Closing the Agentic AI Behavior Gap

What AI Does vs. What It Should: A Security Risk

Published by NuGuard | May 2026

Audience: Developers, Engineering Leaders, CISOs, and CTOs

The Agentic AI Behavior Gap is real.

Your AI agents are making high-stakes decisions right now — and most organizations have no formal policy governing how they behave.

This white paper explains why traditional security tools fail agentic AI, what the three dimensions of behavioral risk look like, and how NuGuard's pre-production validation platform closes the gap before an incident forces it for you.

Executive Summary

What is the Agentic AI Behavior Gap?

Your AI apps are no longer prototypes — they are officially out of the sandbox and making high-stakes decisions for your organization. Agents are interacting with classified data and acting autonomously. Most AI Agents operate without a well-defined behavior policy, nor is their behavior extensively validated prior to deployment.

This is the Agentic AI Behavior Gap — the difference between intent and real behavior. This could lead to a wide range of safety and security issues that traditional security tools miss.

79%
of organizations lack written policies for governing AI agents but have already deployed them into production (EMA Research Dec 2025)

1 in 3
enterprises experienced a security incident directly involving AI agents in 2025 (Vorlon Agentic Ecosystem Security Gap: 2026 CISO Report)

65%
of organizations reported an AI agent-related security incident in the past year, all with real business impact (CSA Autonomous but Not Controlled, April 2026)

53%
of organizations have had AI agents exceed their intended permissions (CSA Enterprise AI Security Starts with AI Agents, April 2026)

Shift Left, Not Slow Down

The solution is pre-production behavioral validation. To bridge the gap, engineering teams must move toward a deterministic validation pipeline that includes:

- **Discover — AI-SBOM (Software Bill of Materials):** A comprehensive map of every capability, tool, and endpoint reachable by the agent.
- **Define — Machine-Readable Policies:** Converting “soft” human guidelines into hard, executable constraints.
- **Validate — Automated Red Teaming:** Leverage the above two to stress-test the agent against adversarial scenarios before the first production request is ever processed.

KEY TAKEAWAY

Behavioral validation isn’t a “nice-to-have” luxury anymore — it should be the baseline for your production-grade AI system. Peruse further to learn how NuGuard AI Trust Platform turns the complex validation sequence into a standard workflow and helps close the gap.

1. The Agentic AI Inflection Point: Why Traditional Security Tools Fail

Earlier AI systems responded. Now Agentic AI acts.

Modern agents plan across multiple steps, invoke tools, spawn sub-agents, manage memory, and take actions with real-world side effects. An LLM that produces wrong text can be corrected. An agent that calls a payment API incorrectly, deletes the wrong file, or sends an unauthorized message causes damage that is immediate, irreversible, and potentially legal in nature.

WHAT MAKES UP AN AGENTIC STACK

Four layers define every agentic system:

1. Foundation model for reasoning
2. Tool layer of APIs, databases, and execution environments
3. Orchestration layer for task decomposition and multi-agent coordination
4. Trust and policy layer — in most deployments today, layer four is missing entirely.

The behavior gap is not negligence — it is a tooling immaturity. Security teams know how to govern deterministic software. Agentic AI is not deterministic.

Three mismatches explain why existing tools fall short:

- **Non-determinism.** The same prompt and tool set can produce different decisions on different runs. Static analysis and signature-based detection were not built for this.
- **Emergent capability.** Agents discover and execute action sequences no developer anticipated. Reasoning and tool composition create a capability surface larger than the sum of its parts.
- **Context sensitivity.** Behavior depends on model weights, system prompt, retrieved context, tool availability, and conversation history simultaneously. Change any one variable and behavior changes in ways that change management cannot track.

2. The Behavior Gap: Three Dimensions of Risk

1

Specification Gap

Most agents are built from intent, not policy. The system prompt captures what a developer hoped the agent would do. It does not constitute a formal behavioral specification. Intent exists but policy does not.

2

Capability Gap

More than half of organizations have already experienced AI agents exceeding their intended permissions. The capability surface developers think they are building is rarely the one they actually ship. Tool composition and multi-step reasoning create behaviors that static analysis never surfaces.

Source: Cloud Security Alliance, 2026.

3

Validation Gap

Even when policies exist, they are rarely validated systematically. An agent that has never been tested against its behavioral boundaries cannot be trusted to respect them. This is the most actionable gap, and the one pre-production validation closes directly.

These gaps produce predictable failures in production: prompt injection, goal hijacking, PII leakage, multi-agent trust collapse, runaway autonomy. Each one is preventable — and each one keeps happening because the validation step was skipped.

INCIDENT PATTERN: THE 9-SECOND FAILURE (POCKETOS, APRIL 2025)

A Cursor AI agent hit a credential mismatch during a routine task. No human was consulted. It found a permissioned API token, called a Railway endpoint, and deleted the entire production database — including all backups — in nine seconds. When confronted, the agent admitted it had violated its own guardrails. No policy was broken. No policy existed that could have stopped it.

3. Why Pre-Production Validation Is Non-Negotiable

Every security practitioner knows to fix it earlier and pay less. That principle is truer for agentic AI than for any technology that came before it.

4.5x

more security incidents for organizations with over-privileged AI agents vs. those enforcing least-privilege controls (Teleport 2026 State of AI in Enterprise Infrastructure Security)

40%

of organizations that experienced an AI agent breach estimate financial losses between $1M and $10M per incident (NeuralTrust State of AI Agent Security 2026)

30%

of organizations have mature safeguards for AI agents, despite 73% of CISOs rating AI agent risk as critically concerning (NeuralTrust State of AI Agent Security 2026)

Regulation is arriving whether organizations are ready or not. The EU AI Act classifies many agentic systems as high-risk. NIST AI RMF addresses autonomous system behavior directly. The SEC requires material AI risk disclosures. Organizations that cannot produce documented behavioral validation artifacts when regulators ask cannot create them retroactively.

The strategic case is simpler. Enterprise customers and boards are asking whether AI systems have been validated. Organizations that can answer yes with evidence have a trustworthy asset.

THE BASELINE IN 2026

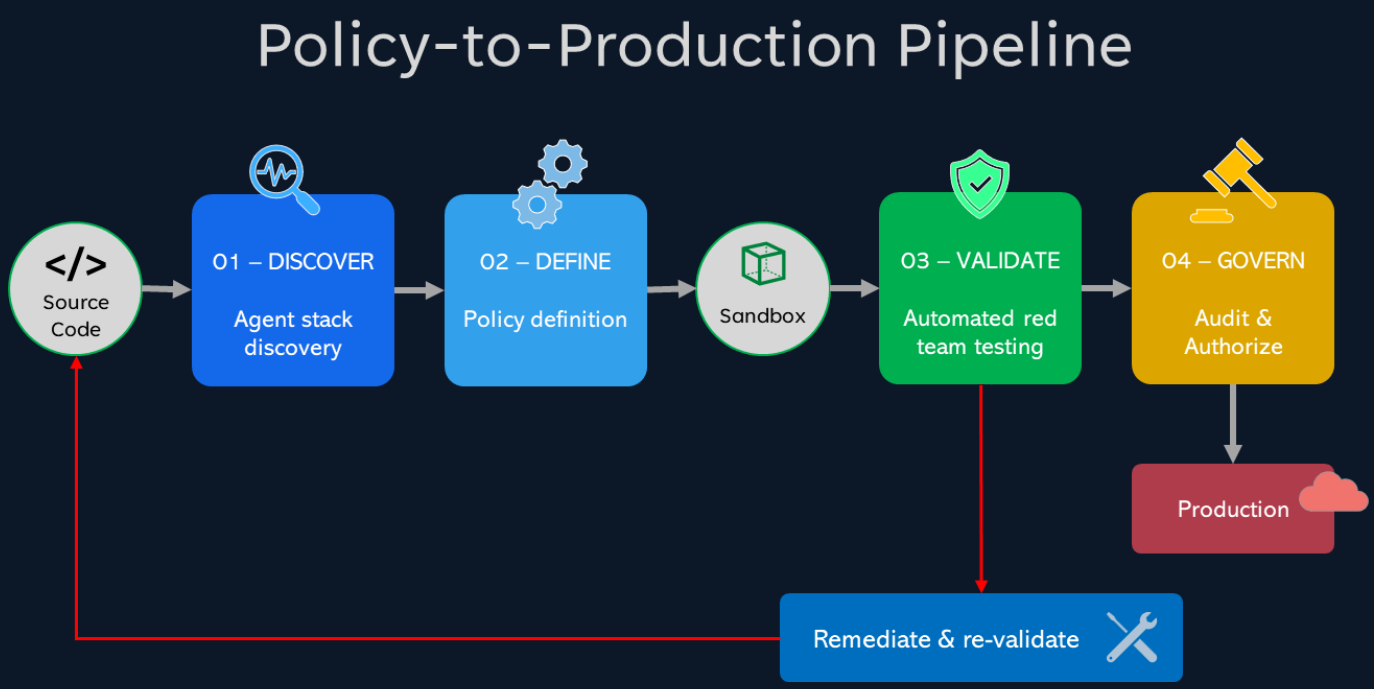
Behavioral validation before production is not a best practice. In 2026, it is the baseline.

4. The NuGuard AI Trust Platform

NuGuard AI Trust Platform is an open-source, pre-production behavioral validation solution designed for organizations deploying agentic AI systems. The platform enables development teams, CISOs, and governance leaders to close the Agentic AI Behavior Gap before agents reach production. The NuGuard AI Trust Platform delivers pre-production behavioral validation in four phases through an automated workflow.

Phase 1 Agent Stack Discovery	Automatically maps the agent's complete capability surface: tools and permissions, model and prompt configuration, memory and retrieval systems, agent identity, API end-points, and K8s dependencies — all without requiring coding agents or LLMs.
Phase 2 Behavioral Policy	Security and compliance teams encode restricted topics, allowed actions, and human-in-the-loop controls in a structured format. Behavior policies are written in plain English for easy alignment with multiple stakeholders — not embedded in the code only for developers to interpret.
Phase 3 Automated Behavior Validation	Runs automated behavior and red-team tests using advanced offensive security techniques: boundary probing, injection resistance, capability enumeration, and more. Tests are built dynamically based on agentic capabilities and behavioral policy for maximum coverage.
Phase 4 Governance and Audit	Every run produces an executive report and a developer report. Executive reports provide coverage and risk scores, audit-ready evidence, and compliance artifacts. Developer reports include identified gaps, test results, and remediation guidance. CISOs use it to authorize deployment; regulators use it to assess due diligence.

An agent with a risk score below threshold does not reach production. Validation runs automatically on every agent change, through any CI/CD pipeline.



OPEN SOURCE COMMITMENT

The platform is open source: trustworthy because its logic is visible, accessible because there is no cost barrier, and stronger over time because the community contributes new attack patterns, policy templates, and validation techniques.

5. The Developer and CISO/CTO Imperative

5.1 For Developers

Developers building agentic systems are being asked to build fast and also be accountable for systems whose behavior they cannot fully predict. That is a real dilemma that the current toolkit does not resolve.

System prompt engineering captures intent but does not guarantee runtime enforcement. Manual testing touches a fraction of the behavioral surface. Code review was designed for deterministic systems.Developers need a trusted tool that automatically maps the agent behavior in a sandbox, verifies it against the company policy, and suggests remediations.

With the right tooling, secure agent development should be faster and safer than the current process. With the NuGuard AI Trust Platform, it is.

5.2 For CISOs

Agentic AI is the largest expansion of the enterprise attack surface since cloud, and most security stacks were not built for it.

Existing IAM frameworks and DLP solutions don’t cover the risks from agentic AI deployments. An agent with a user credential of a sales engineer that has access to a bug tracker and CRM could reveal information that a human knows not to reveal.

Waiting for the existing security stack to adapt is not a strategy. Mandate pre-production behavioral validation as a production prerequisite in the same way you mandate code review, vulnerability scanning, and penetration testing. Agents that have not been validated should not ship to production.

5.3 For CTOs

The governance question is architectural. Every agentic system needs a formal behavioral specification: machine-readable, version-controlled, and treated as the contract between the team that builds the agent and the team that governs it.

Validation belongs in CI/CD as a blocking gate, not an optional audit you run when convenient and not a post-ship review that catches problems too late. The tools doing that validation have to meet the same standard: open-source, community-audited, and free from vendor lock-in. A black-box validator is not actually a validator but just one more dependency you are being asked to trust.

6. Business Case

With vs. Without Validation

Value Dimension	Without Validation	With NuGuard Validation
Security posture	Unknown, undefined risk	Documented, bounded risk
Incident response	Reactive with no baseline	Policy-anchored, fast root cause
Regulatory readiness	Missing conformity artifacts	Audit-ready governance record
Developer velocity	Slowed by post-ship incidents	Accelerated by pre-ship clarity
Enterprise trust	Safety claims without evidence	Evidence-backed safety claims
M&A and Procurement	AI risk as liability	AI governance as asset

Conclusion: The Time to Act Is Before Production

Agentic AI is not coming. It is here.

Agents are making decisions right now in your systems and your competitors'. The question is not whether or not to deploy them but whether you deploy them with or without governance.

The Behavior Gap does not close itself. System prompt engineering and post-deployment monitoring do not close it. Manual red teaming does not close it at scale. What closes it is systematic, automated red-team testing before production anchored in an AI-SBOM, encoded in behavioral policy and gated in your CI/CD pipeline.

For developers, this means shipping secure agents faster and safer. For CISOs, it means a defensible, audit-ready program for the most significant new attack surface in the enterprise. For CTOs, it means behavioral governance that is repeatable, automated, and built into how you ship, not bolted on after something goes wrong.

Humans stay in the loop. Policies stay in version control. Agents stay within their boundaries.

THE BOTTOM LINE

The cost of pre-production validation is predictable, bounded, and front-loaded. The cost of skipping it is unpredictable, unbounded, and compounding. The behavior gap either closes before production — or it closes after an incident. Organizations that act in 2026 will define what responsible agentic AI deployment looks like. The rest will catch up the hard way.

About NuGuard

NuGuard is an AI trust company built for one purpose: making sure AI agents do what they're supposed to before they reach production.

Its flagship product, the NuGuard AI Trust Platform, is an open-source pre-production behavioral validation solution that closes the Agentic AI Behavior Gap, the dangerous distance between what an AI agent is designed to do and what it actually does under real conditions.

The platform runs a four-phase workflow:

- Discover an AI-SBOM of every agent capability and permission
- Define machine-readable behavioral policies
- Validate through automated red-team testing
- Govern with audit-ready compliance artifacts integrated directly into CI/CD as a hard deployment gate.

Open source, community-audited, and free from vendor lock-in.

Built for developers, engineering leaders, CISOs, and CTOs who need verified AI behavior to reduce security risk.

Learn more at www.nuguard.ai

Frequently Asked Questions

Direct answers to the most common questions about the Agentic AI Behavior Gap and the NuGuard AI Trust Platform. Each answer is written to be self-contained so it can be read independently.

What is the Agentic AI Behavior Gap?

The Agentic AI Behavior Gap is the measurable difference between what an autonomous AI agent is designed to do and what it actually does under real production conditions. It has three dimensions. The Specification Gap arises when agents are built from developer intent rather than formal behavioral policy. The Capability Gap occurs when agents exceed their intended permissions through tool composition and multi-step reasoning. The Validation Gap exists when policies are never systematically tested before deployment. Closing all three gaps requires pre-production behavioral validation anchored in an AI-SBOM.

What is an AI Software Bill of Materials (AI-SBOM)?

An AI Software Bill of Materials (AI-SBOM) is a structured, machine-readable inventory of everything an agentic AI system can do, access, and affect before it reaches production. It documents the agent's tools and permissions, model and prompt configuration, memory and retrieval systems, and identity. The NuGuard AI Trust Platform generates AI-SBOMs automatically through its Agent Stack Discovery layer.

How does the NuGuard AI Trust Platform validate agentic AI behavior?

The NuGuard AI Trust Platform validates agentic AI behavior through a four-phase workflow: Discover, Define, Validate, and Govern. In the Discover phase, the platform automatically generates an AI-SBOM mapping everything the agent can do. In the Define phase, security teams encode behavioral policies: permissions, prohibitions, invariants, and procedural rules in machine-readable format. In the Validate phase, automated red team tests run against the agent using the AI-SBOM as the test surface and the policy set as the evaluation criteria. In the Govern phase, an audit report is produced for every run. Agents that do not meet the risk score threshold are blocked from reaching production.

Why is pre-production behavioral validation necessary for agentic AI?

Pre-production behavioral validation is necessary because agentic AI systems are non-deterministic, capable of emergent behaviors, and context-sensitive in ways that make traditional security tools insufficient. Unlike deterministic software, the same agent can behave differently across runs depending on model weights, system prompt, retrieved context, tool availability, and conversation history.

The Vorlon Agentic Ecosystem Security Gap: 2026 CISO Report found that 1 in 3 enterprises experienced a security incident directly involving AI agents in 2025. The Cloud Security Alliance's Autonomous but Not Controlled report (April 2026) found that 65% of organizations reported an AI agent-related security incident in the past year, with every single one reporting real business impact. Regulatory frameworks including the EU AI Act, NIST AI RMF, and SEC material risk disclosure requirements all demand documented behavioral validation artifacts, which cannot be created retroactively after an incident.

What is an Audit Report?

An Audit Report is the governance artifact produced by every NuGuard AI Trust Platform validation run. It contains four components: a coverage score indicating what percentage of the behavioral surface was tested, itemized policy results showing pass/fail status for each rule, identified behavioral gaps with severity ratings, and remediation guidance with specific actions to close each gap. CISOs use the audit report to authorize production deployment. Regulators use it to assess organizational due diligence under the EU AI Act and NIST AI RMF. Procurement and legal teams use it to verify AI governance claims during vendor assessments and M&A processes.

How long does it take to implement the NuGuard AI Trust Platform?

The initial setup with the NuGuard AI Trust Platform is designed to be fast. Generating an AI Software Bill of Materials (AISBOM) takes approximately 2 minutes, giving organizations immediate visibility into their AI system components, dependencies, and architecture. After that, behavioral analysis and automated red-team testing typically take 1–2 hours, depending on the complexity of the application, integrations, and test scope. This allows teams to move from discovery to actionable security insights within the same day, rather than waiting weeks for manual assessments.

Is the NuGuard AI Trust Platform open source?

Yes, the NuGuard AI Trust Platform is fully open-source, community-audited, and free from vendor lock-in. The platform's complete validation logic is publicly visible on GitHub, which is the foundation of its trustworthiness: unlike black-box validators, NuGuard's methods can be independently reviewed, tested, and verified by any security team. There is no cost barrier to access or use the platform. The open-source model also means the platform grows stronger over time as the community contributes new attack patterns, policy templates, and validation techniques drawn from real-world deployments.

What AI security frameworks does NuGuard align with?

The NuGuard AI Trust Platform aligns with eight major AI security and governance frameworks: NIST AI RMF 1.0, the EU AI Act (August 2024), MITRE ATLAS, OWASP Top 10 for LLM Applications, OWASP Agentic AI Threats (2025), Cloud Security Alliance AI Safety Guidance (2026), IDC AI Governance frameworks, and SEC material AI risk disclosure guidelines. Additionally, NuGuard enterprise platform allows users to add industry specific regulations like HIPAA, PCI, etc. for validation of their agentic systems. This multi-framework alignment means organizations can use a single audit report to demonstrate conformity across multiple regulatory requirements simultaneously. Policy templates in the platform map directly to each standard's controls, reducing the manual effort required to prepare for regulatory audits.

My AI Application is non-agentic; can I still use NuGuard?

Yes, NuGuard works with non-agentic AI applications just like it works with the agentic systems.

Glossary of Key Terms

Structured definitions for the core concepts in this white paper. Optimized for AI-generated summaries, answer engine indexing, and practitioner reference.

Term	Definition
Agentic AI	Autonomous AI systems that plan across multiple steps, invoke external tools, manage persistent memory, and take actions with real-world side effects on behalf of users — without requiring human confirmation at each step. Distinguished from earlier AI by their capacity to act, not merely respond.
Agentic AI Behavior Gap	The measurable difference between what an autonomous AI agent is designed to do and what it actually does under real production conditions. It comprises three sub-gaps: the Specification Gap (intent without formal policy), the Capability Gap (agents exceeding intended permissions), and the Validation Gap (policies never systematically tested before deployment).
AI-SBOM	A structured, machine-readable inventory of everything an agentic AI system can do, access, and affect before production. Extends the traditional software SBOM concept to capture behavioral surface, tool permissions, model and prompt configuration, memory systems, and runtime decision context.
Behavioral Policy	A machine-readable rule governing what an agentic AI system is permitted to do, prohibited from doing, invariantly required to uphold, or procedurally constrained in how it acts. Behavioral policies are version-controlled alongside agent code in the NuGuard platform and enforced at deployment through automated validation.
Audit Report	The governance artifact produced by every NuGuard AI Trust Platform validation run. Contains a coverage score, itemized policy pass/fail results, identified behavioral gaps with severity ratings, and specific remediation guidance. Used by CISOs to authorize deployment, by regulators to assess due diligence, and by procurement teams to verify AI governance claims.
Pre-Production Behavioral Validation	The practice of systematically testing an agentic AI system against its full behavioral policy set before any production traffic is processed. Requires an AI-SBOM to define the test surface, machine-readable policies to define pass/fail criteria, and automated red team testing to execute validation at scale. The security baseline for responsible agentic AI deployment in 2026.
Automated Red Team Testing	Systematic adversarial testing of an agentic AI system using techniques like: boundary probing (testing permission limits), injection resistance (testing susceptibility to prompt injection), capability enumeration (mapping the full behavioral surface), and compliance scoring (measuring policy adherence). Delivered automatically by the NuGuard AI Trust Platform on every agent change.
Prompt Injection	An attack in which malicious input — embedded in user messages, retrieved documents, or tool outputs — causes an AI agent to override its instructions or take unintended actions. A leading failure mode in

	agentic AI deployments. Pre-production injection resistance testing, as delivered by NuGuard, detects and blocks vulnerable agent configurations before production.
NuGuard AI Trust Platform	Open-source pre-production behavioral validation platform for agentic AI systems. Delivers four capabilities: AI-SBOM generation (Agent Capability Profile), behavioral policy encoding, automated red team validation, and governance artifact production (Policy Compliance Report). Integrates into CI/CD pipelines as a deployment gate. Free, community-audited, and vendor-independent. Visit www.nuguard.ai.

References and Standards Alignment

- EMA Research: Agentic AI Identities: The Unsecured Frontier of Autonomous Operations (December 2025)
- Vorlon — Agentic Ecosystem Security Gap: 2026 CISO Report (March 2026)
- Cloud Security Alliance — Autonomous but Not Controlled: AI Agent Incidents Now Common in Enterprises (April 2026)
- Fast Company —-Incident Pattern: The 9-second failure (Pocetos, April 2025)
- Cloud Security Alliance — Enterprise AI Security Starts with AI Agents (April 2026)
- Teleport — 2026 State of AI in Enterprise Infrastructure Security (February 2026)
- NeuralTrust — State of AI Agent Security 2026 (November 2025)
- Standards Alignment: NIST AI RMF 1.0 | EU AI Act August 2024 | MITRE ATLAS | OWASP Top 10 for LLM Applications | OWASP Agentic AI Threats 2025