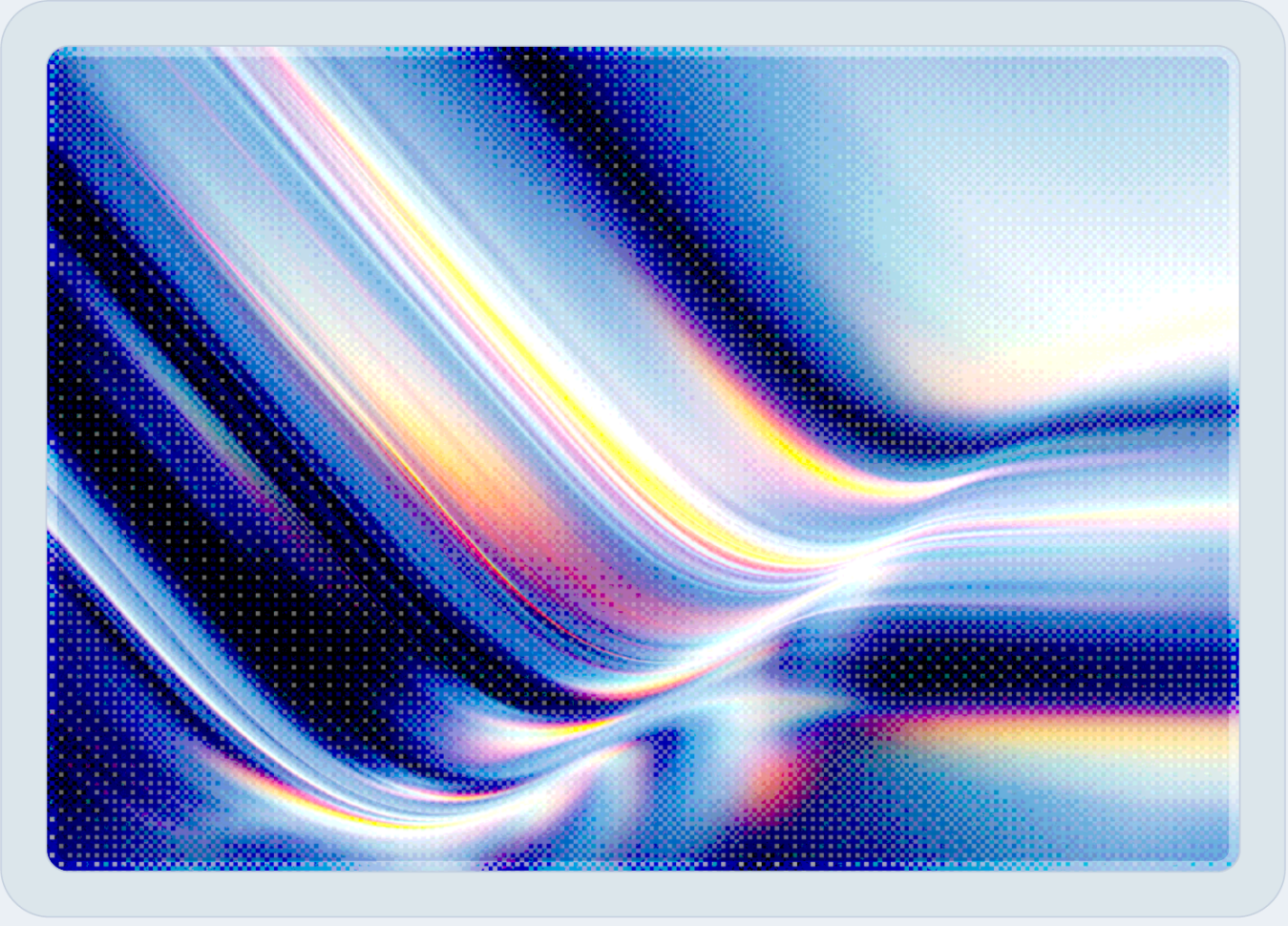
CSAI Foundation | Cloud Security Alliance

NIST: Static AI Guardrails Cannot Achieve Universal Robustness

Governance Implications of the Continuous-Monitor-and-Update Security Model

2026-06-26

AI-assisted Rapid Research



CSAI | CSA cloud security alliance®

Key Takeaways

- On June 9, 2026, NIST announced a peer-reviewed theoretical argument by senior scientist Apostol Vassilev, published in *IEEE Security & Privacy*, providing strong theoretical grounds for concluding that no finite set of AI guardrails can be universally robust against adversarial prompts [1].
- The paper argues by analogy from Gödel's incompleteness theorems that, because the adversarial input space of natural language is effectively unbounded while any guardrail set remains necessarily finite, no static rule-based safety framework can achieve complete adversarial coverage — a rigorous peer-reviewed argument, though the formal analogy between mathematical logic and AI inference remains an active area of theoretical inquiry [1].
- Empirical evidence corroborates the theoretical finding: fine-tuning attacks that instruct models to first refuse, then comply, achieved bypass rates of 72% against Claude Haiku and 57% against GPT-4o [3], and prompt injection has ranked first in the OWASP LLM Top 10 for two consecutive editions [4].
- NIST recommends transitioning from "deploy-and-forget" AI security to a continuous-monitor-and-update model built on three operational pillars: persistent red-team discovery, rapid guardrail updating, and resilience engineering designed for damage containment rather than treating complete breach prevention as an achievable goal [1].
- AI security compliance frameworks, procurement standards, and audit methodologies must be redesigned to evaluate the quality of a continuous discovery-and-update cycle, not the existence of static controls at a point in time.

Background

For the first several years of production AI deployment, organizations understood guardrails largely as a configuration artifact: a bounded set of rules, classifiers, and refusal instructions applied at model inference time to prevent harmful outputs. This framing made intuitive sense. Safety constraints on other software systems — input validation, access control lists, output encoding rules — are indeed finite and, once correctly implemented, provide reliable protection. The natural assumption was that AI guardrails could work the same way.

The compliance posture that followed from this assumption was equally intuitive. Organizations deploying AI systems would specify their safety requirements, vendors would demonstrate that their models incorporated appropriate guardrails, and auditors would verify compliance at a point in time. Security teams treating AI as just another software system applied familiar controls: a penetration test at deployment, periodic reviews, and a policy attesting to the guardrails in place. Regulators drafting AI governance requirements worked within the same paradigm, seeking to establish minimum control standards rather than operational process standards.

This framing has now been fundamentally challenged by a peer-reviewed theoretical argument. On June 9, 2026, NIST announced a paper demonstrating that the category of static guardrail frameworks faces a theoretically insurmountable limit on universal robustness [1]. The paper, authored by Apostol Vassilev, a senior scientist in NIST's Information Technology Laboratory, was published in the May/June 2026 issue of *IEEE Security & Privacy* under the title "Robust AI Security and Alignment: A Sisyphean Endeavor?" [1]. Its implications are not incremental. Vassilev does not argue that existing guardrails are poorly designed or that vendors need to do better work within the existing paradigm. The argument establishes that the paradigm itself is bounded by a theoretical limit that cannot be engineered away.

Security Analysis

The Mathematical Foundation

Vassilev's argument draws on Kurt Gödel's incompleteness theorems, first published in 1931. Gödel demonstrated that any sufficiently expressive formal system built from a finite set of axioms will necessarily contain true statements that cannot be proven within that system. Completeness and consistency cannot simultaneously be guaranteed from a finite rule base — there will always be truths the system cannot reach.

Vassilev applies this insight to AI guardrails by analogy. Guardrails function as a bounded ruleset: a finite collection of classifiers, instructions, and constraints that attempt to enumerate the space of inputs an AI system should refuse to act on. Natural language, however, does not behave like a closed mathematical domain. Its ambiguity is effectively infinite. The syntactic and semantic space through which an adversary can frame a harmful request is vast and continually expanding, shaped by context, cultural reference, newly coined terminology, and the compounding creativity of human expression. As Vassilev states in the paper, "the complexity and richness of the language makes compliance-checking built on a finite set of rules infinitely ambiguous" [1].

The consequence Vassilev derives from this analogy is this: for any finite set of guardrails, a prompt is likely to exist that causes the AI system to disregard them [1]. Adding new rules to address a discovered bypass does not close the gap — it shifts it. The system becomes like a finite formal system trying to enumerate all integers: no matter how many rules are added, the adversarial input space will always extend beyond them. As Vassilev put it in the NIST announcement, "You can't escape Gödel in math, and in AI you likely can't patch an AI system like an LLM and then expect to be OK forever." [1]

It is important to understand the scope of this claim. The argument does not conclude that guardrails are useless. Its implication is that no finite set of guardrails can be *universally* robust — which means that treating any current guardrail configuration as permanently sufficient is an indefensible posture. Guardrails remain valuable as real-time mitigations and cost-raising mechanisms against opportunistic attackers. What must change is the assumption that guardrails can be statically deployed and forgotten; they must be continuously evolved as the adversarial input space evolves.

Empirical Corroboration

The theoretical finding is not an isolated result; empirical research has been converging on the same conclusion from the opposite direction. A 2025 study on fine-tuning-based safety bypass demonstrated that "refuse-then-comply" attacks — in which a model is fine-tuned to perform an initial refusal before providing harmful content, thereby evading output-layer classifiers — achieved success rates of 72% against Claude Haiku and 57% against GPT-4o [3]. These attacks demonstrated that safety mechanisms embedded at training time are not immune to post-deployment adversarial modification. The attack vector was accessible enough that the research team reported it to OpenAI, which awarded a $2,000 bug bounty [3].

The attack surface for static guardrails has also been visible in industry-wide vulnerability tracking for years. Prompt injection — the technique of embedding adversarial instructions in user input that the model interprets as system-level directives — ranked first in OWASP's LLM Top 10 in its inaugural 2023 edition and retained that position in the 2025 update [4]. OWASP's own analysis notes that it remains unclear whether fool-proof prevention is achievable, a practitioner observation that Vassilev's theoretical argument now places on a more rigorous footing [4].

The convergence of theoretical argument and empirical measurement points to a consistent conclusion [5]. The attack surface against static AI guardrails is not an engineering deficit that better tooling will close. It is a structural property of the problem, grounded in the theoretical limits Vassilev identifies and confirmed by the empirical bypass rates researchers have documented, that requires a fundamentally different operational model.

Compliance and Governance Implications

The security implications of the argument extend beyond defensive architecture into governance and legal risk. An organization that deploys an AI system with static guardrails, treats that deployment as a fulfilled safety obligation, and conducts no ongoing adversarial testing is not meeting the standard that the formal and empirical evidence now establishes. As awareness of the Vassilev paper spreads through regulatory and legal communities, this posture is increasingly likely to be characterized as negligent in the event of a guardrail bypass that causes harm.

While NIST AI RMF already incorporates iterative monitoring as a core function through its GOVERN, MAP, MEASURE, and MANAGE activities, its practical implementation guidance and associated audit tools have not yet been calibrated to reflect the adversarial testing cadence that Vassilev's argument implies is necessary. EU AI Act conformity assessments, meanwhile, are structured around point-in-time evaluation, making them more directly due for revision as regulatory awareness of this theoretical limit grows. Organizations that wait for regulatory guidance to catch up before improving their own posture are accepting meaningful risk in the interim.

The theoretical finding provides the foundation for a future shift in governance expectations: as regulatory and legal awareness of Vassilev's argument spreads, organizations relying solely on static guardrails may increasingly face scrutiny to justify why continuous adversarial testing is not warranted. Proactive organizations can get ahead of this shift by establishing documented testing cadences now, creating a defensible record before regulators formalize the standard.

Recommendations

Immediate Actions

Security and AI governance teams should convene promptly — ideally within a standard risk-assessment cycle — to assess whether their current AI deployments rely on static guardrail configurations that have not been adversarially tested since deployment. For each deployment, the key question is whether anyone with an operational mandate is actively searching for new prompt bypasses against that system. If the answer is no, the system should be treated as carrying unquantified residual risk regardless of the vendor's safety certifications.

Organizations should review the adversarial testing posture of AI vendors in their current portfolio. Vendor security questionnaires and service level agreements should be audited to determine whether they include any commitment to ongoing red-team activity, guardrail update cadence, or disclosure of

discovered bypasses. Vendors that cannot demonstrate active programs of this kind should be engaged directly, and that engagement should be documented as part of the organization's vendor risk record.

Incident response plans for AI system failures should be reviewed to ensure they include guardrail bypass scenarios. Plans that assume breaches are primarily infrastructure events rather than inference-layer events will be structurally mismatched to the risk profile the argument describes. Tabletop exercises should include scenarios in which production AI outputs are manipulated through adversarial prompting.

Short-Term Mitigations

The most operationally significant architectural change organizations can make in the near term is decoupling their guardrail controls from application code and embedding them in a gateway layer that supports hot updates without redeployment cycles. When a new bypass is discovered, an update to a gateway-layer control can be deployed in minutes or hours; a patch requiring model retraining or application redeployment may take days or weeks. That gap is the window of maximum exposure, and shortening it is one of the highest-leverage near-term mitigations available.

Adversarial testing should be integrated into existing continuous integration and deployment workflows rather than treated as a separate security activity. Nancy Wang, CTO of 1Password, has publicly advocated for making "continuous validation part of the engineering lifecycle" (as quoted in [2]), and the Vassilev argument makes the operational case for exactly this. Automated adversarial prompt libraries, updated with newly published jailbreak techniques, can serve as a baseline regression suite that executes with every deployment. The intent is not to eliminate the need for skilled red teamers but to ensure that known bypasses are caught continuously rather than discovered by adversaries after they have become stale.

Organizations deploying AI in high-risk contexts — healthcare decision support, financial fraud analysis, legal research tools, critical infrastructure monitoring — should establish minimum update frequency commitments for their guardrail configurations and monitor vendor compliance against those commitments as part of ongoing security operations.

Strategic Considerations

The long-term governance implication of the argument is a necessary shift in the unit of measurement for AI security posture. Current frameworks measure the presence of controls. The evidence now requires measuring the effectiveness of a continuous discovery-and-update cycle. This distinction matters enormously for audit design. An audit that asks "does the system have guardrails?" and "were they tested at deployment?" does not provide meaningful assurance. An audit that asks "what is the

cadence of adversarial testing?", "how quickly are discovered bypasses addressed?", and "what is the current mean time to guardrail update?" begins to reflect the operational reality that the theoretical analysis describes.

Procurement standards should be updated to require AI vendors to disclose their red-teaming programs, guardrail update cadences, and responsible disclosure processes as conditions of procurement approval. Organizations in regulated industries — particularly financial services, healthcare, and critical infrastructure — have an opportunity to drive this standard through their vendor requirements before regulators mandate it, creating first-mover advantage in supply chain security posture.

AI risk programs should also reckon honestly with the irreducible residual risk that the theoretical analysis establishes. Because no guardrail set is universally robust, some adversarial prompts will succeed despite best efforts. Organizations should invest in detection and containment capabilities — behavioral monitoring of AI outputs, anomaly detection on inference patterns, and user-reporting mechanisms for unexpected AI behaviors — that allow rapid identification and response when guardrails are bypassed. The goal, as Vassilev frames it, is economic equilibrium: raising the cost of attack until adversarial exploitation becomes financially impractical for most threat actors, while building the response capacity to limit harm when sophisticated actors do succeed.

CSA Resource Alignment

The findings of Vassilev's argument map directly onto several active CSA research and framework initiatives that practitioners can use to operationalize the recommended shift.

The **AI Controls Matrix (AICM) v1.1** [6] provides the most immediately applicable governance structure. The AICM's 247 control objectives, distributed across 18 security domains, already reflect a process-oriented approach to AI security rather than a purely control-inventory approach. Organizations can use the AICM to structure their transition to continuous-monitor-and-update postures by focusing on control domains governing AI system testing, incident detection, and ongoing model governance. The AICM's alignment to NIST AI RMF 1.0 makes it a natural bridge between the theoretical shift Vassilev describes and the NIST risk management vocabulary that many practitioners already use.

The **MAESTRO framework** [7] for agentic AI threat modeling provides the adversarial analysis methodology most directly suited to the red-team activities the argument's recommended response requires. MAESTRO's layer-by-layer approach to identifying vulnerabilities in AI agent architectures gives red teams a structured taxonomy for discovering bypass categories that may not be apparent from

informal testing. Its emphasis on how layers interact is particularly relevant to multi-model and agentic pipelines, where a bypass against one component may cascade through the system in ways that a single-system guardrail evaluation would miss.

The **CSA STAR for AI** program [8] offers a certification pathway that can evolve to reflect continuous-monitor-and-update posture as an assurance criterion. Currently, STAR for AI assessments provide point-in-time attestation of AI security controls. The theoretical argument makes a strong case for extending that framework to include assessment of a vendor's adversarial testing cadence, mean time to guardrail update, and disclosed bypass history — the operational indicators that reflect genuine ongoing security commitment rather than a snapshot of controls at certification time.

CSA's broader **Zero Trust** guidance also reinforces the recommended posture. The Zero Trust principle of "never trust, always verify" — applied continuously to network access decisions — provides the governing philosophy for how AI inference should be treated: not as a trusted process that, once configured, can be assumed safe, but as an ongoing interaction that must be monitored, tested, and treated as potentially compromised. The extension of Zero Trust principles to AI inference pipelines, rather than just network boundaries, is a natural evolution that the Vassilev argument now makes urgent.

References

[1] NIST. "NIST Mathematical Proof Supports Transition to a Continuous-Monitor-and-Update Security Model for AI Systems." NIST, June 9, 2026.

[2] Help Net Security. "Every set of AI guardrails can be broken by the right prompt." Help Net Security, June 10, 2026.

[3] Xiong et al. "No, of Course I Can! Deeper Fine-Tuning Attacks That Bypass Token-Level Safety Mechanisms." arXiv:2502.19537, February 2025.

[4] OWASP. "OWASP Top 10 for Large Language Model Applications 2025." OWASP, 2025.

[5] TechXplore. "Mathematical proof reveals why fixed AI guardrails can never block every jailbreak." TechXplore, June 2026.

[6] Cloud Security Alliance. "AI Controls Matrix v1.1." CSA, 2025.

[7] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA, February 2025.

[8] Cloud Security Alliance. "CSA STAR for AI." CSA, 2025.