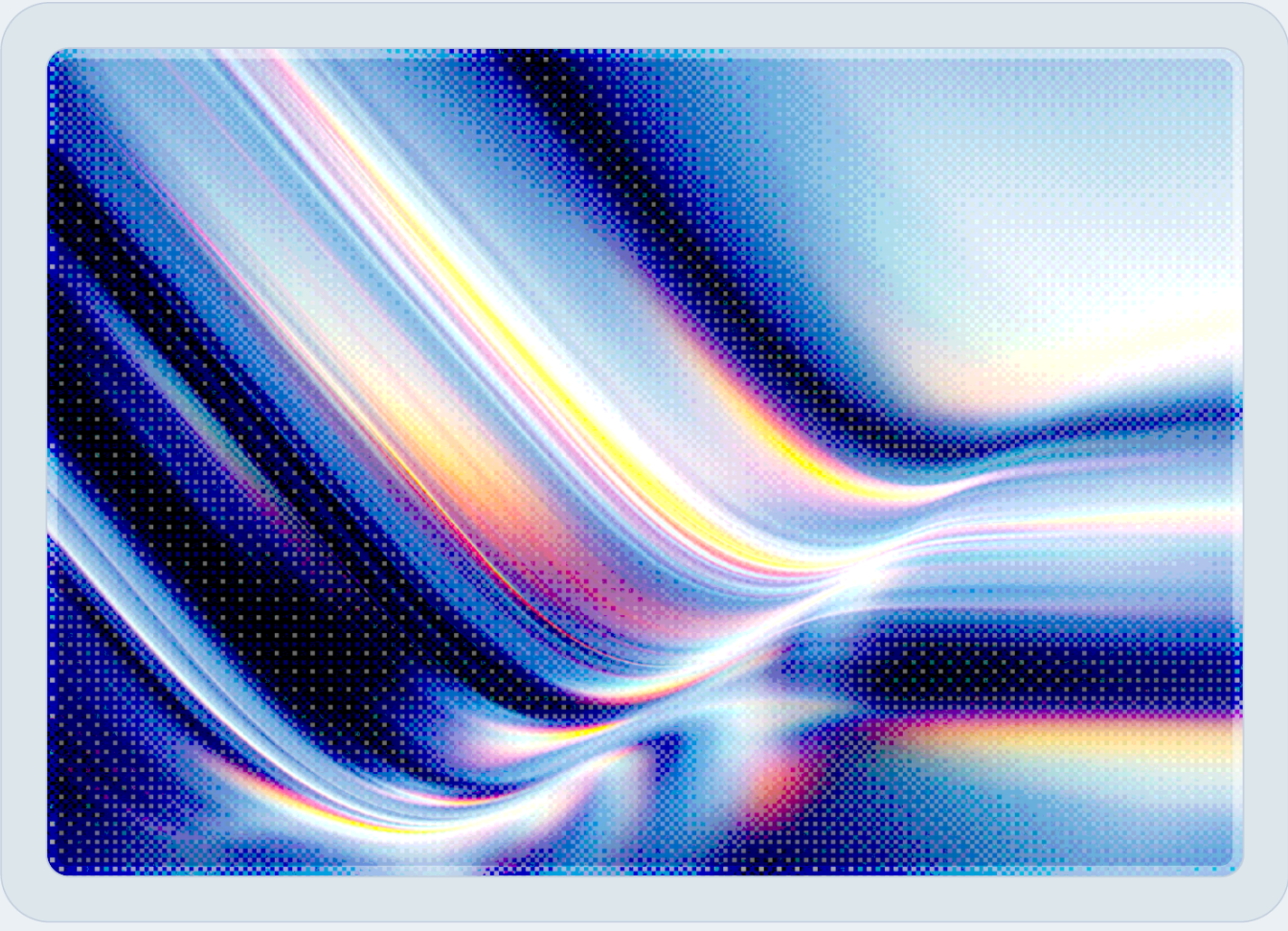
CSAI Foundation | Cloud Security Alliance

Auditable Agents: CERT-In's AI Auditing Guidelines and Global Assurance

What India's tiered AI security audit framework signals for enterprises, auditors, and agent builders worldwide

2026-11-06

AI-assisted Rapid Research



CSAI | CSA cloud security alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

On November 4, 2026, the Indian Computer Emergency Response Team (CERT-In) published its *Artificial Intelligence (AI) Security Auditing Guidelines*. The document sets out a risk-tiered framework for independently auditing AI systems that are developed, procured, deployed, or operated in India, or that process Indian data or serve Indian users [1][2][3]. A national cybersecurity agency has issued an audit methodology, not a set of principles. That choice moves the conversation from "should AI be governed" to "what evidence will an auditor ask for, and how often." Five points stand out for security and compliance leaders.

Audit depth scales with risk. Systems are scored against eight factors, including autonomy, reversibility, and adversarial exposure, and placed in one of four tiers. Obligations range from self-assessment or internal review at the lowest tier to independent audits by CERT-In-empanelled auditors, continuous monitoring, and red-team exercises at the highest [1][2][3].

Agents are explicitly in scope. The guidelines add checks for agentic and multi-agent systems that cover excessive agency, unsafe tool use, persistent memory, least privilege, and prompt injection. Because autonomy is itself a scored factor, it can push a system into a higher tier [2][3][5].

Existing incident duties now extend to AI. CERT-In's April 2022 directions already require six-hour incident reporting and 180-day log retention. The new guidelines apply these duties to reportable AI security incidents and expect time synchronization across all AI components [1][6].

AI-assisted audit findings still need human judgment. The guidelines permit AI-assisted audit procedures, subject to controls, but state that AI-generated outputs cannot replace sufficient and appropriate evidence or the auditor's independent judgment [1].

The global direction is convergence. The structure resembles risk-based regimes elsewhere, including the EU AI Act, ISO/IEC 42001, and the NIST AI RMF [8][9][10]. CSA's view is that organizations building one evidence base mapped to a common control framework, such as CSA's AI Controls Matrix (AICM) v1.1 [11], will face less duplicated effort than those treating each regime as a separate project.

Background

CERT-In is India's national agency for cyber incident response. It operates under the Ministry of Electronics and Information Technology, with authority under Section 70B of the Information Technology Act, 2000 [4][6]. Its April 28, 2022 directions under Section 70B(6) set short operational deadlines: listed cyber incidents must be reported within six hours of being noticed, and system logs must be retained for a rolling period of 180 days [6]. In July 2025, CERT-In issued its *Comprehensive Cyber Security Audit Policy Guidelines*, which listed AI system audits and AI bill of materials (AIBOM) audits among the recognized types of cybersecurity audit [12][13]. The November 2026 AI guidelines build on that base, taking AI auditing from one category within a general audit policy to a dedicated methodology.

The new document is version 1.0, runs to 107 pages, and was prepared by Sub-group 03 of the Working Group on Indian Standards for Cybersecurity Auditing [1][3]. CERT-In's listing page uses the reference code CISG-2026-04 [1], while the copy distributed by MediaNama carries the filename CIGU-2026-0004 [2]; both appear to refer to the same document, which is linked from the CERT-In listing page. The intended audience is broad. It includes government bodies, critical infrastructure operators, regulated entities, private organizations, AI developers and providers, internal and external auditors, and risk and compliance functions [3]. Research AI and consumer personal use are excluded, although research AI is encouraged to adopt the framework voluntarily [1][3]. MSMEs and startups receive proportionality relief for Tier 1 and Tier 2 systems, but Tier 3 and Tier 4 systems remain subject to the full applicable assurance requirements regardless of organizational size [1].

The legal weight of the guidelines needs careful reading. Legal analysis notes that the document uses mandatory language and sets minimum requirements by tier, but does not state its own legal enforceability and prescribes no penalty [3]. Other coverage similarly advises treating the guidelines as a security benchmark rather than a standalone AI law [5][7]. It is also unclear from available material how the guidelines would apply to organizations that serve Indian users or process Indian data but have no presence in India. Even so, CERT-In-empanelled auditors can be expected to apply this methodology, and the incident and logging duties it references already carry statutory force under the 2022 directions. In practice, the guidelines are likely to become the reference point for what a competent AI audit in India must cover, whatever their formal status.

Security Analysis

A Scored, Tiered Model of AI Risk

The core of the framework is a quantitative risk-scoring model. Each AI system is scored from 1 to 4 on eight inherent-risk factors: impact severity, autonomy, data sensitivity, scale and exposure, reversibility, sector criticality, adversarial exposure, and complexity and auditability [1][3]. The composite score, from 8 to 32, determines the tier, and systems can be reclassified upward when their use or exposure changes [2]. The table below consolidates the tier boundaries and minimum audit obligations set out in the primary text, supplemented by secondary coverage where noted. Some secondary sources summarize the model differently; MediaNama, for example, lists a longer set of factors [2], so readers should rely on the primary text where accounts diverge.

Tier	Composite score [1][3]	Audit obligation	Notes
Tier 1 – Low	8–12	Self-assessment or basic internal review with documented evidence [1][4]	Typically advisory or assistive systems with limited autonomy [2]
Tier 2 – Moderate	13–19	Documented internal audit plus independent audit by a CERT-In-empanelled auditor at least annually [1][2][3]	Consistently reported across sources
Tier 3 – High	20–26	Independent third-party audits with enhanced governance and monitoring, up to every six months where risk or material change warrants [1][2][4]	Frequency not stated in one legal analysis [3]
Tier 4 – Critical	27–32	Comprehensive audit with continuous monitoring and event-triggered reassessment, "rather than by the calendar" [1][3]	Testing includes advanced adversarial testing and red-team exercises [1]

Two design choices stand out. First, autonomy and reversibility are scored separately. An agent that executes irreversible actions, such as moving funds, deleting records, or changing production infrastructure, will accumulate risk points that a comparable read-only assistant would not. Second,

"complexity and auditability" is itself a scored factor [3]. Opaque systems are therefore pushed toward higher scrutiny, which gives organizations an incentive to design for auditability from the start. The model also defines remediation consequences. An unsatisfactory assurance opinion, or two or more Critical findings, triggers a mandatory re-audit within three months, or sooner if a regulator directs [1]. MediaNama further reports corrective action plans for Critical, High, and Moderate findings and monthly progress reporting on Critical and High items [2].

The model also leaves open questions that organizations and auditors will need to work through. Scoring each factor on a four-point scale involves judgment, and two assessors may reasonably score the same system differently, which makes consistency across audit firms a concern. Systems scored close to a tier boundary create an incentive to argue for the lower score. Whether India has enough CERT-In-empanelled auditors with AI expertise to meet demand, and what the audit cost burden will be for mid-sized organizations with several Tier 2 or Tier 3 systems, are practical questions the guidelines alone do not answer.

Nine Domains Across the Full Lifecycle

The guidelines organize audit coverage into nine domains: AI governance and accountability; data lifecycle, quality, and privacy; model risk management; ethics, bias, and fairness; explainability and transparency; security, adversarial resilience, and incident reporting; continuous monitoring, drift management, and validation; human oversight and decision accountability; and third-party and AI supply-chain assurance [1][2][3]. These domains apply across seven lifecycle stages, from design and procurement through decommissioning and retirement [3]. Audits are expected to document a system's intended use, operating boundaries, architecture, risk classification, and security requirements. Training data, configurations, prompts, and post-deployment changes should be traceable throughout the lifecycle [5][7]. Coverage describes the inventory of models, datasets, prompts, tools, and software components as an AI bill of materials [7], building on the AIBOM auditing that the 2025 audit policy introduced [12].

The breadth here matters for security teams in particular. Because security, adversarial resilience, and incident reporting form one domain within a governance-heavy framework, a security-only evidence package is unlikely to satisfy an auditor. Fairness, explainability, and human-oversight evidence will be assessed alongside penetration-test results and access reviews. Legal analysis also notes that some domain coverage is marked as recommended rather than mandatory, so the mandatory core may be narrower than the document's length suggests [3]. Organizations will need to read the final text closely to separate minimum requirements from good practice.

Agentic AI: Where the Audit Gets Harder

Agentic systems receive the most distinctive treatment. Beyond the general domains, the guidelines add requirements for generative AI, large language models, retrieval-augmented generation, multimodal systems, and autonomous agents [4][5]. Reported agent-specific audit points include scrutiny of tool use and excessive agency, least-privilege access controls, review of persistent memory, and assessment of unsafe autonomous actions [2]. Named risks include prompt injection, data poisoning, sensitive information leakage, insecure software components, excessive AI permissions, and unsafe communication between multiple agents [5]. News coverage also describes expectations for clear permission boundaries, human approval for sensitive actions, behavior monitoring, and emergency shutdown capability [7], details that readers should confirm against the primary text.

The guidelines reportedly ask auditors to assess whether the chosen human oversight model, whether human-in-the-loop, human-on-the-loop, or human-over-the-loop, fits the system's risk [2]. For agent builders, this changes the oversight question. It is no longer enough to state that a human is "involved." The organization will need to show that the placement of human review fits the system's autonomy and reversibility scores, and that the review actually takes place. For multi-agent architectures, the focus on inter-agent communication, including authentication between agents [1], suggests auditors will look for evidence about delegation chains, agent identity, and trust boundaries between agents, not only between users and models. In CSA's assessment, this is an area where many current deployments are likely to have gaps.

Incident Reporting and Forensic Readiness

The six-hour reporting window and 180-day log retention are not new obligations; they come from the 2022 directions [6]. What the AI guidelines add is their explicit application to AI security incidents, together with a 180-day minimum log retention requirement that calls for time synchronization across all AI components to support investigation [1][5][7]. Coverage notes that an ordinary model error or inaccurate output does not automatically count as a reportable incident [5]. Organizations still need a defined triage path to decide quickly whether a prompt-injection-driven data exfiltration, a poisoned retrieval corpus, or an agent's unauthorized action meets CERT-In's reportable criteria. Six hours leaves little time to classify a novel AI failure. Pre-agreed decision criteria, and logging that captures prompts, tool calls, retrieved context, and agent actions with correlated timestamps, are therefore operational prerequisites, not refinements.

Guarding the Auditors' Own Use of AI

One provision addresses a risk that is easy to overlook. The guidelines permit AI-assisted audit procedures, subject to conditions on scope, confidentiality, tool validation, traceability, and human review, but state that AI-generated outputs shall not replace sufficient and appropriate evidence or the auditor's independent judgment [1][5]. This matters most when AI tools are used to evaluate other AI systems, where errors in automated conclusions could go unnoticed. As audit firms and compliance platforms adopt LLM-based evidence review, the provision draws a line that will shape how those tools are used. AI may speed up evidence collection and analysis, but the audit opinion must rest on verifiable evidence and accountable human assessors.

What This Signals Globally

India's framework fits a broader movement toward risk-proportionate, evidence-based AI assurance. The EU AI Act sets conformity assessment, logging, human oversight, and post-market monitoring obligations for high-risk systems [8]. ISO/IEC 42001 provides a certifiable AI management system standard [9], and the NIST AI RMF organizes AI risk activities around its Govern, Map, Measure, and Manage functions [10]. CERT-In's approach differs from these in three ways that may influence other regulators. It assigns tiers using a scored, multi-factor model rather than enumerated use cases. It names agentic risks such as excessive agency and inter-agent communication directly. And it relies on an existing general cyber-incident regime with a six-hour reporting window, much shorter than the serious-incident timelines, measured in days, under Article 73 of the EU AI Act [8]. Whether other jurisdictions adopt similar scoring is uncertain. Still, the combination of independent audit, continuous monitoring for the highest-risk systems, and explicit agent controls suggests where global assurance expectations are heading.

For multinational organizations, the practical concern is duplicated work. A system serving Indian users may also be high-risk under the EU AI Act, inside an ISO/IEC 42001 certification scope, and subject to customer due diligence questionnaires. Each regime uses different terms, but most of the evidence they require is the same: inventories, risk assessments, access controls, test results, monitoring records, and oversight procedures. Mapping that evidence once to a shared control framework, and then cross-walking it to each regime, is in our assessment the most sustainable approach.

Recommendations

Immediate Actions

Organizations with AI systems in scope should first obtain the official guidelines from CERT-In [1] and confirm the clause-level requirements that apply to them, particularly the agent-specific checks that this note draws partly from secondary reporting. Next, they should build or update an inventory of AI systems that touch Indian users or data, including third-party and embedded AI services, and give each system a provisional score against the eight risk factors. Agentic systems with write access to production systems, financial functions, or sensitive data should be prioritized. These systems tend to score high on autonomy, reversibility, impact severity, and data sensitivity together, which can place them in Tier 3 or Tier 4. Finally, incident response teams should confirm that AI-specific scenarios are covered by six-hour reporting playbooks, and that prompts, tool calls, and agent actions are logged with synchronized timestamps and kept for at least 180 days [1][6].

Short-Term Mitigations

Over the next one to two quarters, organizations should close the agent-specific control gaps that auditors are most likely to examine. This work begins with enforcing least-privilege tool permissions and distinct agent identities, and with documenting the human oversight model chosen for each system along with a rationale tied to its risk score. Persistent memory and retrieval corpora should be brought under change control, and pre-deployment testing for prompt injection, data leakage, and unsafe tool use should be in place; CSA's Agentic AI Red Teaming Guide gives a practical test methodology for these scenarios [14]. In parallel, organizations should start building an AI bill of materials covering models, datasets, prompts, tools, and dependencies, and should extend vendor due diligence to AI suppliers in line with the guidelines' supply-chain assurance domain.

Strategic Considerations

Over the longer term, organizations should design AI systems for auditability rather than retrofitting evidence before each audit cycle. Because complexity and auditability is a scored factor, architectural choices that improve traceability may lower a system's tier and its audit burden. Organizations should also adopt one control framework as the backbone of a multi-regime evidence program and map CERT-In, EU AI Act, ISO/IEC 42001, and NIST AI RMF obligations to it. For Tier 4 systems, where continuous monitoring is expected, automating control monitoring can reduce the effort of change- and incident-

triggered reassessments. Organizations that use AI in their own compliance and audit work should keep human review and evidence verification in place for every AI-assisted finding, in line with CERT-In's position on AI-generated audit outputs.

CSA Resource Alignment

CSA's most directly relevant work is the set of **AICM v1.1 Auditing Guidelines**, which translate each AI Controls Matrix control into concrete verification procedures for specific actors in the AI supply chain. The *AICM v1.1 Auditing Guidelines for Model Providers (MP)* [15] cover the full model lifecycle, from data sourcing and training through validation, deployment, and retirement, and address much of the ground in CERT-In's data lifecycle, model risk management, and continuous monitoring domains. The *AICM v1.1 Auditing Guidelines for Orchestrated Service Providers (OSP)* [16] address organizations that coordinate multiple models, tools, and services, which is the architecture CERT-In's agentic and inter-agent requirements target. Because CERT-In's scope extends to AI that is procured and deployed, the *AICM v1.1 Auditing Guidelines for AI Customers (AIC)* [17] are relevant to the many organizations that consume rather than build AI, and the *AICM v1.1 Auditing Guidelines for Application Providers (AP)* [18] serve builders of AI-enabled applications. Auditors and auditees preparing for a CERT-In engagement can use these documents as a starting point for evidence collection and test procedures.

At the control-framework level, the **AI Controls Matrix (AICM) v1.1** [11] provides the common backbone recommended above. AICM's domains, including governance, data security, model security, supply-chain management, identity and access, logging and monitoring, and incident response, align closely with several of CERT-In's nine audit domains. Organizations should check coverage for CERT-In's ethics and fairness, explainability, and human-oversight domains in particular, since these were not cross-walked for this note. AICM is also mapped to other frameworks, including ISO/IEC 42001 and the EU AI Act, which supports the build-once, map-many approach that multinational organizations will need.

For independent assurance, CSA's **STAR for AI** program and the *Auditors' Guidance Document for STAR for AI Certification* [19] show how third-party auditors can assess AICM implementation alongside ISO/IEC 27001 and ISO/IEC 42001 using a management-capability maturity model. CERT-In's Tier 4 expectation of change- and incident-triggered reassessment is consistent with the approach described in *The Evolution of STAR: Introducing Continuous Auditing* [20], which describes a move from periodic point-in-time attestation to continuous assurance. Organizations operating critical AI systems should treat continuous auditing as a concrete design requirement.

References

[1] CERT-In. "Artificial Intelligence (AI) Security Auditing Guidelines (CISG-2026-04)." Indian Computer Emergency Response Team, November 4, 2026. Full text: AI_Security_Auditing_Guidelines.pdf.

[2] MediaNama. "CERT-In's AI Security Auditing Guidelines Set Out Tiered Framework." MediaNama, November 6, 2026.

[3] Sansa Legal. "CERT-In Issues AI Security Auditing Guidelines with a Four-Tier Risk Model." Sansa Legal, November 2026.

[4] Success Corner. "CERT-In Issues AI Security Audit Guidelines with Four Risk Tiers for AI Systems." Success Corner, November 5, 2026.

[5] The420.in. "CERT-In Issues AI Security Audit Guidelines, Sets Four Risk Categories." The420.in, November 4, 2026.

[6] CERT-In. "Directions under Sub-section (6) of Section 70B of the Information Technology Act, 2000." Indian Computer Emergency Response Team, April 28, 2022.

[7] NewsKart. "CERT-In's New AI Audit Framework: Why AI Agents May Face Tighter Security Checks." NewsKart, November 5, 2026.

[8] European Union. "Regulation (EU) 2024/1689 (Artificial Intelligence Act)." Official Journal of the European Union, July 12, 2024.

[9] ISO. "ISO/IEC 42001:2023 Information Technology — Artificial Intelligence — Management System." International Organization for Standardization, December 2023.

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

[11] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, 2026.

[12] CERT-In. "Comprehensive Cyber Security Audit Policy Guidelines." Indian Computer Emergency Response Team, July 25, 2025.

[13] MediaNama. "CERT-In Sets Cybersecurity Audit Rules for India." MediaNama, July 31, 2025.

[14] Cloud Security Alliance. "Agentic AI Red Teaming Guide." Cloud Security Alliance, 2025.

© 2026 Cloud Security Alliance. AI-assisted Rapid Research.

[15] Cloud Security Alliance. "AICM v1.1 Auditing Guidelines for Model Providers (MP)." Cloud Security Alliance, 2026.

[16] Cloud Security Alliance. "AICM v1.1 Auditing Guidelines for Orchestrated Service Providers (OSP)." Cloud Security Alliance, 2026.

[17] Cloud Security Alliance. "AICM v1.1 Auditing Guidelines for AI Customers (AIC)." Cloud Security Alliance, 2026.

[18] Cloud Security Alliance. "AICM v1.1 Auditing Guidelines for Application Providers (AP)." Cloud Security Alliance, 2026.

[19] Cloud Security Alliance. "Auditors' Guidance Document for STAR for AI Certification." Cloud Security Alliance, February 2026.

[20] Cloud Security Alliance. "The Evolution of STAR: Introducing Continuous Auditing." Cloud Security Alliance, 2024.