CIRCIA's AI Blind Spot: Closing the Mandatory Reporting Gap

Definitional Gaps in AI Incident Reporting for Critical Infrastructure

Unofficial AI-assisted Research

Cloud Security Alliance AI Safety Initiative

2026-03-27

Key Takeaways

- CIRCIA's proposed "covered cyber incident" definition relies on traditional CIA-triad language — confidentiality, integrity, and availability of information systems — that does not explicitly address AI-specific attack vectors such as data poisoning, model weight manipulation, adversarial machine learning inputs, and supply-chain-compromised model repositories. Critical infrastructure operators currently have no explicit regulatory guidance on whether AI-specific security events trigger CIRCIA's mandatory reporting obligations.
- CISA published a Federal Register notice on February 13, 2026 announcing sector-specific town halls to gather stakeholder input on CIRCIA implementation [1]. Those town halls were subsequently postponed when the DHS shutdown began on February 14, 2026. The final rule's effective date now targets May 2026 at the earliest [2] [3]. Given the unresolved DHS funding situation and the absence of a rescheduled stakeholder engagement calendar, further delay is plausible.
- A fundamental visibility crisis underlies the reporting gap. According to HiddenLayer's 2026 AI Threat Landscape Report, 31% of IT and security leaders do not know whether their organization experienced an AI security breach in the past 12 months [4]. You cannot report what you cannot detect.
- The enforcement culture compounds the definitional ambiguity: 85% of the same survey respondents support mandatory AI breach reporting in principle, yet 53% admit to having withheld breach reports — a gap between stated values and operational behavior that signals systemic uncertainty about what must be disclosed [4].
- Critical infrastructure operators deploying AI in operational technology (OT) environments face compounded exposure. Published guidance from CISA, NSA, and six international partners identifies AI-specific OT risks — including model drift, adversarial inputs causing physical-process misclassification, and limited explainability in safety-critical decisions — that do not map cleanly to traditional incident response frameworks [5].
- Operators should not wait for definitional clarity to act: building AI-specific incident detection and response capabilities now creates compliance readiness for when the rule takes effect in final form, while also reducing operational risk independent of how the final rule resolves the definitional questions.

Background

CIRCIA and Its Rulemaking Trajectory

The Cyber Incident Reporting for Critical Infrastructure Act of 2022 (CIRCIA), signed into law on March 15, 2022, established the statutory foundation for a mandatory federal cyber incident reporting regime covering sixteen critical infrastructure sectors. Congress directed CISA to publish a Notice of Proposed Rulemaking within 24 months of enactment and finalize the rule 18 months thereafter. CISA met the first deadline by publishing the NPRM in the Federal Register on April 4, 2024 [6]. The public comment period closed July 3, 2024, with more than 2,000 comments submitted from industry, government, and civil society [6].

The proposed rule requires covered entities to report a "covered cyber incident" to CISA within 72 hours of reasonably believing one has occurred, and to report ransomware payments within 24 hours regardless of whether a separate cyber incident is reportable [6]. Covered entities are defined using a combination of size-based criteria (exceeding SBA small business thresholds) and sector-specific criteria calibrated to the sixteen PPD-21 critical infrastructure sectors: Chemical, Commercial Facilities, Communications, Critical Manufacturing, Dams, Defense Industrial Base, Emergency Services, Energy, Financial Services, Food and Agriculture, Government Services and Facilities, Healthcare and Public Health, Information Technology, Nuclear Reactors and Materials, Transportation Systems, and Water and Wastewater [7]. CISA estimated that more than 30,000 entities would qualify under size-based criteria alone.

The October 2025 statutory deadline for the final rule passed without publication. CISA self-extended its target to May 2026, citing the volume of public comment and a stated intent to "streamline CIRCIA's requirements consistent with feedback" [8]. The DHS funding lapse that began February 14, 2026 — forcing CISA to postpone the sector-specific stakeholder town halls it had scheduled for March and April 2026 — introduces further schedule risk [2][3]. As of this writing, no revised town hall schedule has been announced and no new final rule date has been committed.

The Definitional Problem

CIRCIA's proposed definition of a "covered cyber incident" characterizes qualifying events as those resulting in substantial loss of confidentiality, integrity, or availability of an information system or network; serious impact on the safety and resilience of operational systems; disruption of the entity's ability to deliver critical services; or unauthorized access facilitated through a cloud service provider, supply chain compromise, or managed service provider [6]. This definition reflects the vocabulary of traditional IT security — a vocabulary developed long before the deployment of AI and machine learning systems into safety- and mission-critical environments.

AI-specific attack vectors do not map cleanly onto these categories. Consider four illustrative cases. First, an adversary who poisons a model's training data — gradually shifting the decision boundary of an anomaly detection system in an energy management SCADA environment — may never trigger a traditional security alert. No unauthorized access event is logged; no availability disruption occurs; the system continues operating in a degraded state that benefits the attacker. Second, model weight manipulation introduced through a compromised model repository — a technique documented in 35% of AI security breaches according to HiddenLayer's 2026 survey [4] — may produce an artifact that passes all standard software supply chain checks because the weights are not executable code, and software composition analysis tooling has historically focused on package manifests and binary signatures rather than model artifact inspection [12]. Third, adversarial inputs engineered to cause misclassification in an AI system performing industrial process control could produce physical consequences — a valve opened at the wrong time, a safety interlock bypassed — that manifest as operational failures rather than cybersecurity incidents, routing the event to operational reliability processes that do not feed into CIRCIA reporting pipelines. Fourth, model drift in a healthcare AI diagnostic system may represent a subtle integrity compromise that no single event makes reportable, even though the cumulative effect constitutes a material degradation in the system's trustworthiness.

None of these scenarios clearly triggers the 72-hour reporting clock under the current NPRM language. This is not a hypothetical concern. The gap exists now and widens as AI systems are deployed into critical infrastructure ahead of any regulatory update to the reporting framework.

The Detection Gap

The reporting problem is inseparable from a prior detection problem. CIRCIA's 72-hour clock runs from when a covered entity "reasonably believes" a reportable incident has occurred — a standard that presupposes meaningful visibility into whether AI systems have been compromised. That visibility does not reliably exist. HiddenLayer's 2026 AI Threat Landscape Report, drawn from a vendor-commissioned survey of 250 IT and security leaders across AI-deploying enterprises, found that 31% do not know whether they experienced an AI security breach in the past twelve months [4]. In a critical infrastructure context, that figure is particularly troubling: an operator cannot discharge a reporting obligation for an incident it has no means of detecting.

The NIST report on challenges to monitoring deployed AI systems (NIST AI 800-4), published in March 2026, independently corroborates this finding at the technical level, identifying "immature information sharing ecosystems for incident data," a lack of trusted guidelines for AI monitoring methodologies, and the absence of standardized approaches to post-deployment AI oversight as critical systemic gaps [9]. Detection of traditional cyber incidents has benefited from decades of investment in endpoint detection and response, security information and event management, and network traffic analysis — tooling that generates the observable artifacts that feed CIRCIA reporting. The NIST report identifies these same tooling gaps as systemic [9], consistent with a market for AI-specific monitoring that remains early-stage and unevenly deployed across sectors.

Security Analysis

AI Attack Vectors and Their Relationship to CIRCIA Thresholds

Understanding where AI-specific attacks do and do not engage CIRCIA's proposed thresholds requires mapping the primary adversarial ML attack categories — authoritatively defined in NIST AI 100-2, the Adversarial Machine Learning taxonomy published in March 2025 [10] — against the proposed reporting criteria.

Data poisoning attacks corrupt the integrity of training data or fine-tuning datasets to degrade model performance or introduce hidden behaviors triggered by specific inputs. Because poisoning occurs during the training process and produces a model that subsequently operates in a degraded state, the attack does not generate a discrete intrusion event; the compromise is baked into the artifact. The CIRCIA definition requires an incident affecting "an information system or network," which could encompass a compromised training pipeline, but the boundary conditions are unclear and no guidance has been issued.

Evasion attacks, in which carefully crafted inputs are designed to cause model misclassification without modifying the model itself, leave no trace in traditional security telemetry. An adversarial image that causes a medical imaging AI to miss a pathological finding, or an adversarial sensor reading that causes a grid management AI to make an incorrect load-balancing decision, produces an output error that looks operationally identical to a natural model mistake. These attacks are difficult to distinguish from benign failures in the absence of dedicated adversarial ML testing, as documented in the NIST adversarial ML taxonomy [10] — and, as noted above, only 9% of energy sector organizations conduct AI red-teaming [11].

Model supply chain compromises, in which backdoored model weights are distributed through public repositories, represent what HiddenLayer identifies as the largest single source category of AI breaches — 35% of those reported [4]. The DoD and NSA published dedicated guidance on AI/ML supply chain risks and mitigations in March 2026, identifying public model repositories as a significant and underaddressed attack surface and noting that model weight inspection remains absent from most enterprise software supply chain security programs [12]. Under CIRCIA, a supply chain compromise explicitly constitutes a qualifying pathway to a covered cyber incident, but only if the organization detects that the compromise occurred. Given that 93% of organizations continue using open model repositories despite known risks [4], many supply-chain-sourced AI compromises will not be detected — and therefore not reported.

The OT-AI Nexus: Compounded Exposure in Critical Infrastructure

The convergence of AI and operational technology in critical infrastructure sectors — energy, water, transportation, healthcare — creates a compounded risk environment where the consequences of undetected AI compromise extend beyond data exposure into physical safety domains. CISA, NSA, the FBI, and six international partner agencies addressed this directly in December 2025 guidance on the secure integration of AI in OT environments [5]. The guidance identifies AI model manipulation, data poisoning, prompt injection into AI-assisted OT controls, model drift, and limited explainability in safety-critical contexts as distinct risk categories requiring integration into OT incident response plans and AI asset inventories.

The current state of AI security practice in critical infrastructure OT environments falls significantly short of this standard. According to a 2026 industry assessment cited in HSToday [11], 91% of organizations lack network isolation for their AI systems, 73% lack the ability to shut down AI systems rapidly if compromise is detected, 64% lack purpose binding to constrain what AI systems are permitted to do, and only 14% maintain AI-specific incident response playbooks. CIRCIA will impose reporting obligations on many of these same organizations beginning on the rule's effective date. Organizations that cannot detect AI-specific incidents and have no AI-specific response procedures face severe compliance risk: they are unlikely to identify reportable events, and unlikely to report them within the 72-hour window if they do.

The Disclosure Culture Gap

The HiddenLayer survey reveals a striking disconnect between stated values and actual behavior that goes beyond detection limitations. Eighty-five percent of surveyed security leaders support mandatory AI breach reporting — yet 53% report having withheld AI breach reports [4]. One plausible explanation is that AI security failures are being treated as reputational risks to be managed internally rather than as shared threat intelligence that benefits the broader ecosystem. The survey data does not capture respondent-stated reasons for withholding, so causal interpretation remains inferential — legal counsel advising silence pending definitional clarity, absence of formal incident classification processes, and uncertainty about reporting channels are equally plausible contributing factors [4]. CIRCIA's mandatory reporting regime is specifically designed to overcome this tendency, but it will only be effective to the extent that operators understand what they are required to report.

The 76% of respondents who cite shadow AI — undisclosed AI systems deployed outside formal IT governance — as a definite or probable problem in their organizations represents a related challenge [4]. CIRCIA's reporting requirements will apply to covered entities regardless of whether their AI systems are formally inventoried. An AI system that causes or contributes to a covered cyber incident without appearing in any asset register creates exactly the compliance exposure that organizations most want to avoid: a reportable event that the organization has no visibility into and no procedure for handling.

Recommendations

Immediate Actions

Critical infrastructure operators should begin by constructing a comprehensive AI asset inventory that identifies every deployed AI system — including shadow AI — and documents its role, the data it processes or acts upon, and its potential failure modes. This inventory is the foundation for any reliable CIRCIA compliance exposure assessment. Operators that are already engaged in CIRCIA compliance preparation for traditional IT and OT systems should extend that preparation to include AI assets, mapping each system against the proposed "covered cyber incident" definition and identifying which failure modes could plausibly constitute a reportable event. The uncertainty about AI-specific thresholds is not an exemption; it is a risk that organizations need to manage proactively.

Operators should also review CISA's December 2025 joint guidance on AI in OT environments and assess their current gap against the recommended controls, particularly around AI asset inventorying, AI-specific incident response procedures, and integration of AI systems into existing OT monitoring and detection frameworks [5].

Short-Term Mitigations

Organizations should develop AI-specific incident response playbooks that address the scenarios most likely to produce ambiguous CIRCIA reporting situations: detected or suspected data poisoning, anomalous model output patterns suggestive of adversarial inputs, supply chain alerts on model artifacts, and model drift exceeding defined operational thresholds. These playbooks should specify who makes the determination that a reportable event has occurred, what evidence must be preserved, and what the CIRCIA reporting pathway looks like — including the responsible individual designated to interface with CISA.

On the detection side, operators should evaluate whether their existing security monitoring infrastructure generates sufficient telemetry to detect AI-specific compromise. Meaningful detection requires logging of model inputs and outputs to establish behavioral baselines, monitoring of model artifact integrity, and instrumentation of training pipelines if those pipelines operate within the covered entity's environment. The December 2025 NIST IR 8596 preliminary draft on a Cybersecurity Framework Profile for AI explicitly recommends preserving logs, inputs, outputs, and decision chains of AI systems as the foundation for incident analysis [13].

Operators should also begin systematic review of AI dependencies sourced from public model repositories. The DoD/NSA March 2026 supply chain guidance provides a practical framework for model provenance verification, checksum validation, and controlled deployment processes that reduce the risk of consuming backdoored model artifacts [12].

Strategic Considerations

Beyond internal preparation, operators and their trade associations can shape the rule itself. CISA has indicated intent to reschedule the postponed town halls when the DHS funding situation resolves, and the final rule — whenever it is published — will shape mandatory reporting requirements for years. Critical infrastructure operators, their trade associations, and the broader AI security community have both the opportunity and the responsibility to advocate for AI-specific provisions in

the final rule. Constructive engagement should focus on three concrete needs: a clarification that AI-specific attack vectors including data poisoning, model manipulation, adversarial ML inputs, and supply chain compromise of model artifacts constitute qualifying cyber incidents when they result in substantial impact on safety or reliability; a recognition that the 72-hour reporting clock should begin at the time of reasonable detection rather than at the time of the underlying compromise for attacks like data poisoning where the attack and the observable impact are temporally separated; and guidance on what constitutes sufficient AI monitoring to satisfy the "reasonable belief" standard that governs the reporting obligation.

Longer term, the 73% of organizations that report internal conflict over ownership of AI security controls — between IT security, data science, and operational teams — need to resolve that question in advance of CIRCIA's effective date [4]. Regulatory compliance responsibility for cyber incidents cannot be distributed across organizational functions without clear escalation paths. CIRCIA requires a designated point of contact for reporting; the organizational groundwork for that designation should be laid before the rule takes effect, not after the first incident.

CSA Resource Alignment

The CSA AI Safety Initiative has produced a body of work directly applicable to organizations navigating the intersection of AI security and mandatory incident reporting obligations. The AI Controls Matrix (AICM) — CSA's superset of the Cloud Controls Matrix that extends CCM into AI-specific risk categories — provides a structured framework for mapping AI security controls to compliance requirements across multiple regulatory regimes simultaneously [14]. Organizations subject to CIRCIA would benefit from conducting an AICM-based control gap assessment specifically scoped to incident detection and response capabilities for AI systems, treating the CIRCIA reporting requirements as one of the compliance dimensions in the mapping.

The CSA Capabilities-Based Risk Assessment (CBRA) for AI Systems offers a methodology for systematically evaluating which AI system capabilities create reporting-relevant risk — a useful tool for prioritizing which systems in a covered entity's AI inventory require immediate attention versus those with lower risk profiles [15]. CBRA's emphasis on what AI systems can do, rather than what they are labeled as, is particularly relevant for identifying shadow AI that may not appear in formal procurement records.

CSA's MAESTRO threat modeling framework for agentic AI systems addresses the emerging category of AI breach that HiddenLayer identifies as growing fastest: agentic system compromises, which now account for approximately one in eight AI security incidents [4]. As critical infrastructure operators deploy agentic AI for functions ranging from grid optimization to threat detection, MAESTRO provides the adversarial modeling vocabulary needed to design detection and response capabilities that can generate the evidence required for CIRCIA reporting.

The AI Organizational Responsibilities series — covering governance, risk management, compliance, and cultural transformation — speaks directly to the organizational culture gap documented by the HiddenLayer survey [4][16]. Bridging the gap between the 85% of organizations that support mandatory reporting and the 53% that actually withhold reports is fundamentally a governance and culture challenge, not a technical one, and it requires the kind of structured organizational accountability that the AI Organizational Responsibilities guidance describes.

Finally, CSA STAR certification provides a vehicle for covered entities to demonstrate to regulators, customers, and partners that their AI and cloud security posture meets established standards — a capability that will become increasingly valuable as CIRCIA compliance creates transparency into which organizations have and have not invested in security governance for their AI systems.

References

[1] CISA, "Cyber Incident Reporting for Critical Infrastructure Act (CIRCIA) Rulemaking Town Hall Meetings," Federal Register Document 2026-02948, February 13, 2026. https://www.federalregister.gov/documents/2026/02/13/2026-02948/cyber-incident-reporting-for-critical-infrastructure-act-circia-rulemaking-town-hall-meetings

[2] Federal News Network, "CISA Delays Cyber Incident Reporting Town Halls Due to Shutdown," March 2026. https://federalnewsnetwork.com/cybersecurity/2026/03/cisa-delays-cyber-incident-reporting-town-halls-due-to-shutdown/

[3] MeriTalk, "CISA Says DHS Shutdown Will Likely Further Delay CIRCIA Rule," March 10, 2026. https://www.meritalk.com/articles/cisa-says-dhs-shutdown-will-likely-further-delay-circia-rule/

[4] HiddenLayer, "2026 AI Threat Landscape Report: Spotlighting the Rise of Agentic AI and the Expanding Attack Surface of Autonomous Systems," March 18, 2026 (vendor-commissioned survey, n=250 IT and security leaders). https://www.hiddenlayer.com/news/hiddenlayer-releases-the-2026-ai-threat-landscape-report-spotlighting-the-rise-of-agentic-ai-and-the-expanding-attack-surface-of-autonomous-systems

[5] CISA, NSA, FBI, and international partners, "Principles for the Secure Integration of Artificial Intelligence in Operational Technology," December 3, 2025. https://www.cisa.gov/resources-tools/resources/principles-secure-integration-artificial-intelligence-operational-technology

[6] CISA, "Cyber Incident Reporting for Critical Infrastructure Act (CIRCIA) Reporting Requirements – Notice of Proposed Rulemaking," Federal Register Document 2024-06526, April 4, 2024. https://www.federalregister.gov/documents/2024/04/04/2024-06526/cyber-incident-reporting-for-critical-infrastructure-act-circia-reporting-requirements

[7] CISA, "CIRCIA Covered Cyber Incident One-Pager," April 2024. https://www.cisa.gov/sites/default/files/2024-05/24-0630-CCI-One-Pager-20240410-2-508c.pdf

[8] Davis Wright Tremaine, "CISA Delays Cyber Incident Reporting Rules Until May 2026," September 17, 2025. https://www.dwt.com/blogs/privacy--security-law-blog/2025/09/cisa-delays-cyber-incident-reporting-rules-2026

[9] NIST, "Report on Challenges to Monitoring Deployed AI Systems – NIST AI 800-4," March 2026. https://www.nist.gov/news-events/news/2026/03/new-report-challenges-monitoring-deployed-ai-systems

[10] NIST, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations – NIST AI 100-2 E2025," March 2025. https://csrc.nist.gov/pubs/ai/100/2/e2025/final

[11] HSToday, "Agentic AI Expands Critical Infrastructure Attack Surface Beyond Governance," 2026. https://www.hstoday.us/subject-matter-areas/ai-and-advanced-tech/agentic-ai-and-the-critical-infrastructure-attack-surface-that-lacks-governance/

[12] NSA and DoD, "AI/ML Supply Chain Risks and Mitigations," March 4, 2026. https://media.defense.gov/2026/Mar/04/2003882809/-1/-1/0/AI_ML_SUPPLY_CHAIN_RISKS_AND_MITIGATIONS.PDF

[13] NIST, "Cybersecurity Framework Profile for Artificial Intelligence – NIST IR 8596 (Preliminary Draft)," December 2025. https://csrc.nist.gov/pubs/ir/8596/iprd

[14] Cloud Security Alliance, "AI Controls Matrix (AICM)," Cloud Security Alliance, 2024–2025. https://cloudsecurityalliance.org/artifacts/ai-controls-matrix

[15] Cloud Security Alliance, "Capabilities-Based Risk Assessment for AI Systems," Cloud Security Alliance, 2025. https://cloudsecurityalliance.org/artifacts/capabilities-based-risk-assessment-cbra-for-ai-systems

[16] Cloud Security Alliance, "AI Organizational Responsibilities: Governance, Risk Management, Compliance, and Cultural Aspects," Cloud Security Alliance, 2024. https://cloudsecurityalliance.org/artifacts/ai-organizational-responsibilities-governance-risk-management-compliance-and-cultural-aspects

