DSCI
PROMOTING DATA PROTECTION

Brief on

CERT-In AI Security Auditing Guidelines v1.0, November 2026



Brief on CERT-In AI Security Auditing Guidelines v1.0, November 2026

1. Overview

CERT-In (the Indian Computer Emergency Response Team) issued AI security auditing guidelines on the 4th of November 2026, highlighting risks that ordinary IT and cybersecurity audits do not fully cover. The document by the working committee at Cert-In lists poisoned training data, adversarial attacks, data leakage, bias, unexplainable outputs, model drift, deepfakes, hallucinations and weak human oversight.

Who it applies to. AI developed, procured, deployed or operated by organizations in India, AI that processes Indian data, and AI that serves Indian users. That includes MSMEs (micro, small and medium enterprises), large companies, government bodies and PSUs (public sector undertakings), startups, and Indian or foreign AI service providers. It covers in-house, purchased, open-source and API-based models alike. It excludes undeployed academic research, purely rule-based automation, and personal consumer use, though providers of consumer AI tools stay in scope.

The seven guiding principles:

1. **Security-by-Design:** Security is built in from design to retirement, not added after launch.
2. **Accountability and Human Oversight:** Every production AI system has a named owner, and people can approve or override consequential outputs.
3. **Transparency and Explainability:** Enough documentation and logging exists for auditors to understand what the system does and where it fails.
4. **Fairness and Bias Management:** Unintended performance gaps between groups are tested for, fixed and monitored.
5. **Data Security and Privacy:** Data is authorized, classified, protected, kept only as long as needed and disposed of securely.
6. **Auditability and Traceability:** You can reconstruct which model, data, prompt and setting produced any outcome.
7. **Resilience and Continuous Assurance:** Systems are monitored after launch, and drift or change triggers re-checks or rollback.

2. How the framework works?

Every AI system is placed in one of four risk tiers, and the tier sets how deep, how often and how independently it is audited. A single standard for all systems would overburden low-risk tools and under-govern high-impact ones.

The four tiers, with illustrative examples:

- **Low:** Advisory or assistive output with limited autonomy. *Example: an internal tool that summarizes meeting notes.*
- **Moderate:** Influences business processes under defined human oversight, with recoverable impact. *Example: a support chatbot that hands complex cases to staff.*
- **High:** Makes or materially influences consequential decisions. *Example: a bank's loan-approval model.*
- **Critical:** Critical infrastructure, national security or safety-critical use, where failure could be systemic or irreversible. *Example: AI balancing load on a power grid.*

The same technology can land in different tiers depending on sector and use.

The eight risk factors are each scored 1 (Low) to 4 (Critical), before giving credit for any controls: F1 impact severity, F2 autonomy and actionability, F3 data sensitivity, F4 scale and exposure, F5 reversibility and recoverability, F6 operational or sector criticality, F7 adversarial exposure, and **F8 complexity and auditability** (how hard the system is to understand and assess, so black boxes score higher). Missing evidence cannot justify a low score. Certain traits, such as autonomous safety-critical action or no way to interrupt or roll back, can push a system up a tier whatever its score. The Executive Summary of the released CERT-In guidelines* also highlights related considerations, such as transparency and incident-response readiness.

Tier (total score)	Primary audit	Frequency	AI Risk and Impact Assessment (AIIA)	Governance body
Low (8–12)	Self-assessment or basic internal review	Annual review	Optional, recommended	Designated AI owner
Moderate (13–19)	Documented internal audit	Continuous assessment, plus a CERT-In-empaneled audit at least annually	Mandatory, simplified	AI governance function
High (20–26)	Independent third-party audit by a CERT-In-empaneled organization	Continuous; every six months where risk or material change warrants	Mandatory, comprehensive	AI Governance Committee
Critical (27–32)	Independent CERT-In-empaneled audit with continuous monitoring	Comprehensive audit, continuous monitoring, event-triggered reassessment	Mandatory, comprehensive, independent review where appropriate	Board or senior-management oversight

Tiers are reassessed at least annually and after material changes, incidents, regulator directions, or a move from pilot to production. MSMEs get simpler paths for Low and Moderate systems, but no exemption for High or Critical ones.

The audit lifecycle has six phases:

1. **Engagement initiation:** Agree objectives, audit type, a signed engagement letter, and confirm the auditor has no conflict of interest.
2. **Risk assessment and scoping:** Review the AI inventory and independently re-check each tier; if auditor and organization disagree, the higher tier applies. The auditor then writes a Statement of Applicability, which lists which controls apply and why any are excluded.
3. **Detailed planning:** Set the test procedures, evidence requests and samples, such as prompts, agent tool calls and human overrides.
4. **Fieldwork:** Test each control for design and for real-world operation, since a policy document alone proves nothing. AI tests must be reproducible, recording the model hash, prompt version and settings.
5. **Analysis and findings:** Find root causes and rate severity.
6. **Reporting and follow-up:** Issue the report, track fixes and verify closure.

High and Critical audits need at least one AI/ML specialist and one certified auditor (CISA, CIA, ISO 27001 Lead Auditor or equivalent); single-auditor engagements are not allowed. Lack of access to a vendor's model is never treated as proof that it conforms. Auditors may use AI tools, but AI output is not evidence, and the human auditor stays accountable.

3. The nine core audit domains

The nine domains are the main control baseline every in-scope system is audited against, with depth set by tier. Governance and Security are mandatory even for Low-tier systems; all nine are mandatory from Moderate upward. For each domain: what is asked, what the auditor checks, and an everyday comparison.

3.1 Governance and Accountability. *Asks:* a board-approved AI policy, a cross-functional AI committee, a RACI matrix (who is Responsible, Accountable, Consulted, Informed), an AIIA, and a complete AI inventory. *Auditor checks:* the committee meets at least quarterly with minutes, and AIIAs exist for all Moderate-to-Critical systems before go-live. *Like:* a company's finance controls, which need a ledger and named signatories before anyone spends.

3.2 Data Lifecycle, Quality and Privacy. *Asks:* each dataset has an authorized purpose and approved source, plus classification, least-privilege access, lineage, poisoning protection, quality and representation checks, retention rules, and separation of training from live data. *Auditor checks:* data can be traced from source to model, and consent evidence exists where consent is an applicable requirement. *Like:* a food supply chain where you know the farm, the date and every handler.

3.3 Model Risk Management. *Asks:* a model registry, version control that blocks unauthorized changes, risk reviews, validation (independent of developers for High and Critical), model cards, a decommissioning process, and an AIBOM (AI Bill of Materials: a parts list of the model, datasets,

prompts, libraries and tools). *Auditor checks:* the registry includes pre-production models and decommissioning has been exercised. *Like:* an aircraft maintenance log.

3.4 Ethics, Bias and Fairness. *Asks:* an ethics policy, pre-launch bias tests using metrics such as disparate impact ratio and equalized odds, screening for proxy variables (features that indirectly stand in for a protected trait), mitigation and live monitoring. *Auditor checks:* counterfactual tests on high-impact models, where changing only a protected attribute must not change the outcome, plus live dashboards for customer-facing Moderate-to-Critical models. *Like:* a blind audition for orchestra musicians.

3.5 Explainability and Transparency. *Asks:* an explanation method suited to the model, with its limits documented, decision logs, plain-language explanations for people affected, and up-to-date model cards. *Auditor checks:* logs can rebuild individual decisions, and explanations are consistent across similar cases and understandable to non-experts. *Like:* a doctor who explains a diagnosis instead of only handing over a prescription.

3.6 Security, Adversarial Resilience and Incident Reporting. *Asks:* AI threat modelling, adversarial robustness tests, AI-specific VAPT (Vulnerability Assessment and Penetration Testing), model-file integrity checks, supply-chain scanning and API security. *Auditor checks:* AI VAPT done in the past 12 months, a tested CERT-In reporting process, and logs kept at least 180 days with synchronized clocks. *Like:* testing a lock against lock-picks, not just confirming it is fitted.

3.7 Continuous Monitoring, Drift and Validation. *Asks:* performance metrics; data-drift tests (statistical checks that live data still resembles training data); concept and fairness drift tracking; re-validation triggers; and controlled retraining. *Auditor checks:* automated drift alerts with thresholds, and retraining that passes approval gates. *Like:* a GPS that needs map updates when roads change.

3.8 Human Oversight and Decision Accountability. *Asks:* an oversight model for each system: HITL (human-in-the-loop, a person approves before the output takes effect), HOTL (human-on-the-loop, a person supervises and can intervene), or HOVL (human-over-the-loop, a person can constrain or shut the system down). Also, approval gates, override, emergency stop, reviewer training and override logs. *Auditor checks:* overrides are tested and re-tested, and decisions link to a named approver. *Like:* autopilot with a trained pilot who can actually take the controls.

3.9 Third-Party and AI Supply-Chain Assurance. *Asks:* vendor due diligence, provenance checks, AIBOM and SBOM (software bill of materials), contracts covering incident, vulnerability and change notification plus audit rights, monitoring of AI-as-a-Service, and exit plans. *Auditor checks:* contract clauses exist and fallback plans for critical vendors are tested. *Like:* outsourcing your kitchen still leaves you responsible for food poisoning.

4. Advanced AI, cybersecurity and contestability

Advanced AI. Generative AI, RAG (Retrieval-Augmented Generation, where a model looks up documents before answering), agentic, multimodal, multi-agent and autonomous or cyber-physical systems face extra requirements on top of the nine domains. The key threats:

- **Prompt injection:** Hidden instructions in user input, retrieved documents, tool output or media hijack the model. Input filtering alone is not accepted as the main defense.
- **Excessive agency:** An agent holds more permissions or tools than its task needs, so an error or attack becomes a real action.
- **Data and model poisoning:** Corrupted training, fine-tuning, RAG or memory data plants wrong behavior.
- **Vector and embedding weaknesses:** The document store behind a RAG system leaks across users or serves poisoned, stale or deleted content.
- **Unsafe tool use:** An agent calls unapproved or unauthenticated tools, plugins or MCP (Model Context Protocol) services.
- **Hallucination (misinformation):** Confident but false output, including fabricated citations.

Agents must not inherit user permissions automatically, and high-risk agents need emergency stop, credential revocation and rollback. Multi-agent systems need a unique identity per agent and controls against runaway delegation. Cyber-physical AI needs safety interlocks independent of the AI, a safe state and manual takeover. Organizations must also tell users when they are dealing with AI and label synthetic media.

Cybersecurity and resilience:

- **Infrastructure hardening:** Network segmentation, container security, GPU isolation, secrets management, AES-256 and TLS 1.2+ (preferably 1.3) or equivalent, and risk-based patching.
- **Secure MLOps/LLMOps** (the pipelines that build and deploy models): Segregated environments, signed model files, evaluation gates, separation of duties, rollback, and no automatic promotion of retrained models to production.
- **AI-specific VAPT:** Adds tests for model extraction, RAG, agents, memory, multimodal input, poisoning, prompt injection and privacy attacks, repeated after material change.
- **SOC integration** (Security Operations Centre): AI logs feed the SIEM (security event monitoring system), with AI detection rules, playbooks, annual analyst training, and an annual red team for Critical systems.
- **Incident response:** AI incidents meeting CERT-In reportable criteria **must be reported within 6 hours**. Logs are kept at least 180 days. High and Critical systems need tested rollback and business-continuity plans with RTO and RPO (recovery time and recovery point objectives).

Contestability is the ability to send an AI decision, recommendation or action for review, override, correction or escalation when it is questioned. Organizations must define review triggers such as low confidence or high impact, give reviewers enough information, and feed repeated overrides back into model governance. Auditors test whether approval gates can be bypassed through other interfaces and will not treat oversight as effective just because a manual procedure exists.

Contestability works as an organizational assurance mechanism: it makes sure consequential AI outputs pass through competent, well-informed and empowered human review before or after they take effect.

5. Reporting, remediation and maturity

Each severity level carries a defined response: Critical and High findings call for immediate escalation and containment, while Moderate and Low findings follow set remediation timelines.

- **Critical** (immediate or severe risk, such as discriminatory AI running in production): Immediate escalation to senior management, prompt containment or restriction of the system, and a Corrective Action Plan on an expedited basis.
- **High** (significant gap not needing shutdown, such as a broken human-override mechanism): Tracked through governance and risk escalation until closure.
- **Moderate** (control exists but is incomplete): Remediation plan within 30 business days and implementation within 60 days.
- **Low/Observation** (minor improvement): Addressed within 3 months or folded into continuous improvement, with no mandatory escalation.

A finding's severity alone does not trigger external reporting; that happens only if the condition independently meets an external requirement. Progress on Critical and High findings goes to the AI Governance Committee monthly, Moderate quarterly. Overdue Critical and High findings go to senior management or the Board. An "Unsatisfactory" opinion or two or more Critical findings triggers a mandatory re-audit within 3 months.

Corrective Action Plans are required for every Critical, High and Moderate finding. Each needs a named individual owner (not a team), target date, milestones, defined evidence and a verification method. A finding closes only after the auditor confirms the root cause, not just the symptom, is fixed.

Assurance opinions are Satisfactory, Satisfactory with Reservations, Requires Significant Improvement, or Unsatisfactory. They weigh materiality and systemic impact, not a count of findings.

Five maturity levels: (1) Initial/Ad Hoc, (2) Developing/Repeatable, (3) Defined/Standardized, (4) Managed/Measured, (5) Optimizing/Leading. Each domain is scored separately, and the overall level equals the weakest domain. Maturity never changes a finding's severity. Regulators and CERT-In may use it to set minimum expectations by sector or tier.

6. Standards alignment

Only one external standard, ISO/IEC 42001 (AI management systems), is labelled normative: the primary alignment standard, mapped through the audit methodology, the core domains and the reporting rules. The others are informative, and none becomes mandatory just by being listed. ISO/IEC 23894 guides AI risk management, and the NIST AI RMF (US AI Risk Management Framework) supplies its Govern, Map, Measure and Manage functions. ISO/IEC 27001 holders can extend their existing security management system to AI. MITRE ATLAS (a catalogue of real attacks on AI) and the OWASP Top 10 lists for ML, LLM and agentic applications supply threat categories. The glossary also draws definitions from the EU AI Act, the IT Act 2000, the DPDPA 2023 (Digital Personal Data Protection Act) and the CERT-In Directions 2022.

How the two Indian instruments fit together. MeitY's India AI Governance Guidelines (2025) answer "what should India's AI governance look like?"; this framework answers "how do we verify organizations are doing it?". MeitY's seven "sutras" map onto the seven principles, and the risk tiering delivers the India-specific risk assessment framework MeitY called for. The closing acknowledgement credits CERT-In's leadership and the working-group contributors.

7. What this means for Indian industry?

The biggest shift is that any AI system scoring Moderate or above needs an external audit by a CERT-In-empaneled organization. Every organization also has to inventory and risk-score all of its AI first.

Impact by group

- **Large enterprises:** Comprehensive obligations, with independent third-party audits for High and Critical systems.
- **Regulated sectors** (BFSI, meaning banking, financial services and insurance, plus health, telecom and critical infrastructure): High data-sensitivity and criticality scores (F3, F6) push many systems into High or Critical. RBI, SEBI, IRDAI and TRAI are named as intended users.
- **Government and PSUs:** The same tiering applies, including to citizen services and law enforcement.
- **AI developers and SaaS vendors:** Audited both as providers and as suppliers to their customers. Customers must demand AIBOMs, notification clauses and audit rights.
- **Startups and MSMEs:** Simplified paths for Low and Moderate systems, with no size-based exemption for High or Critical.
- **CERT-In-empaneled auditors:** A new mandated workload, needing AI/ML specialists on every High or Critical audit.