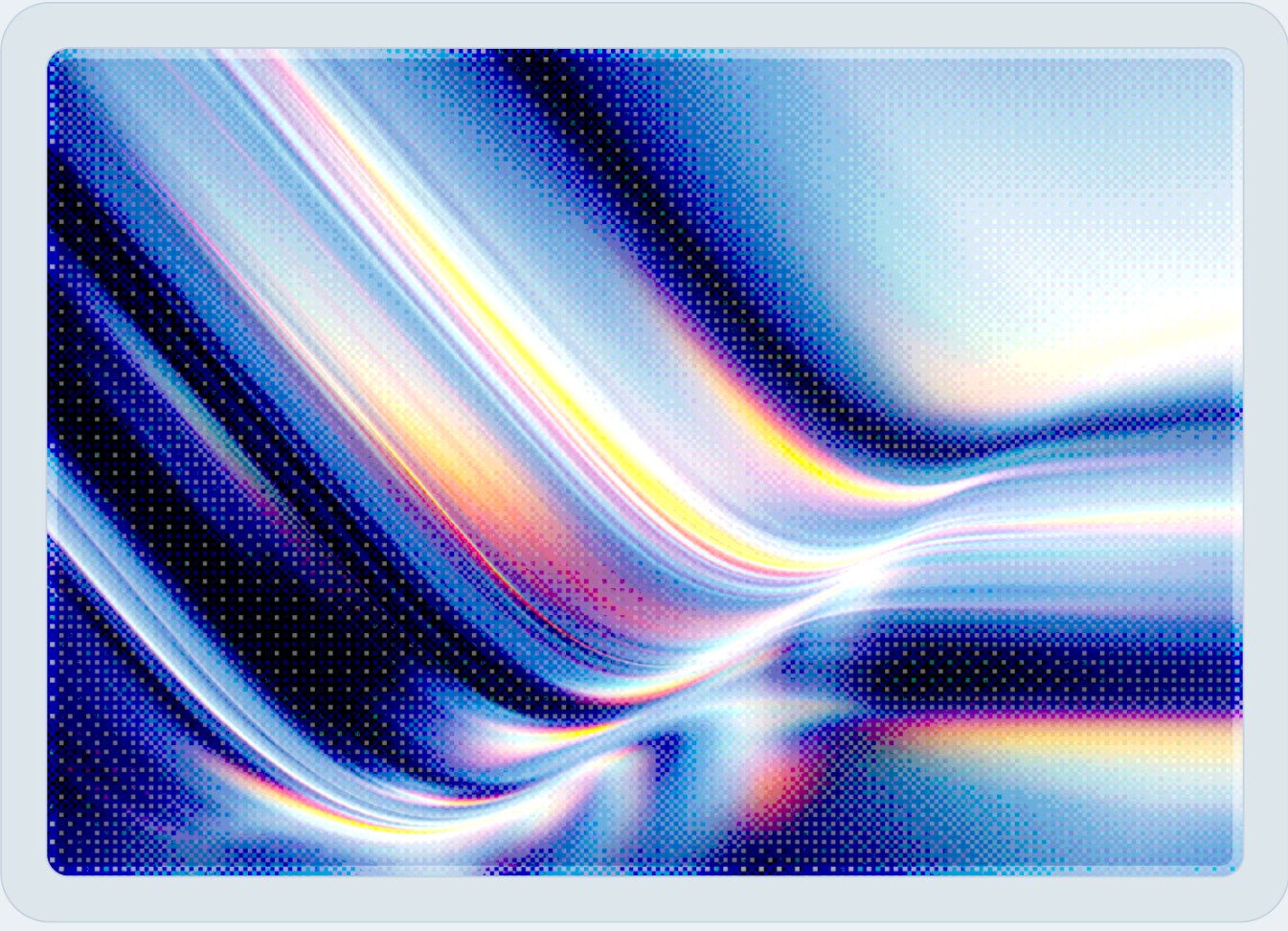
CSAI Foundation | Cloud Security Alliance

Static AI Guardrails: The NIST Incompleteness Proof

Why No Finite Rule Set Can Fully Secure AI Systems and What Governance Frameworks Must Require Instead

2026-06-30

AI-assisted Rapid Research



CSAI

cloud security alliance®

Key Takeaways

- On June 9, 2026, NIST announced a peer-reviewed mathematical proof establishing that no finite set of AI guardrails can be universally robust against adversarial prompts — the formal demonstration that static AI safety controls have an irreducible structural limit.
- The proof, by NIST Senior Scientist Apostol Vassilev, formally extends Gödel's incompleteness theorems to AI safety systems: because the adversarial prompt space is infinite while guardrail rule sets are necessarily finite, complete protection is mathematically impossible.
- Empirical evidence supports the theoretical finding: a recent fine-tuning technique called NOICE (No, of Course I Can Execute) bypassed safety controls in 72% of Claude Haiku evaluations and 57% of GPT-4o evaluations.
- Static compliance models built on one-time guardrail documentation can no longer treat guardrail presence as adequate evidence of AI safety, given a federal standards body's peer-reviewed finding that static guardrail sets have irreducible structural limits.
- Governance frameworks must shift from "enumerate and prevent" to "monitor, detect, and adapt" — mandating continuous adversarial red-teaming, dynamic guardrail update cycles, and operational resilience controls as first-class requirements.

Background

A common enterprise pattern in AI security has been to build defensive posture primarily around guardrails: rule-based systems that screen inputs or outputs to block harmful requests. These take many forms — keyword filters, classifiers trained on harmful content, constitutional AI constraints, content moderation APIs, and system prompts instructing models to refuse certain categories of requests. Vendors have widely marketed these controls as the foundation of safe AI deployment, and many enterprise AI governance frameworks, including third-party audit checklists and procurement questionnaires, have adopted guardrail presence as a primary compliance criterion.

This posture carries intuitive appeal. If harmful outputs can be defined, they can presumably be blocked; if adversarial inputs can be characterized, they can be filtered. The practice mirrors traditional software security, where allowlists and denylists have proven effective for constrained, predictable input domains

such as SQL queries or file-path traversal strings.

Large language models do not operate over constrained, predictable domains. They process natural language — a medium whose expressive range is effectively infinite. The same semantic intent can be encoded in an unbounded number of syntactically distinct formulations: multilingual rephrasing, indirect analogies, hypothetical framings, encoded or obfuscated text, role-playing scenarios, adversarial formatting, and techniques that exploit model context in ways no fixed rule anticipates. No finite enumeration of prohibited patterns can cover this space exhaustively.

For years, practitioners understood this limitation intuitively, and adversarial prompt injection has ranked first on the OWASP Top 10 for LLM Applications for two consecutive editions [1]. But intuition is not proof. On June 9, 2026, it became proof.

Security Analysis

The Vassilev Proof

Apostol Vassilev, a Senior Scientist at the National Institute of Standards and Technology specializing in adversarial machine learning, published "Robust AI Security and Alignment: A Sisyphean Endeavor?" in IEEE Security & Privacy (DOI: 10.1109/MSEC.2026.3678214) [2][3]. The paper, also available as an arXiv preprint, provides a formal mathematical argument that the security limitations of AI guardrails are not engineering deficiencies that better tooling can remedy, but structural properties of finite systems operating over infinite input spaces.

The proof formalizes guardrails as a checker function that evaluates whether a given prompt violates a set of behavioral constraints. Vassilev establishes five theorems building on Gödel's incompleteness theorems and Chaitin's extension of them. Theorem 2 is the central result: for any such guardrail checker, there exist constraints that the system will fail to enforce correctly against some adversarial input. This holds not because any particular guardrail implementation is flawed, but because the property is a structural consequence of the finite/infinite asymmetry. Theorem 3 tightens the result for practical deployments: systems operating under real-world finite context windows face additional binding limitations, because the information available to the guardrail function is itself bounded even before the infinite input space is considered.

The intuition mirrors Gödel's original incompleteness insight: a finite set of axioms cannot produce a complete, consistent formal system across an unbounded domain. Characterizing the Gödel analogy at the heart of the proof, Vassilev stated in NIST's announcement: "You can't have a finite set of statements and create a theory that is complete and consistent without contradictions." [3] Applied to AI safety, no

matter how carefully constructed, a static guardrail set will contain gaps. Those gaps may be difficult to locate in practice, but adaptive adversaries have demonstrated the ability to find them — and as models evolve, are fine-tuned, or encounter novel input distributions, the gap set itself changes.

The practical consequence is significant. Static compliance documentation that lists "guardrails implemented" as a completed control should no longer be treated as sufficient evidence of ongoing AI security assurance. Guardrails remain useful defensive layers — the proof does not argue they should be abandoned — but they must be understood as controls with documented theoretical boundaries, not as sufficient safeguards.

Empirical Corroboration

Kazdan et al. (2025) demonstrated a fine-tuning attack technique called NOICE, which exploits LLMs' formulaic refusal patterns [4]. The attack trains models on a dataset in which the model briefly refuses a request on safety grounds before complying regardless. By encoding this "refuse-then-comply" pattern into model weights, NOICE bypasses token-level safety mechanisms that depend on detecting refusal signals at inference time — a category that includes many deployed guardrail systems. In controlled evaluations, the technique succeeded in 72% of tests against Claude Haiku and 57% against GPT-4o, two commercially deployed frontier models [4].

The NOICE results illustrate what the Vassilev proof predicts: as adversaries identify and map the structure of existing guardrails, they find paths through the gaps. The attack is notable because it succeeds against closed-source frontier models, not only open-weight models where direct weight access is available. This suggests that vendor-provided guardrails do not transfer safety assurance to operators in cases where the model has been fine-tuned or otherwise adapted downstream. Notably, the NOICE paper was published in February 2025, ten months before the Vassilev preprint appeared — the theoretical incompleteness result thus accounts for empirical evasion already observed in practice, not merely future threats.

Prompt injection has persisted as the top-ranked vulnerability class in the OWASP Top 10 for LLM Applications across editions [1] — a pattern the Vassilev proof now explains theoretically: static input filters cannot keep pace with an effectively infinite adversarial input space. The defense surface is finite; the attack surface is not.

Implications for Current Compliance Models

The most direct governance consequence of this proof concerns how organizations document and audit AI safety controls. Compliance programs that treat guardrail configuration as a one-time activity — deploy, document, pass audit — can no longer treat point-in-time guardrail documentation as sufficient

evidence of ongoing safety assurance. The Vassilev proof provides a formal theoretical basis for this position: organizations and auditors that continue to accept "guardrails implemented" as a complete control can no longer claim ignorance of its theoretical limits. Best practice should shift toward requiring continuous demonstration of guardrail adequacy through adversarial testing.

This shift carries cascade effects across organizational functions. AI risk assessments must account for guardrail decay: the window between when a bypass technique is discovered and when it is reflected in guardrail updates. Security teams must maintain adversarial testing cadences that generate measurable signal about this decay rate. Procurement processes for AI services should evaluate vendors on the velocity and transparency of their guardrail update cycles, not merely on the presence of controls at a point in time. Legal and compliance functions should assess whether their AI risk management documentation satisfies the EU AI Act's ongoing conformity assessment and robustness requirements [5] for high-risk AI systems — particularly given that the Act contemplates continuous monitoring rather than point-in-time evaluation.

The proof also has implications for organizations using fine-tuning or custom model configurations. As the NOICE research demonstrates, fine-tuning can alter safety properties at the model weight level in ways that bypass inference-time guardrail checks entirely [4]. Organizations permitting fine-tuned variants in production — whether first-party or third-party — must treat those variants as requiring independent safety evaluation, not inherited safety assurance from the base model's published guardrails.

Recommendations

Immediate Actions

Organizations deploying AI systems should audit whether their existing AI security documentation includes a baseline adversarial testing record. If current evidence of AI safety relies solely on point-in-time control documentation without red-team testing results, that gap should be escalated to the AI risk owner. Initial red-team exercises should be prioritized for AI systems handling sensitive data, making autonomous decisions, or operating in regulated environments.

Security teams should also review AI systems for any reliance on token-level refusal detection as a primary safety mechanism, given the demonstrated ability of NOICE-class attacks to defeat this category of guardrail. Systems relying solely on inference-time output filtering without model-level alignment verification represent elevated residual risk that should be documented and disclosed to relevant stakeholders.

Short-Term Mitigations

Organizations should establish a guardrail update cadence — a defined frequency at which AI safety configurations are reviewed against new adversarial research and updated accordingly. This should be treated analogously to patch management: not an ad-hoc activity triggered by incidents, but a scheduled operational control with defined ownership and escalation criteria. Vendor contracts and SLAs for AI platforms should be reviewed to determine whether guardrail update timelines are disclosed and whether operators receive notification when material safety updates occur.

Red-teaming should be formalized as a recurring practice rather than a one-time assessment. Effective red teams must remain independent from the teams that configure guardrails, be briefed on current adversarial research, and operate with a mandate to identify bypass techniques specifically relevant to the organization's AI deployment context. External red-team services specializing in adversarial LLM testing provide a useful supplement to internal capability, particularly where dedicated AI security staff are not available.

Operational resilience planning for AI systems should explicitly address guardrail evasion as a first-class failure scenario. Incident response playbooks should define detection criteria for guardrail bypass events, escalation paths, and containment measures — including the ability to modify or suspend AI system behavior in response to an active evasion campaign without requiring a full deployment cycle.

Strategic Considerations

At the program level, organizations should revisit the architecture of their AI governance frameworks to distinguish between static controls (configuration documentation, design reviews, access controls, training data governance) and dynamic controls (adversarial testing, behavioral monitoring, anomaly detection, response, and recovery). The Vassilev proof establishes the structural limit of static controls, and the practical implication is that dynamic controls — adversarial testing, behavioral monitoring, and response capabilities — cannot be treated as merely supplemental. In any AI system processing adversarial inputs, they become operationally necessary to compensate for the irreducible limits of static guardrails. Governance frameworks that mandate only static controls — without requiring ongoing adversarial testing or behavioral monitoring — leave organizations exposed to the irreducible gaps the proof identifies, and should be revised to include dynamic assurance requirements.

AI procurement and vendor management functions should incorporate guardrail transparency into their evaluation criteria. Relevant questions include: How frequently are guardrails updated? What is the typical time from public disclosure of a new bypass technique to its mitigation in production? Are

customers notified of material guardrail changes? Is adversarial testing conducted by the vendor, and at what frequency and depth? The absence of any substantive answer to these questions is itself informative about a vendor's actual AI safety posture.

For organizations subject to AI-specific regulatory requirements — including the EU AI Act, emerging sector-specific guidance in financial services and healthcare, and AI-related provisions in existing privacy and critical infrastructure regulations — legal and compliance teams should assess whether regulatory interpretations of "robustness" and "ongoing monitoring" are likely to evolve in light of this proof. The publication of a mathematical incompleteness result by a federal standards body creates a shared baseline of knowledge that regulators may eventually incorporate into guidance. Proactive engagement and documentation of dynamic monitoring programs may be warranted in high-stakes deployment contexts.

CSA Resource Alignment

MAESTRO (Agentic AI Threat Modeling Framework): MAESTRO's seven-layer threat model addresses Layer 5 (Evaluation and Observability) and Layer 6 (Security and Compliance) — precisely the layers implicated by the Vassilev proof [6]. Practitioners applying MAESTRO should ensure that threat model refreshes are scheduled at regular intervals rather than conducted once at system design time. The proof reinforces MAESTRO's emphasis on observability as a first-class security requirement: detection and response capability is not an operational convenience, but a structural necessity when prevention cannot be guaranteed.

AI Controls Matrix (AICM): The AICM v1.1 provides control domains for AI Model Providers, Application Providers, and Orchestrated Service Providers, covering governance, monitoring, and incident response [7]. The Vassilev proof speaks most directly to AICM's monitoring and behavioral assessment controls. Organizations using the AICM for compliance mapping should ensure that monitoring-related controls are not satisfied by static documentation alone — audit evidence should include adversarial testing records, guardrail update logs, and incident response exercise results. The AICM's distinction between Model Provider and Application Provider responsibilities is also directly relevant: Application Providers cannot assume that base model guardrails transfer safely through fine-tuning or customization.

STAR (Security, Trust, Assurance, and Risk) Registry: STAR assessments should evolve to reflect the distinction between static AI safety documentation and dynamic AI safety demonstration. A STAR entry listing guardrails as implemented without corresponding evidence of an ongoing adversarial testing program represents an incomplete picture of an AI provider's actual security posture. The STAR program

is well-positioned to define tiered assurance criteria for AI-specific submissions that differentiate point-in-time documentation from continuous assurance, analogous to the existing SOC 2 Type I/Type II distinction.

Zero Trust Alignment: The Vassilev proof maps directly to Zero Trust's foundational premise of "never trust, always verify." Applying Zero Trust to AI safety means treating every static guardrail configuration as unverified until demonstrated by current adversarial testing. The assumption of *complete* coverage — "we have guardrails, therefore we are fully protected" — is precisely the assumption the proof invalidates. Continuous verification of safety control effectiveness, not the presence of controls at a moment in time, is the applicable Zero Trust standard.

NIST AI RMF (GOVERN, MAP, MEASURE, MANAGE): Vassilev's publication from within NIST and the paper's explicit governance recommendations align directly with the AI RMF's MEASURE and MANAGE functions, which call for ongoing AI risk monitoring and iterative mitigation [8]. The proof provides the theoretical grounding for treating AI MEASURE controls as continuous activities rather than periodic snapshots — a framing consistent with NIST's own public statement that organizations need to move away from "one and done" security models.

References

[1] OWASP. "OWASP Top 10 for LLM Applications 2025." OWASP Gen AI Security Project, November 2024.

[2] Apostol Vassilev. "Robust AI Security and Alignment: A Sisyphean Endeavor?" arXiv:2512.10100, December 2025. Published in *IEEE Security & Privacy*, May/June 2026. DOI: 10.1109/MSEC.2026.3678214.

[3] NIST. "NIST Mathematical Proof Supports Transition to a Continuous-Monitor-and-Update Security Model for AI Systems." National Institute of Standards and Technology, June 9, 2026.

[4] Joshua Kazdan, Abhay Puri, Rylan Schaeffer, Lisa Yu, Chris Cundy, Jason Stanley, Sanmi Koyejo, Krishnamurthy Dvijotham. "No, of Course I Can! Deeper Fine-Tuning Attacks That Bypass Token-Level Safety Mechanisms." arXiv:2502.19537, February 2025.

[5] European Parliament and Council. "Regulation (EU) 2024/1689 – Artificial Intelligence Act." Official Journal of the European Union, July 2024.

[6] Ken Huang. "Agentic AI Threat Modeling Framework: MAESTRO." Cloud Security Alliance, February 6, 2025.

[7] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, 2025.

[8] NIST. "AI Risk Management Framework 1.0." National Institute of Standards and Technology, January 2023.