demands enforceable boundaries. This document formally adopts the five agent-safety principles published by NVIDIA as part of the Open Agent Safety Platform (announced September 28, 2026) as the Staffinity Agent Safety Standard, maps each principle to the controls already operating in the Staffinity agent platform, and sets the roadmap toward formal participation in the Open Agent Safety Platform ecosystem and the Open Secure AI Alliance.

2. Background: The NVIDIA Open Agent Safety Platform

NVIDIA's Open Agent Safety Platform is an open software platform and reference system design for securing AI agents from testing to deployment. It has two components:

- **OpenShell —** an open-source secure runtime that executes agents in sandboxed environments with kernel-level isolation, declarative YAML policies, formally verified policy changes, and brokered credentials the agent never sees.
- **Sentry —** a reference design for an out-of-band hardware watchdog (BlueField DPU) that monitors agent behavior and quarantines drift in milliseconds, invisible to the agent itself.

More than 100 organizations — including Anthropic, Microsoft, SAP, Salesforce, ServiceNow, CrowdStrike, and Palo Alto Networks — are collaborating on the platform. The platform's thesis matches the architecture Staffinity has operated since its first client deployment: an agent cannot be expected to govern its own behavior, so enforcement must live outside the agent.

3. The Five Principles, Adopted

Staffinity adopts all five principles as written. The table below states each principle and the Staffinity control that implements it today.

NVIDIA principle	Staffinity implementation (operating today)
1. Policies must be verifiable	Declarative, reviewable policy: SAP table whitelist, blocked-table list, and SQL-construct preflight rules are configuration, version-

	controlled and auditable. The independent judge verifies every factual claim in an agent reply against the underlying SAP rows before delivery; unverifiable claims fail closed and are never sent.
2. Enforcement must live outside the agent	The judge, the SQL preflight guard, and the tool sandbox run outside the agent's own reasoning loop. The agent cannot disable, argue with, or route around them. Human approval gates sit outside the agent for every external action (email, posting, writes).
3. Control the path to the model	Agents reach model providers only through the platform's brokered channel with operator-controlled provider allowlists. Tool processes never hold model credentials; clients' SAP credentials live in AWS Secrets Manager and are injected per-task, never exposed to the agent's context.
4. Scale authority only with inspectability	Every tool call is logged with structured fields (userId, conversationId, tool, table) to CloudWatch. A failure-signature monitor watches hourly for retry loops, fail-closed cascades, promised-but-not-executed actions, and safety-net catches — first in shadow mode, then enforced. New capabilities ship to shadow before they ship to production.
5. Shared responsibility across the ecosystem	Safety is a contract with clients: onboarding includes a permission review, agents are scoped to least-privilege SAP access, and every honest failure is communicated as such — never papered over. Staffinity commits to contributing findings and engineering work back to the Open Secure AI Alliance and the Shared AI Findings Exchange (SAFE).

4. Operating Evidence

These controls are not aspirational. On the morning of September 28, 2026, the production failure-signature monitor flagged a retry loop and three judge blocks in a single client conversation. Triage showed the enforcement layer working exactly as the principles require:

- **Preflight policy:** six mutated SQL attempts against a sandbox that rejects CTEs, blocked tables, and correlated subqueries — every one stopped before execution;
- **Fail-closed enforcement:** three independent judge blocks of unverifiable claims (an alphanumeric vendor code typed as a number, a database name typed as a name, a user-echoed date range) — none delivered to the user;
- **Honest failure:** one promised-but-not-executed action caught and communicated honestly — the user was told the email was not sent, rather than being told it was.

No unverified figure, fabricated claim, or silent failure reached the client at any point.

5. Roadmap to Ecosystem Participation

Phase	Milestone	Target
1 — Adopt	Adopt the five principles as the Staffinity Agent Safety Standard (this document). Publish the alignment internally and in client-facing materials.	Complete — Sept 2026
2 — Evaluate	Technical evaluation of OpenShell against the Staffinity agent runtime: policy model mapping, sandbox isolation for tool execution, and use of OpenShell for Staffinity's own development/coding agents as a low-risk first deployment.	Q4 2026
3 — Contribute	Contribute engineering work and operational findings: judge/enforcement patterns, failure-signature taxonomy, and OpenShell issues/PRs. Submit findings to the Shared AI Findings Exchange (SAFE).	Q1–Q2 2027
4 — Recognize	Formal participation: Open Secure AI Alliance membership (Linux Foundation), NVIDIA Inception / Partner Network application, and co-developed reference material with ecosystem partners (SAP practice alignment).	2027

6. Standing Commitments

- **1.** Enforcement outside the agent is non-negotiable: no client agent ships without an independent verification and policy layer.
- **2.** Fail closed, always: when verification cannot complete, the safe message goes out — never the unverified one.
- **3.** New agent authority is earned in shadow mode before it is exercised in production.
- **4.** Every client can see what their agent did: structured logs, audit trails, and honest failure messages are part of the product, not an add-on.
- **5.** Findings that make the ecosystem safer get shared — through SAFE, the Open Secure AI Alliance, and upstream contributions.

Approved: Brent Sudeck, Founder — Staffinity
Review cycle: quarterly, or upon material change to the NVIDIA Open Agent Safety Platform.