SRIDA IT PROFESSIONAL REFERENCE PUBLICATION

AI Security Auditing

CERT-In AI Security Auditing Guidelines v1.0

Requirements, Implementation Guidance & Audit Testing Procedures

DOCUMENT ID	VERSION	STATUS
SRIDA-CERTIN-AISA-RTP-001	1.0	Final — Approved for Publication

Rajendra Bodda
Chief Information Security Officer (CISO)
Srida IT · 2026

Independent professional guidance.
Not a CERT-In publication.

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-certin-ai · © 2026 Srida IT



CERT-In AI Security Auditing Guidelines v1.0

Requirements, Implementation Guidance & Audit Testing Procedures

SAMPLE

Prepared by

Rajendra Bodda

Chief Information Security Officer (CISO)

Published by Srida IT

SRIDA-CERTIN-AISA-RTP-001 · Version 1.0 · November 2026

This publication is an independently developed professional guidance document prepared by Srida IT. It is not a CERT-In publication, is not endorsed by or affiliated with the Indian Computer Emergency Response Team (CERT-In) or the Ministry of Electronics and Information Technology, does not replace the CERT-In AI Security Auditing Guidelines, and does not confer any approval, empanelment, certification or assurance opinion.

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-certin-ai · © 2026 Srida IT

Document Information and Control

Field	Value
Document title	CERT-In AI Security Auditing Guidelines v1.0
Subtitle	Requirements, Implementation Guidance & Audit Testing Procedures
Document identifier	SRIDA-CERTIN-AISA-RTP-001
Version	1.0
Status	Final — Approved for Publication
Classification	Srida IT Professional Reference Publication
Author	Rajendra Bodda, Chief Information Security Officer (CISO)
Publisher	Srida IT
Publication year	2026
Issue date	November 2026
Reference document	CERT-In Artificial Intelligence (AI) Security Auditing Guidelines, Version 1.0, 04 November 2026 (CERT-In Sub-group 03, Working Group on Indian Standards for Cybersecurity Auditing), 107 pages
Related documents	CERT-In Directions under section 70B(6) of the IT Act, 2000 (28 April 2022); India AI Governance Guidelines (MeitY, 2025); ISO/IEC 42001:2023; NIST AI RMF 1.0; Digital Personal Data Protection Act, 2023
Companion publications	SRIDA-AIMS-42001-RTP-001; SRIDA-AIS-27090-RTP-001; SRIDA-ISMS-27001-RTP-001; SRIDA-DPDP-RTP-001; SRIDA-PIMS-27701-RTP-001
Next review	On publication of a new version of the CERT-In AI Security Auditing Guidelines, and at least annually

Version history

Version	Date	Author	Description of change
1.0	November 2026	Rajendra Bodda	First published edition, approved for publication.

Approval

Role	Name	Date
Approved for publication	Rajendra Bodda, Chief Information Security Officer (CISO)	9 November 2026

Distribution

Released for publication by Srida IT. Distribution and use are subject to the permitted-use terms in the Copyright, Legal and Usage Notices.

© 2026 Srida IT · Version 1.0 · Independent guidance, not a CERT-In publication

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-certin-ai · © 2026 Srida IT

Table of Contents

Part I Foundations 19

Chapter 1 Introduction to the CERT-In AI Security Auditing Guidelines 20
Chapter 2 Structure of the Guidelines and How This Manual Maps to Them 25
Chapter 3 Scope, Applicability, AI System Types and Exclusions 30
Chapter 4 Risk Tiering and Audit Intensity 37
Chapter 5 Framework Architecture and the Seven Guiding Principles 45
Chapter 6 Terminology, Definitions and Abbreviations 50
Chapter 7 How to Use This Manual 58

Part II Audit Methodology 63

Chapter 8 Engagement Initiation, Auditor Competence and Independence 64
Chapter 9 Risk Assessment, Scoping, the Statement of Applicability and External Requirements 71
Chapter 10 Detailed Audit Planning, Evidence Requests and Sampling 78
Chapter 11 Fieldwork, Evidence, Technical Test Records and Scope Limitations 83
Chapter 12 AI-Assisted Audit Procedures 88

Part III Core Audit Domains 93

Chapter 13 Domain 8.1 — AI Governance & Accountability 94
8.1-C1 AI Governance Policy 96
8.1-C2 AI Governance Committee 102
8.1-C3 Roles and Responsibilities 108
8.1-C4 AI Risk and Impact Assessment (AIIA) 113
8.1-C5 AI System Inventory 120
8.1-C6 Policy Review and Update 127
8.1-C7 External Requirements Management 132
Chapter 14 Domain 8.2 — Data Lifecycle, Quality & Privacy 138
8.2-C1 Data Authorisation and Purpose Control 140
8.2-C2 Data Classification and Access Control 146
8.2-C3 Data Provenance and Lineage 151
8.2-C4 Data Integrity 156
8.2-C5 Data Quality 161
8.2-C6 Bias and Representation Assessment 166
8.2-C7 Retention and Secure Disposal 171
8.2-C8 Training / Inference Data Separation 176
Chapter 15 Domain 8.3 — Model Risk Management (MRM) 181
8.3-C1 Model Registry 183
8.3-C2 Version Control 189
8.3-C3 Model Risk Identification 194
8.3-C4 Model Validation 199
8.3-C5 Model Performance Monitoring 206
8.3-C6 Model Decommissioning 211
8.3-C7 Model Documentation 216
8.3-C8 AI Bill of Materials (AIBOM) 221
Chapter 16 Domain 8.4 — AI Ethics, Bias & Fairness 228

SAMPLE

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-certin-ai · © 2026 Srida IT

8.4-C1 Ethics Policy 230
8.4-C2 Pre-Deployment Bias Testing 236
8.4-C3 Proxy Variable Screening 242
8.4-C4 Bias Mitigation Techniques 247
8.4-C5 Ongoing Fairness Monitoring 252
8.4-C6 Adversarial Fairness Audit 257

Chapter 17 Domain 8.5 — Explainability & Transparency 262
8.5-C1 Explainability Technique Selection 264
8.5-C2 Decision Logging 270
8.5-C3 User-Facing Explanations 276
8.5-C4 Auditor Explanation Access 281
8.5-C5 Transparency Reporting 286
8.5-C6 Consistency Verification 291
8.5-C7 Explanation Quality 296

Chapter 18 Domain 8.6 — Security, Adversarial Resilience & Incident Reporting 301
8.6-C1 AI Threat Modelling 303
8.6-C2 Adversarial Robustness Testing 310
8.6-C3 AI-Specific VAPT 316
8.6-C4 CERT-In Incident Reporting 321
8.6-C5 Model Integrity Monitoring 328
8.6-C6 Supply Chain Security 333
8.6-C7 API Security 338
8.6-C8 Infrastructure Hardening 343

Chapter 19 Domain 8.7 — Continuous Monitoring, Drift Management & Validation 348
8.7-C1 Performance Monitoring 350
8.7-C2 Data Drift Detection 355
8.7-C3 Concept Drift Detection 360
8.7-C4 Fairness Drift Monitoring 365
8.7-C5 Re-Validation Triggers 370
8.7-C6 Retraining Governance 375

Chapter 20 Domain 8.8 — Human Oversight & Decision Accountability 380
8.8-C1 Oversight Model 382
8.8-C2 Approval Gates 388
8.8-C3 Manual Override 393
8.8-C4 Emergency Stop / Safe State 400
8.8-C5 Reviewer Competence 405
8.8-C6 Override Logging 410
8.8-C7 Effectiveness Monitoring 415

Chapter 21 Domain 8.9 — Third-Party and AI Supply-Chain Assurance 420
8.9-C1 Vendor Due Diligence 422
8.9-C2 Model and Component Provenance 428
8.9-C3 AIBOM/SBOM 433
8.9-C4 Contractual Security Requirements 438
8.9-C5 Provider Change Management 443
8.9-C6 AI-as-a-Service Monitoring 448
8.9-C7 Dependency and Vulnerability Monitoring 453
8.9-C8 Exit, Portability and Fallback Strategy 458

Part IV Advanced AI Security Requirements 463

Chapter 22 Generative AI (Sections 9.1–9.5) 464

SAMPLE

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-certin-ai · © 2026 Srida IT

CHAPTER 1

Introduction to the CERT-In AI Security Auditing Guidelines

This chapter explains where the CERT-In *Artificial Intelligence (AI) Security Auditing Guidelines* come from, what they set out to achieve, whom they address, and what kind of instrument they are. It closes by describing what this Srida IT manual adds to them: testing procedures, evidence expectations, templates and crosswalks that turn the Guidelines' control expectations into an executable audit programme.

1.1 Origin of the Guidelines

The Indian Computer Emergency Response Team (CERT-In) constituted a Working Group on "Indian Standards for Cybersecurity Auditing" to develop a standardised, India-specific approach to cybersecurity auditing. Within that Working Group, Sub-group 03 prepared the AI Security Auditing Guidelines. Version 1.0 of the document is dated 04 November 2026 and runs to 107 pages across sixteen sections, from the Preamble to an Acknowledgement that records the mentorship of CERT-In leadership and the contribution of Working Group members drawn from CERT-In and from private-sector security firms.

The document's cover carries the identifier "AI Security Auditing Guidelines v1.0" and a reference to the CERT-In AI Cyber Defence Centre. The Preamble separately mentions a dedicated AI cyber defence "War Room" that CERT-In has set up to bring together technical capability for coordinated threat assessment and response, and explains that the framework draws on the operational experience gained through that and similar initiatives. Readers should note that the two names are used in different places; the Guidelines do not explain whether they denote the same facility.

Throughout this manual, "the Guidelines" means that Version 1.0 document. Where this manual refers to a section, it uses the Guidelines' own section numbers (for example, Section 5.4 for the scoring methodology). Chapter 2 explains why this manual cites by section rather than by table number.

SAMPLE

1.2 Purpose of the Guidelines

The Guidelines set out to give Indian organisations and their auditors a common, risk-based way of examining whether AI systems are governed, secured and operated responsibly across their whole lifecycle. In Srida IT's reading, the purpose has five practical strands:

- **One audit baseline regardless of sourcing.** The same expectations apply whether a model is built in-house, licensed from a vendor, taken from an open-weight repository or consumed through a cloud API. The deploying organisation cannot escape scrutiny by outsourcing the model.
- **Proportionate assurance.** The depth, frequency, independence and technical intensity of an audit are driven by an assessed risk tier rather than applied uniformly (Section 5).
- **AI-specific threat coverage.** The audit is expected to reach threats that a conventional IT audit tends to miss: poisoning of data or models, adversarial evasion, model theft and inversion, prompt injection, abuse of tools by agents, and compromise of the model supply chain.
- **Lifecycle verification.** Controls are examined from design and procurement through training, validation, deployment, monitoring, retraining and eventual retirement.
- **Consistency and interoperability.** The Guidelines aim for repeatable scoping, evidence handling, testing, finding classification and reporting, and for results that can be related to recognised international standards.

The Guidelines express these aims through six strategic objectives in Section 4.2: audit consistency, risk-based assurance, technical resilience, lifecycle traceability, continuous assurance and standards interoperability. Chapter 3 discusses them in the context of scope.

© 2026 Srida IT · Version 1.0 · Independent guidance, not a CERT-In publication

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-certin-ai · © 2026 Srida IT

1.3 The AI cyber defence context

The Preamble frames the Guidelines against a simple observation: AI is no longer a niche technology but a component of banking, healthcare, telecommunications, education, transport, manufacturing, policing, defence and citizen services. AI systems now predict, recommend, decide and generate content, and some are permitted to act on their own in digital or physical environments.

The Preamble then argues that the resulting risks are not fully handled by conventional IT and cybersecurity assurance. It lists, among others, manipulated training data and models, adversarial inputs, unauthorised access to models, leakage of sensitive information, algorithmic bias, opaque decision-making, drift, weak third-party dependencies, misuse of generative and agentic capability, synthetic media and hallucination, and insufficient human oversight. These can affect not only the classical confidentiality, integrity and availability triad but also safety, fairness, reliability and accountability.

Two consequences follow for the auditor. First, an AI audit under the Guidelines is not a narrow penetration test, nor an ethics review, nor a management-system certification audit; it combines organisational accountability, technical testing, data and model governance, monitoring, human oversight, incident response and third-party assurance. Second, the Guidelines position AI security auditing alongside CERT-In's operational work on AI-enabled threats. Section 14.4, for example, points auditors to CERT-In advisories and to the CERT-In *Blueprint for Reducing Exposure and Defending against AI-Assisted Vulnerabilities Exploitation*, which addresses digital infrastructure, and expects auditors to check whether frontier-AI threats are reflected in the auditee's threat modelling, testing and incident response. Chapter 25 of this manual covers those technical references.

> **Two different subjects of "AI" in the Guidelines**
>
> The Guidelines deal with AI in two roles. Most of the document treats AI as the *subject* of the audit. Section 7.10 deals with AI as an *aid* to the auditor (AI-assisted audit procedures) and sets safeguards for that use. This manual keeps the two apart: Chapters 13 to 29 cover auditing AI systems; Chapter 12 covers the auditor's own use of AI tools.

1.4 Intended audiences

Section 4.3 of the Guidelines identifies a wide stakeholder community. Srida IT groups it into five audiences, each of which uses the Guidelines differently.

Table 1-1. Audiences of the Guidelines and how each typically uses them

Audience	Who is included (as the Guidelines describe them)	Typical use
Regulators and public bodies	CERT-In, MeitY and sectoral regulators such as RBI, SEBI, IRDAI and TRAI	A reference point when framing sector expectations or directions on AI assurance; source of the classification guidance that can trigger tier reassessment (Section 5.6)
Auditors	Internal and external auditors; CERT-In-empanelled auditing organisations assessing AI-based systems	Audit criteria, scoping logic, tier-driven intensity, finding severities, report and opinion structure
Governance, risk and compliance functions	Chief Risk Officers, compliance heads, AI governance committees, Data Protection Officers, information security teams	Designing the AI governance programme, the AI inventory, the AI Risk and Impact Assessment (AIIA) and the tiering process
Builders and operators	Data scientists, ML engineers, AI architects, product owners, MLOps/LLMOps and platform teams, cybersecurity, VAPT and SOC teams	Engineering controls: provenance, versioning, adversarial testing, monitoring, rollback, logging and integrity protection
Suppliers	AI vendors, SaaS and cloud providers, data-labelling firms, outsourced development partners	Supplying the assurance evidence their customers will be asked for, and preparing for audit in their own right as AI service providers

Section 4.3 also describes auditors as conducting "certification evaluations" under the framework. The Guidelines do not, however, establish or describe any certification scheme, accreditation body or certificate. Srida IT therefore treats an audit

© 2026 Srida IT · Version 1.0 · Independent guidance, not a CERT-In publication

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-certin-ai · © 2026 Srida IT

SRIDA IT REF. 8.8-C3 · GUIDELINES SECTION 8.8

Manual Override

CERT-In requirement WP-8.8-C3

A Requirement Reference

CERT-In Artificial Intelligence (AI) Security Auditing Guidelines, Version 1.0, 04 November 2026, 8.8 Human Oversight & Decision Accountability

Source pages 58–60.

Summaries are in Srida IT's own words and do not reproduce CERT-In text. Refer to the Guidelines for the authoritative wording.

Ref.	Control area	Class	What the Guidelines ask (Srida IT summary)
8.8-C3	Manual Override	CERT-IN REQUIREMENT	The Guidelines require that authorised people are able to stop an AI action in progress, change it, or replace it with their own decision. The auditor checklist adds that override and intervention mechanisms are tested periodically and re-tested after material changes that affect them. Section 12.1 gives a non-functional human override mechanism as an example of a High finding.

B Auditor Checklist (Guidelines)

- Confirm that override and intervention mechanisms are tested on a defined cycle and re-tested after any material change affecting them.
- Confirm that override capability is implemented and tested in proportion to the system's risk, autonomy and potential impact.
- Confirm, as part of contestability testing (Section 11.3), that authorised personnel can reject, correct or override AI outputs and that reviewer permissions are appropriately restricted.
- Confirm that override accuracy is periodically measured (shared with 8.8-C7).

Checklist items restated in Srida IT's words.

SAMPLE

C Expectations by Risk Tier

Tier 1 · Low	Tier 2 · Moderate	Tier 3 · High	Tier 4 · Critical
Domain 8.8 is recommended at Tier 1, and the Guidelines suggest basic intervention or override capability where applicable. Srida IT accepts a documented manual workaround (for example, disregarding an advisory output) as sufficient.	Table 58 lists override capability and escalation for Tier 2, and Section 11.4 states that override shall be available at this tier, so Srida IT treats it as a requirement. Srida IT recommends at least an annual functional test and a test after each material change.	The Guidelines require override to be implemented and periodically tested. Override must be technically effective (the overridden outcome is what executes), authorised and logged.	The Guidelines require emergency override to be independently tested, together with intervention, safe-state, fallback and recovery. Srida IT reads 'independently' as testing by a party not responsible for building or operating the mechanism; the external auditor may perform or observe it.

D Principles and Standards Alignment

Guiding principles served

- P2 Accountability and Human Oversight
- P3 Transparency and Explainability

Standards alignment

- CERT-In stated: ISO/IEC 42001 Clause 5 (Leadership)
- CERT-In stated: ISO/IEC 42001 Annex A.3 (Internal organisation)
- CERT-In stated: EU AI Act (human oversight)
- CERT-In stated: NIST AI RMF GOVERN function
- Srida IT crosswalk: ISO/IEC 42001:2023 A.9.2 (processes for responsible use of AI systems)

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-certin-ai · © 2026 Srida IT

- Srida IT crosswalk: ISO/IEC 42001:2023 A.6.2.6 (AI system operation and monitoring)
- Srida IT crosswalk: ISO/IEC 27001:2022 A.8.2 (privileged access rights) and A.8.32 (change management)
- Srida IT crosswalk: NIST AI RMF MANAGE 2.4 (mechanisms to supersede, disengage or deactivate AI systems)
- Srida IT crosswalk: EU AI Act Article 14 (ability to disregard, override or reverse output, and to interrupt)
- Srida IT crosswalk: OWASP LLM Top 10 2025 LLM06 Excessive Agency

"CERT-In stated" items reproduce the alignment the Guidelines give for the domain. "Srida IT crosswalk" items are indicative only; mapping a control does not by itself establish conformity with any standard or law.

E Objective

When an authorised person decides that an AI decision, output or action is wrong, unsafe or inappropriate, that person can stop it, change it or substitute their own decision quickly enough to prevent the harm the risk assessment anticipates; the human decision is what actually takes effect downstream; the AI cannot silently reassert its own outcome; and only authorised people can exercise the override.

F Interpretation

Override has three distinct capabilities, and the auditor tests each one separately: interrupt (halt a decision or action before or during execution), modify (amend the AI output, for example a reduced credit limit or an edited response, before it takes effect) and override (replace or reverse the AI outcome with a human decision, including after execution where reversal is possible). Section 5.3 tells assessors to test approval and override effectiveness separately when scoring F2, and Srida IT applies the same separation in fieldwork: a working approval gate (8.8-C2) proves nothing about override, because override is needed precisely when something has already slipped past or through the gate. Emergency stop of the whole system is 8.8-C4; this control concerns intervention in individual decisions and actions.

Technical effectiveness. An override button that writes a flag nobody reads is the classic non-functional mechanism. The auditor traces an override through to the system of record: does the downstream ledger, ticket, message or actuator reflect the human decision? Common failure patterns are (a) the AI recomputes on the next batch run and overwrites the human value, (b) a cache or queue already holds the AI action and executes it anyway, (c) an agent treats the override as one more input and re-plans to the same end, and (d) a second integration consumes the original model output directly. For agentic systems, override must reach the tool executor, and Section 9.7 expects interruption, credential revocation and rollback for high-risk agents.

Automation bias and rubber-stamping. An override capability that is never used may be a sign of a good model or of reviewers who defer to it. Automation bias is the tendency to accept machine output even when contrary evidence is available; rubber-stamping is its procedural form, where approval or non-intervention becomes the default. The Guidelines name automation bias in the context of AI-assisted auditing (Section 7.10) and, in Section 11.2, require organisations to assess periodically whether reviewers are accepting AI outputs automatically and inappropriately. The auditor therefore looks at the override rate, the time spent before accepting, and whether the user interface makes overriding harder than accepting (extra screens, mandatory free-text justification only for overrides, pre-ticked acceptance). A design that charges a cost only for disagreeing is a design that discourages override. Srida IT recommends that override be as easy to execute as acceptance, that justification be required for both on a sample basis, and that seeded test cases (known-wrong outputs) be inserted periodically to measure whether reviewers catch them. Detailed measurement belongs to 8.8-C7 and 11.2-C6; this control checks that the mechanism does not itself create the bias.

Authority and bypass. Override is a privileged function. It must be restricted to authorised roles, segregated from those who benefit from the outcome, and unavailable through side channels such as direct database edits, administrator consoles or alternate APIs that change the outcome without the override record. The Section 11.3 instruction to test whether mandatory approval gates can be bypassed through alternate interfaces applies equally in reverse here: the auditor checks both that authorised staff can override and that unauthorised routes to change outcomes do not exist. For bought or API-consumed models, override sits in the deployer's workflow; the provider's model cannot be overridden, but its output can be, and the auditor confirms the deployer has not delegated that capability to the vendor's interface alone.

Production versus staging. Testing override in production gives the strongest evidence but can cause real effects; testing only in staging is safe but may miss production-only integrations, caches and permissions. Srida IT recommends a layered approach agreed in the rules of engagement (Appendix C): functional and failure-mode tests in a production-equivalent staging environment with configuration parity demonstrated by comparing exported configuration and version hashes; then, in production, either observation of a real override performed by the auditee during fieldwork, or a controlled override on a designated test record or synthetic transaction that the auditee has authorised and can reverse, with a named auditee

SAMPLE

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-certin-ai · © 2026 Srida IT

contact present and a rollback step agreed in advance. Where production testing is refused for a Tier 3 or Tier 4 system and parity cannot be shown, the auditor records a scope limitation under Section 7.9, not a pass.

G Applicability and Scope

Applies to all systems in which a human may need to stop, change or reverse an AI decision or action, which covers every Tier 2 to Tier 4 system and Tier 1 systems where the risk assessment identifies an action interface. For purely advisory Tier 1 outputs, the ability to disregard the advice is the override and needs only to be documented. Override cannot be marked not applicable for an agentic system with write access or for a cyber-physical system; for the latter, Section 9.10 manual takeover applies in addition. Irreversible actions (for example a payment already settled or a physical action completed) do not remove the requirement; they shift the emphasis to interrupt-before-execution and to the speed of override, and the auditor records which actions are irreversible.

H Implementation Guidance

1. Document, for each action type, which override capabilities exist (interrupt, modify, override after the fact), the authorised roles, the maximum time within which the override must take effect, and whether the action is reversible.
2. Implement override at the system of record or execution layer so that the human decision is authoritative; mark overridden records so that subsequent model runs, batch jobs or agents cannot overwrite them.
3. For agentic systems, implement interrupt at the orchestrator and tool proxy, with credential revocation and rollback for high-risk agents, and confirm the agent cannot resume the cancelled plan without a new human decision.
4. Design the reviewer interface for symmetry: overriding should take no more effort than accepting, and acceptance should not be pre-selected.
5. Restrict override to authorised roles through access control, prevent self-override where segregation is required, and remove or monitor side channels (direct data edits, administrator tools) that change AI outcomes.
6. Log every override with the original AI output, the human decision, the reason, the actor, the time and the downstream effect (see 8.8-C6).
7. Maintain an override test plan covering each capability and each action type, with test cases for the failure patterns of re-assertion, cached execution and secondary consumers.
8. Run the test plan on a defined cycle (Srida IT suggests annually at Tier 2, six-monthly at Tier 3, quarterly at Tier 4) and after every material change to the model, workflow, integration or permissions.
9. Use seeded known-wrong outputs to check that reviewers detect and override them, and feed the results to 8.8-C5 and 8.8-C7.
0. For Tier 4, arrange independent testing by a party separate from the build and operating teams, and keep the test records under Section 7.5 reproducibility standards.

I Roles and Responsibilities

Role	Responsibility
AI system owner	Owns the override design, test plan and test cycle; ensures re-testing after material change.
Authorised reviewer / operator	Exercises override when an AI output or action is wrong or unsafe and records the reason.
Platform engineering	Implements override at the execution layer and prevents AI re-assertion and side-channel changes.
Identity and access management	Restricts override rights to authorised roles and reviews them periodically.
Independent tester (Tier 4)	Performs or witnesses override tests independently of the build and operating teams.
Auditor	Tests design and operating effectiveness within the agreed rules of engagement.

J Documentation and Records

- Override procedure per system, listing capabilities, action types, authorised roles, time limits and reversibility
- Override test plan and test records, including re-tests after material change
- Access-control matrix for override rights and periodic access reviews
- Override logs for the audit period
- Change records identifying changes that affected the override mechanism

SAMPLE

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-certin-ai · © 2026 Srida IT

- Rules of engagement for production testing, with rollback steps
- Configuration-parity evidence where staging is used for testing

K Audit Testing Procedures

Recommended Srida IT audit methodology; the Guidelines set the control expectations but do not prescribe these procedures. Assess design effectiveness and operating effectiveness separately (Guidelines Section 7.5); record a missing evidence item as a scope limitation, not as not applicable.

Table 20-6. Audit testing procedures, Control area 8.8-C3

Test ID	Method	Testing procedure	Result
TP-8.8-C3-01	EXAMINE	Obtain the override procedure and confirm it separately defines interrupt, modify and after-the-fact override for each action type, with authorised roles, time limits and reversibility; record any action type with no override route.	☐ Effective ☐ Partially effective ☐ Ineffective ☐ N/A · scope limit
TP-8.8-C3-02	EXAMINE	Inspect the override test plan and records for the period; confirm tests ran on the defined cycle and after each material change identified in the change log, and that each record states system version, model version or hash, configuration version, test case, expected and actual result.	☐ Effective ☐ Partially effective ☐ Ineffective ☐ N/A · scope limit
TP-8.8-C3-03	TEST	In a production-equivalent environment agreed in the rules of engagement, perform an interrupt, a modify and an override for each sampled action type; trace each to the system of record and confirm the human decision is what took effect. Record environment, versions, payload, timestamps and the observed downstream state.	☐ Effective ☐ Partially effective ☐ Ineffective ☐ N/A · scope limit
TP-8.8-C3-04	TEST	Execute the failure-pattern cases: trigger the next model or batch run after an override and confirm the human value persists; cancel an agent action already queued and confirm it does not execute; check secondary consumers of the model output for the original value.	☐ Effective ☐ Partially effective ☐ Ineffective ☐ N/A · scope limit
TP-8.8-C3-05	OBSERVE	In production, observe an auditee operator perform an override on a designated test record or synthetic transaction (or a real case arising during fieldwork), with rollback agreed in advance; record time from decision to effect and compare with the documented limit.	☐ Effective ☐ Partially effective ☐ Ineffective ☐ N/A · scope limit
TP-8.8-C3-06	TEST	Attempt to override using an account without override rights, and attempt to change an AI outcome through an alternate interface (direct API, administrator console, data edit); confirm refusal or, where the route exists, that it is logged and monitored.	☐ Effective ☐ Partially effective ☐ Ineffective ☐ N/A · scope limit
TP-8.8-C3-07	INSPECT	Inspect configuration-parity evidence (exported configuration, version hashes, integration lists) for any test performed only in staging; where parity is not shown for a Tier 3 or Tier 4 system, record a scope limitation.	☐ Effective ☐ Partially effective ☐ Ineffective ☐ N/A · scope limit
TP-8.8-C3-08	INSPECT	Review the reviewer interface for asymmetry that discourages override (extra steps, pre-ticked acceptance, justification required only for overrides) and record observations for 8.8-C7.	☐ Effective ☐ Partially effective ☐ Ineffective ☐ N/A · scope limit
TP-8.8-C3-09	SAMPLE	Select a sample of overrides from the log and confirm each was made by an authorised user, carries a reason, and is reflected in the downstream record.	☐ Effective ☐ Partially effective ☐ Ineffective ☐ N/A · scope limit
TP-8.8-C3-10	RE-PERFORM	Recalculate override rate and override accuracy (share of overrides later confirmed correct) for the sample, and compare with any figures the auditee reports.	☐ Effective ☐ Partially effective ☐ Ineffective ☐ N/A · scope limit

L Sampling Guidance

Two populations apply: action types (for functional testing) and override events (for operating effectiveness). Srida IT recommends functional testing of every consequential action type at Tier 3 and Tier 4 and of the two highest-impact action types at Tier 2. For override events, sample 25 at Tier 2, 40 at Tier 3 and 60 at Tier 4 across the audit period, stratified by

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-certin-ai · © 2026 Srida IT

reviewer and action type; where fewer events exist, test the whole population and treat a very low count as a possible automation-bias indicator rather than as comfort. Test records must cover the full period and every material change in it.

M Interview Questions

Interviewee	Question
Operator or reviewer	Show me how you would stop this action right now. What happens if you are wrong to override, and what happens if you were wrong not to?
AI system owner	Which changes in the last year affected the override path, and where are the re-test records for each?
Platform engineer	After an operator overrides, what stops the next model run or the agent from putting the original outcome back?
Risk owner	Which actions cannot be reversed once taken, and how is override arranged for those?
Independent tester (Tier 4)	Who designed the test cases, and did you have authority to test in production?

N Technical and System Verification

- Compare the system-of-record value after override with the AI output and the human decision in the override log.
- Inspect batch and scheduler configuration for jobs that rewrite AI-derived fields, and confirm overridden records are excluded.
- Inspect agent orchestrator settings for cancel, interrupt and resume behaviour and for credential revocation on interrupt.
- Enumerate roles and accounts holding override rights and compare with the authorised list.
- Search administrator and database audit logs for direct edits to AI-derived fields outside the override workflow.

O Expected Audit Evidence

- Override procedure and capability matrix
- Override test plan and dated test records with version identifiers
- Auditor technical test records (Appendix E format) for each test performed
- Rules-of-engagement extract and rollback evidence for production tests
- Configuration-parity comparison for staging tests
- Override log sample and downstream-record screenshots
- Access-control matrix and access-review records
- Reviewer interface screenshots

P Evaluation Criteria

Indicators: effective

- Interrupt, modify and override work for each consequential action type, take effect within the documented limit, and persist downstream.
- Only authorised roles can override, side channels are closed or monitored, and every override is logged.
- Tests run on cycle and after each material change, with Tier 4 testing performed independently.

Indicators: partially effective

- Override works but tests are overdue or were not repeated after a material change.
- Override works but the interface discourages its use, or reasons are not captured.

Indicators: ineffective

- An override is recorded but the AI outcome still takes effect or is reasserted.
- No working override route exists for a consequential action type, or unauthorised users can change outcomes.

Q Typical Finding Severity (Guidelines Section 12.1)

- Critical: in a Tier 4 system an operator cannot stop a consequential action in progress, or the AI reasserts its outcome after override, and the condition has occurred in production, creating uncontrolled consequential actions.
- High: a non-functional human override mechanism in a Tier 3 or Tier 4 system, which Section 12.1 itself cites as a High example, even where no harm is evidenced in the period.

SAMPLE

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-certin-ai · © 2026 Srida IT