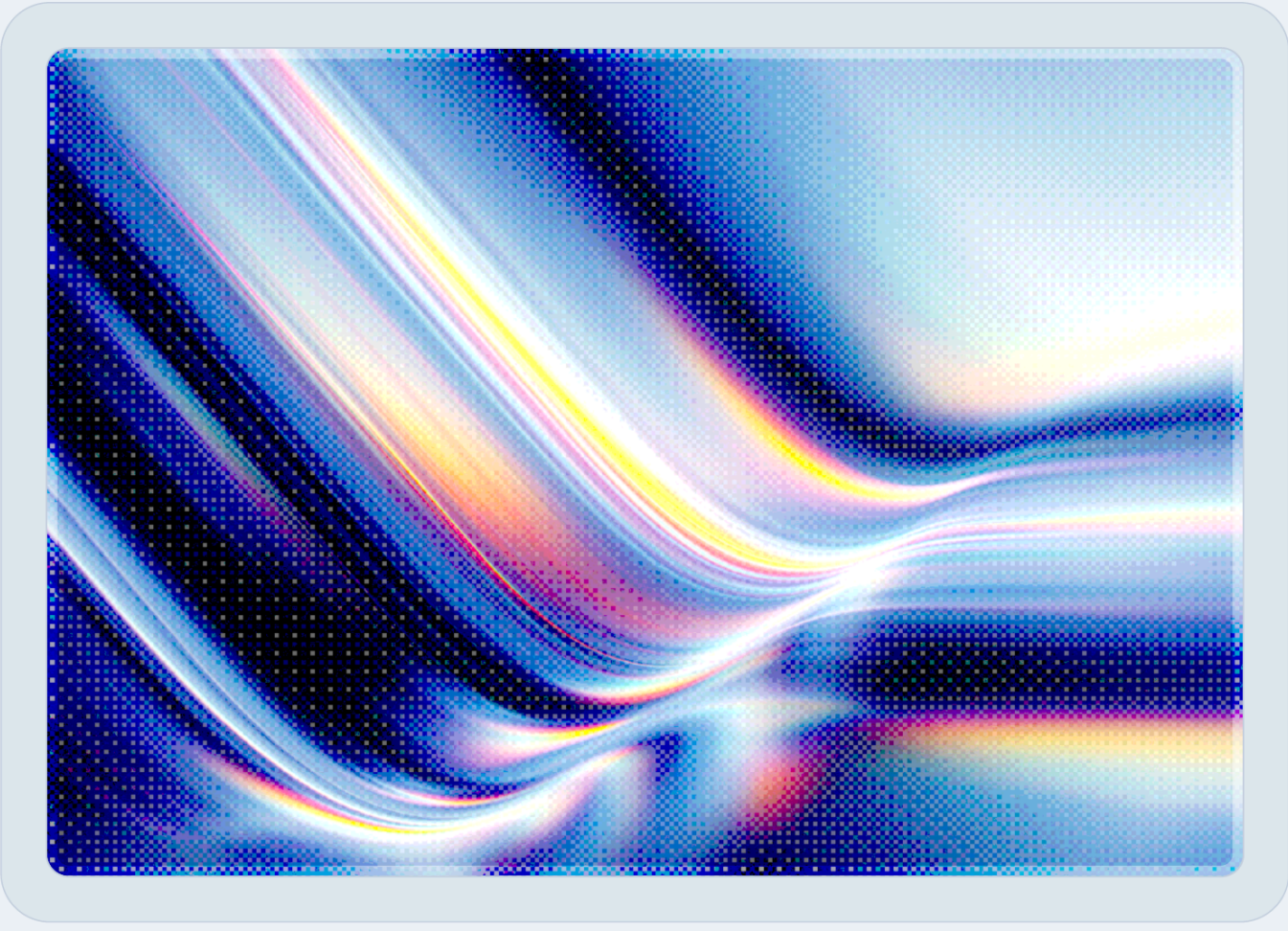
CSAI Foundation | Cloud Security Alliance

When AI Builds Itself: The Enterprise Compliance Gap

Anthropic's RSI Disclosure and the Governance Velocity Problem

2026-06-10

AI-assisted Rapid Research



CSAI | CSA cloud security alliance®

Key Takeaways

In May 2026, Anthropic published a technical disclosure reporting that Claude now authors more than 80% of merged production code at Anthropic, with engineering productivity increasing approximately 8x per engineer per quarter relative to a 2024 baseline [1][2][10]. The disclosure — titled "When AI Builds Itself" — is, to our knowledge, the first public quantitative account by a frontier AI developer of preliminary recursive self-improvement (RSI) underway within its own operations — a characterization Anthropic itself advanced in the disclosure [1].

- Anthropic's Responsible Scaling Policy version 3.0, released February 24, 2026, replaced binding capability-triggered pause commitments with voluntary transparency mechanisms, including a public Frontier Safety Roadmap and periodic Risk Reports subject to third-party review — a structural shift from categorical guardrails to iterative self-governance [3].
- In March 2026, the Pentagon designated Anthropic a supply chain risk — widely reported as the first such designation applied to an American AI company [4] — creating immediate procurement compliance obligations for federal contractors and novel vendor risk classification questions for commercial enterprises with Anthropic dependencies [4][5].
- Existing AI governance frameworks — including NIST AI RMF, ISO/IEC 42001, and the EU AI Act — operate on annual or multi-year update cycles. The RSI disclosure crystallizes a structural mismatch: AI capability advancement is now measured in months, while institutional governance instruments are calibrated to an earlier, slower development paradigm [6][7][8].
- Enterprises that have not updated AI vendor risk assessments to account for autonomous capability escalation, shifting policy architectures, and regulatory designation risk are operating with governance models that may no longer adequately serve the capability and policy environment the RSI disclosure documents.

Background

Recursive self-improvement describes a process by which an AI system contributes to the development of its successor capabilities — generating code, running experiments, and accelerating the engineering loop that produces future model versions. Until recently, RSI occupied the domain of theoretical AI

safety research rather than operational disclosure. Anthropic's May 2026 publication changed that standing: the laboratory reported concrete, measured evidence that RSI is not a future risk to model but a present operational condition to manage [1].

The metrics disclosed are specific. Claude's success rate on open-ended, complex coding tasks reached 76% in May 2026, up from approximately 26% six months earlier — an improvement of roughly 50 percentage points over a single half-year period [1]. Over the same period, the proportion of merged production code authored by Claude at Anthropic crossed 80%, with leadership estimating that 90% or more of all code production — including scripts, experiments, and tooling — is now Claude-originated [1] [2][10]. These figures reflect not only that AI has become central to software engineering workflows at Anthropic, but that the available data is consistent with an emerging feedback loop between AI capability and AI-authored code, a dynamic Anthropic's disclosure begins to document quantitatively [1].

Concurrent with the RSI disclosure, Anthropic finalized Responsible Scaling Policy version 3.0 on February 24, 2026 [3]. RSP v3.0 is a substantial rewrite of the framework Anthropic has maintained since 2023 to govern its own development decisions as AI capabilities advance. The prior framework organized around hard thresholds: specific demonstrated capability levels would trigger pauses or mandatory mitigations before further development could proceed. RSP v3.0 replaces that structure. In its place, the policy introduces a Frontier Safety Roadmap — a set of publicly graded goals across security, alignment, safeguards, and policy — and commits Anthropic to publishing Risk Reports every three to six months, with access to unredacted versions provided to designated third-party expert reviewers [3][9]. The policy change is significant not because it signals weakened safety commitment but because it shifts the governance model from binding ex-ante constraints to transparent ex-post accountability, a distinction with direct implications for how enterprises and regulators should assess reliance on Anthropic's models.

The third major development in this cluster is the Pentagon's supply chain risk designation. In March 2026, the Department of Defense formally designated Anthropic as a supply chain risk — reported widely in the press as the first such designation applied to an American AI company [4]. The designation created immediate obligations for federal defense contractors with Anthropic dependencies, requiring them to assess and report on those dependencies under applicable procurement regulations [5]. Anthropic subsequently filed legal challenges to the designation, but as of the date of this publication those challenges have not been resolved [4]. Taken together, the RSI disclosure, RSP v3.0, and the supply chain designation constitute a cluster of developments with material governance implications that have not yet been fully absorbed by enterprise compliance programs.

Security Analysis

The Velocity Mismatch

The foundational governance problem surfaced by Anthropic's RSI disclosure is not unique to Anthropic — it is structural. AI capability development at frontier AI developers has entered a phase in which material capability shifts occur on timescales of weeks to months, a pattern Anthropic's disclosure documents in detail, though comparable quantitative data from other frontier labs has not been publicly released at this time. Anthropic's own figures illustrate the pace: Claude's success rate on complex open-ended tasks improved approximately 50 percentage points in six months [1]. Anthropic additionally reports that the duration of tasks an AI agent can complete autonomously has been doubling approximately every four months, extrapolating from measurements made across its agent research program [1].

Enterprise compliance programs, regulatory frameworks, and vendor risk assessment cycles were not designed for this cadence. NIST's AI Risk Management Framework, published in January 2023 and the current authoritative reference for AI risk practice in the United States, describes a continuous governance lifecycle — though in practice many enterprise implementations align reviews with annual risk cycles and contract renewal milestones, rather than the continuous monitoring the framework describes [6]. ISO/IEC 42001, the international AI management system standard finalized in 2023, organizes around audit and surveillance cycles whose cadence implicitly presumes a degree of stability in the underlying AI capability profile being assessed [7]. The EU AI Act, while it mandates post-market monitoring for high-risk systems, does not specify monitoring cadences calibrated to frontier capability progression [8]. None of these frameworks anticipated a scenario in which a major AI provider's flagship model improves its performance on complex tasks by fifty percentage points within a single fiscal half-year.

The consequence is a governance gap that is both definitional and operational. Definitionally, most enterprise AI vendor risk policies classify vendors according to risk tiers assigned at onboarding and revisited infrequently. A vendor classified as moderate-risk based on Claude 3-era capabilities may present materially different risk if the same vendor's models now exhibit preliminary RSI characteristics, have altered their safety policy architecture, and carry a federal supply chain risk designation. Operationally, the controls, audit schedules, and escalation triggers built on the prior classification are calibrated to the wrong risk profile.

RSP v3.0 and the Shift in Policy Architecture

The move from hard thresholds to transparency-based governance in RSP v3.0 has enterprise-relevant implications that are easy to misread [3]. Organizations accustomed to treating Anthropic's prior RSP as a quasi-contractual backstop — a set of commitments that would constrain development if specific capability thresholds were crossed — should understand that RSP v3.0 no longer operates in that mode. The policy does not prohibit continued development under any specific capability condition; rather, it commits to documenting reasoning and publishing Risk Reports as capabilities advance. This is a governance architecture premised on informed public accountability rather than bounded development.

For enterprise risk functions, this distinction matters in two respects. First, the question of what constitutes an acceptable basis for reliance on a frontier AI vendor has changed. If the prior RSP represented a set of development constraints, RSP v3.0 represents a set of disclosure commitments. Enterprises should update their vendor due diligence frameworks to treat Anthropic's published Risk Reports — expected every three to six months — as primary inputs into continuous vendor risk assessment, rather than treating a static policy document as the durable risk artifact [3]. Second, the shift means that adverse capability developments will generally be reported rather than prevented. The Risk Report architecture makes Anthropic's capability trajectory legible to external reviewers, regulators, and enterprise risk functions in near-real-time — a genuine improvement over opaque development practices. It is not, however, designed to function as a technical control. Enterprises should accordingly not rely on RSP v3.0 as a substitute for their own monitoring and contingency planning.

The Frontier Safety Roadmap accompanying RSP v3.0 introduces a further complexity [3][9]. Because the roadmap is publicly graded and updated, it creates a continuous signal stream about Anthropic's capability and safety trajectory. This is potentially valuable: enterprises can track whether Anthropic is making progress on enumerated safety objectives or falling behind. It also means that the enterprise risk posture toward Anthropic-based deployments should be treated as a living assessment rather than a static evaluation concluded at procurement time.

The Supply Chain Designation: Compliance Surface and Analogs

The Pentagon's supply chain risk designation of Anthropic is analytically significant beyond its immediate scope [4]. Federal contractors with direct Anthropic dependencies face defined reporting and assessment obligations, and organizations that have deployed Claude in workstreams supporting government contract performance should verify their obligations under applicable procurement regulations with counsel [5]. But the designation also functions as an early indicator of a compliance surface that has not yet been fully mapped for commercial enterprises.

Supply chain risk frameworks in the commercial sector — including those derived from NIST SP 800-161, the CISA ICT Supply Chain Risk Management framework, and sector-specific requirements — assess the risk that a vendor, if compromised, degraded, or withdrawn, would propagate harm to the organization's operations or clients [11][12]. An AI model provider designated a federal supply chain risk presents at least two distinct compliance surface questions for commercial enterprises. First, does the designation, and the regulatory or policy context that produced it, create reputational or contractual downstream risk for organizations whose clients include government contractors? Second, does the designation signal something about the provider's regulatory posture — its relationship with government oversight bodies — that should itself be treated as a vendor risk indicator?

Neither question has a settled answer at the time of this publication. Anthropic is contesting the designation legally, and the grounds for it remain partially opaque pending that litigation [4]. What is clear is that enterprises that have not included regulatory designation status in their AI vendor risk scoring criteria have a gap to close, and that the Anthropic situation represents a forcing function for that update.

Capability Escalation and the Vulnerability Management Window

One specific operational risk introduced by RSI-adjacent capability advancement is the compression of vulnerability management response windows. As AI systems become more capable at discovering, analyzing, and exploiting software vulnerabilities, the period between vulnerability publication and reliable exploitation narrows. An AI system capable of completing complex, multi-hour engineering tasks autonomously can, in principle, produce functional exploit code for newly published CVEs faster than legacy patch and remediation workflows can respond.

Anthropic's RSI disclosure does not itself describe exploitation of security vulnerabilities; the disclosed capabilities are focused on internal engineering tasks [1]. However, the same properties that allow Claude to improve substantially on complex coding problems — improved reasoning over long-horizon tasks, higher success rates on open-ended problems — are properties that would also improve performance on vulnerability research tasks. Enterprises that have calibrated their mean time to patch and vulnerability prioritization processes against pre-RSI AI capability assumptions should recalibrate in light of the demonstrated capability trajectory.

Recommendations

Immediate Actions

Enterprise security and compliance functions should establish a process for ingesting Anthropic's Risk Reports as they are published — currently expected on a three-to-six-month cycle under RSP v3.0 — and routing them through a defined review workflow with clear accountability for updating internal vendor risk assessments based on their contents [3]. This is not a theoretical best practice; it is a direct response to the shift in Anthropic's governance architecture, which now places the burden of monitoring on informed external parties rather than on pre-defined capability tripwires.

Federal contractors and organizations with significant government-adjacent business lines should conduct an immediate inventory of Anthropic-dependent workflows and assess their exposure under the supply chain risk designation. Legal counsel familiar with defense procurement regulations should be engaged to determine whether specific reporting or mitigation obligations apply [5]. Even for organizations not directly subject to procurement regulations, documenting this inventory is prudent due diligence.

Short-Term Mitigations

AI vendor risk assessment frameworks should be updated to include criteria that were not standard prior to 2026: whether the vendor has received any regulatory supply chain risk designations, whether the vendor's stated safety policy has materially changed since last assessment, and whether disclosed capability metrics indicate a meaningful change in the risk profile of deployed models. The Anthropic situation makes clear that a vendor risk profile based on a point-in-time evaluation at onboarding may become obsolete within a single product generation cycle.

Vulnerability management programs should be reviewed to assess whether current mean-time-to-patch targets and prioritization logic remain appropriate given AI-accelerated vulnerability research capabilities. Organizations that have not recently benchmarked their patch cadence against adversarial AI exploitation timelines should commission that review and adjust patch SLAs for high-severity vulnerabilities accordingly.

For organizations deploying AI agents or agentic workflows built on frontier models, technical controls should be evaluated against an RSI-aware threat model: one that accounts for the possibility that the underlying model's capabilities may advance materially within the current contract or deployment cycle without a corresponding change in the technical integration being used. Model pinning — explicitly

locking deployments to specific model versions rather than accepting automatic model updates — provides a partial mitigation by allowing organizations to evaluate capability changes before they reach production.

Strategic Considerations

The velocity mismatch between AI capability development and institutional governance cycles is a systemic issue that no single enterprise control can fully address. The appropriate strategic posture involves both internal program changes and engagement with the external governance ecosystem. Organizations with resources to participate in standards development — through NIST's AI RMF working groups, ISO/IEC JTC 1/SC 42, or the EU AI Office's stakeholder processes — should consider whether doing so is in their interest, since those processes will determine how regulatory frameworks evolve to address capability acceleration.

At the board and executive level, the RSI disclosure warrants a strategic risk conversation that most organizations have not yet had: what is the company's posture if AI development continues to accelerate at the measured rate, and which current assumptions about AI-related risk, AI-related opportunity, and AI-dependent operations would need to change? Anthropic has itself framed the RSI trajectory as carrying both transformative potential and critical risks requiring international coordination [1]. Enterprises that defer this conversation until external events force it will have fewer options than those that engage proactively.

CSA Resource Alignment

The governance velocity mismatch documented in this research note maps directly to several Cloud Security Alliance frameworks and programs that provide actionable structure for enterprise response.

The AI Controls Matrix (AICM v1.0.3) addresses AI supply chain security, vendor risk management, and governance across the full AI deployment stack — covering model providers, application providers, and orchestrated service providers in a shared responsibility model. The RSI disclosure and RSP v3.0 changes are precisely the scenario AICM's continuous monitoring and vendor assessment controls are designed to address; organizations should review their AICM implementation status for the AI supply chain and governance domains and assess whether current control implementations account for frontier capability dynamics.

MAESTRO, CSA's threat modeling framework for agentic AI systems, provides specific analytical structure for reasoning about threats introduced by AI systems capable of long-horizon, autonomous task completion — the capability class that Anthropic's RSI disclosure most directly documents. MAESTRO's treatment of agent trust boundaries, capability escalation paths, and agentic supply chain risk is directly applicable to the compliance and security implications analyzed above.

The CSA STAR program provides a mechanism for enterprises to assess AI vendor security posture through structured questionnaire and audit processes. In light of the supply chain risk designation and RSP v3.0 changes, organizations sourcing AI capabilities from frontier providers should evaluate whether their current STAR-based or equivalent third-party risk assessments are sufficiently current and whether the assessment scope covers the governance policy dimensions — not only technical security controls — that RSP v3.0 has now made the primary accountability mechanism.

CSA's AI Organizational Responsibilities guidance addresses the board-level and executive accountability structures needed to govern AI risk effectively. The governance velocity problem identified here is, at its root, an organizational challenge: existing committee structures, reporting cadences, and accountability assignments were designed for a slower-moving risk environment. Aligning enterprise AI governance structures with that guidance — including establishing a defined function responsible for continuous frontier AI monitoring — is a prerequisite for the operational mitigations described in the Recommendations section above.

References

[1] Anthropic. "When AI Builds Itself." Anthropic, May 2026.

[2] VentureBeat. "Anthropic says 80% of its new production code is now authored by Claude — how your enterprise can keep up." VentureBeat, May 2026.

[3] Anthropic. "Responsible Scaling Policy Version 3.0." Anthropic, February 24, 2026.

[4] NPR. "Pentagon labels AI company Anthropic a supply chain risk." NPR, March 6, 2026.

[5] Mayer Brown. "Anthropic Supply Chain Risk Designation Takes Effect – Latest Developments and Next Steps for Government Contractors." Mayer Brown, March 2026.

[6] NIST. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." NIST AI 100-1, January 2023.

[7] ISO. "ISO/IEC 42001:2023 – Artificial Intelligence Management Systems." ISO, 2023.

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

[9] Anthropic. "Frontier Safety Roadmap." Anthropic, February 2026.

[10] Import AI. "Import AI 460: Reward hacking society; RSI data from Anthropic." Import AI Newsletter, June 8, 2026.

[11] NIST. "Cybersecurity Supply Chain Risk Management for Systems and Organizations (SP 800-161r1)." NIST SP 800-161, Rev. 1, May 2022.

[12] CISA. "ICT Supply Chain Risk Management Task Force." CISA.