CERT-In WORKING GROUP
INDIAN COMPUTER EMERGENCY RESPONSE TEAM

Artificial Intelligence (AI) Security Auditing Guidelines

Framework for Secure, Ethical and Auditable Artificial Intelligence Systems

Document Identifier	AI Security Auditing Guidelines v1.0
Prepared By	CERT-In Sub-group 03 – AI Security Auditing Guidelines Working Group
Version	1.0
Date	04 November 2026



CERT-In AI Cyber Defence Centre

Table of Contents

1. Preamble 3
2. Executive Summary 4
3. Definitions & Glossary 7
4. Purpose & Scope 14
5. Applicability Matrix & Risk Tiering 22
6. Framework Architecture & Guiding Principles 29
7. Audit Planning & Scoping Methodology 35
8. Core Audit Domains 46
9. Advanced AI Security Requirements 62
10. Cybersecurity & Resilience 70
11. Contestability and Human Decision Review 79
12. Audit Reporting & Remediation 82
13. Maturity Level Model 88
14. Referenced Frameworks & Standards 94
15. Integration with the India AI Governance Guidelines 102
16. Acknowledgement 107

1. Preamble

Artificial Intelligence has evolved from a specialised technology into an important component of digital services, business processes, public administration and critical infrastructure. AI systems are increasingly being used in banking & financial services, healthcare, telecommunications, education, transportation, manufacturing, law enforcement, defence, citizen-service delivery and other sectors to generate predictions, recommendations, decisions and content, and in certain cases to autonomously initiate actions in physical or digital environments.

While AI offers significant opportunities for innovation, efficiency and improved service delivery, its adoption also introduces risks that are not fully addressed through conventional information technology and cybersecurity assurance mechanisms. Such risks include manipulation of training data and models, adversarial attacks, unauthorised model access, leakage of sensitive information, algorithmic bias, lack of explainability, model and data drift, insecure third-party dependencies, misuse of generative and agentic AI, deepfakes, hallucinated outputs and inadequate human oversight. These risks may affect the confidentiality, integrity, availability, safety, fairness, reliability and accountability of AI-enabled systems.

The Artificial Intelligence Security Auditing Guidelines have been developed by Sub-group 03 of the Working Group on “Indian Standards for Cybersecurity Auditing” constituted by the Indian Computer Emergency Response Team (CERT-In). The Guidelines establish a common, risk-based and lifecycle-oriented framework for assessing the governance, security, privacy, fairness, transparency, resilience and auditability of AI systems. They are intended to facilitate consistent security auditing practices across sectors while supporting responsible innovation, indigenous capability development and alignment with applicable Indian legal, regulatory and policy requirements and internationally recognised AI governance and cybersecurity standards.

The Guidelines recognise that AI governance cannot be addressed solely through ethical principles or conventional cybersecurity controls. Effective assurance requires organisational accountability, technical testing, data and model governance, continuous monitoring, human oversight, incident response and third-party assurance. Accordingly, the framework provides auditable requirements proportionate to the nature, scale, autonomy and potential impact of each AI system.

CERT-In has also taken initiatives to strengthen coordinated cybersecurity monitoring, threat assessment, technical analysis through the establishment of a dedicated AI cyber defence War Room bringing together relevant technical capabilities for coordinated assessment and response. The framework further draws upon the operational experience and institutional capabilities developed through such initiatives, providing a practical foundation for strengthening the implementation of risk-based AI security auditing and continuous assurance practices.

2. Executive Summary

The **Artificial Intelligence Security Auditing Guidelines** provide a structured framework for conducting security and governance audits of AI systems developed, procured, deployed or operated by organisations in India, as well as AI systems processing Indian data or providing AI-enabled services to Indian users.

The framework adopts a risk-based approach under which AI systems are classified into four risk tiers—Low, Moderate, High and Critical based on factors such as impact severity, autonomy of decision-making, data sensitivity, scale and exposure, transparency, Explainable AI (XAI), incident response readiness, reversibility and recoverability, operational/sector criticality, adversarial exposure and complexity/auditability. The assigned risk tier determines the depth, frequency and independence of the audit, the extent of documentation required, the level of human oversight and the applicable monitoring, reporting and assurance requirements.

The Guidelines establish nine core audit domains:

i. AI Governance and Accountability

ii. Data Lifecycle, Quality and Privacy

iii. Model Risk Management

iv. AI Ethics, Bias and Fairness

v. Explainability and Transparency

vi. Security, Adversarial Resilience and Incident Reporting

vii. Continuous Monitoring, Drift Management and Validation

viii. Human Oversight and Decision Accountability

ix. Third-Party and AI Supply-Chain Assurance

These domains collectively cover the entire AI lifecycle, including design, procurement, data collection, model development, training, validation, deployment, operation, retraining, monitoring and decommissioning. The framework supplements these domains with dedicated requirements for advanced AI architectures, including Generative AI, Retrieval-Augmented Generation (RAG), Agentic AI, Multimodal AI, Multi-Agent AI and Autonomous/Cyber-Physical AI. These requirements address prompt injection, sensitive-information disclosure, excessive agency, data and model poisoning, supply-chain risks, vector and embedding weaknesses, unsafe tool use, persistent-memory risks, inter-agent communication, multimodal attacks, misinformation and improper output handling.

The Guidelines also prescribe a structured audit methodology covering engagement initiation, risk assessment, scoping, audit planning, evidence collection, technical testing, finding

classification, reporting, remediation and follow-up verification. Audit findings are classified as Critical, High, Moderate or Low/Observation, with corresponding containment, remediation, escalation and reporting requirements. A five-level maturity model enables organisations to assess their current AI governance capability and plan progressive improvement from ad hoc practices to measured and continuously optimised governance.

AI-assisted audit procedures may support evidence analysis, control mapping, technical testing and report preparation. Their use is subject to approved scope, confidentiality controls, tool validation, traceable evidence and competent human review. AI-generated outputs shall not replace sufficient and appropriate evidence or the auditor's independent judgement.

The framework is aligned with recognised standards and frameworks, including CERT-In's Blueprint for Reducing Exposure and Defending against AI-Assisted Vulnerabilities Exploitation in Digital Infrastructure, India AI Governance Guidelines by MeitY, ISO/IEC 42001, ISO/IEC 23894, ISO/IEC 27001, NIST AI RMF, MITRE ATLAS and OWASP guidance for machine learning and large language model applications. The framework also provides a structured mechanism for identifying, assessing and mapping applicable CERT-In requirements, sector-specific cybersecurity obligations, contractual security requirements and other external requirements into the audit criteria and Statement of Applicability.

The Guidelines are intended for use by government organisations, public sector entities, critical infrastructure operators, regulated entities, private organisations, AI developers, AI service providers, internal and external auditors, CERT-In empanelled cybersecurity auditing organisations, risk and compliance functions and other stakeholders involved in the development, procurement, deployment or oversight of AI systems.

2.1 What This Framework Provides

This document delivers a comprehensive, actionable and internationally aligned auditing framework structured across the following pillars:

Table 1. Framework Pillars

Framework Pillar	Description
Risk Tiering and Applicability	A four-tier risk classification system (Low, Moderate, High, Critical) calibrated to sector, use case, and impact severity, determining audit intensity and compliance obligations.
Nine Core Audit Domains	Structured audit requirements spanning AI governance, data lifecycle, model risk management, ethics and fairness, explainability, cybersecurity resilience, continuous monitoring,human oversight, and third-party accountability.

Advanced AI Security Requirements	Dedicated audit provisions for large language models, Generative AI, RAG, Agentic AI, Multimodal AI, Multi-Agent AI and Autonomous/Cyber-Physical AI, addressing prompt injection, hallucination, content integrity, training data provenance, and output attribution.
Audit Planning and Scoping Methodology	A risk-based audit lifecycle from engagement planning through fieldwork, evidence collection, and reporting, aligned with ISO 42001 and ISACA/IIA methodologies.
Audit Reporting and Remediation	Structured finding classification, corrective action planning, ISO 42001 Clause 9–10 integration for management review and continual improvement, and escalation protocols.
AI-Assisted Audit Procedures	Controlled use of AI tools to support audit work, with defined authorisation, data protection, tool validation, evidence traceability, human review and auditor accountability.

2.2 Guiding Principles

The framework is anchored on seven foundational principles for secure, ethical and auditable AI systems align with MeitY's India AI Governance guiding principles:

i. Security-by-Design

ii. Accountability and Human Oversight

iii. Transparency and Explainability

iv. Fairness and Bias Management

v. Data Security and Privacy

vi. Auditability and Traceability

vii. Resilience and Continuous Assurance

3. Definitions & Glossary

This section establishes the standardised terminology used throughout the AI Security Auditing Guidelines. Consistent definitions are essential to ensure uniform interpretation by auditors, auditees, AI system owners, security teams and other relevant stakeholders across all audit domains and compliance contexts.

Where definitions are drawn from international standards (ISO/IEC 42001:2023, ISO/IEC 23894:2023, NIST AI RMF 1.0, EU AI Act) or Indian legislation (IT Act 2000, DPDPA 2023, CERT-In Directions 2022), the source is indicated. Where no single authoritative definition exists, this framework provides a harmonised definition synthesised from multiple standards for the Indian regulatory context.

Table 2. Definitions and Glossary

Term	Definition	Source
A. CORE AI CONCEPTS AND TERMINOLOGY		
Artificial Intelligence (AI)	An engineered system that generates outputs such as predictions, recommendations, decisions, or content, that can influence physical or virtual environments. AI systems are designed to operate with varying levels ofautonomy.	ISO/IEC 42001; OECD
AI Lifecycle	The sequence of stages an AI system undergoes from initial conception and requirements definition through design, development, verification, deployment, operation, monitoring, and eventual decommissioning or retirement.	ISO/IEC 42001
AI Management System(AIMS)	The set of interrelated or interacting elements of an organisation to establish policies, objectives, and processes to achieve those objectives, with respect tothe responsible development, provision, or use of AI systems.	ISO/IEC 42001
AI System	A machine-based system that, for explicit or implicit objectives, infers from the input it receives how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.	EU AI Act; OECD
Autonomous System	An AI system that can perform tasks, make decisions,or take actions in the physical or digital world without continuous human supervision, based on its programming, learned behaviour, or environmental inputs.	NIST AI RMF

Foundation Model	A large-scale AI model trained on broad, diverse datathat can be adapted (fine-tuned) to a wide range of downstream tasks. Includes large language models (LLMs), vision models, and multimodal models.	Framework; NIST
Generative AI	AI systems capable of generating novel content—including text, images, code, audio, video, or synthetic data—based on patterns learned from training data. Includes large language models, diffusion models, and generative adversarial networks.	Framework
Hybrid AI System	A system that combines multiple AI approaches (e.g.,rule-based logic with machine learning, or symbolic AI with neural networks) or integrates AI with non-AI decision components to produce outputs.	Framework
Inference	The process by which a trained AI model applies its learned parameters to new, previously unseen inputdata to generate predictions, classifications, recommendations, or other outputs.	Framework
Large Language Model (LLM)	A type of foundation model trained on extensive text corpora using self-supervised learning, capable of generating, summarising, translating, and reasoning over natural language text. Examples include GPT-series, LLaMA, and PaLM-series models.	Framework
Machine Learning (ML)	A set of techniques that enable computer systems to learn from data and improve their performance on a specific task without being explicitly programmed, including supervised, unsupervised, semi-supervised, and reinforcement learning.	ISO/IEC 22989
Model	A mathematical or computational construct that processes input data and produces outputs (predictions, classifications, decisions, or generated content). In this framework, “model” refers to AI/ML models unless otherwise specified.	Framework

B. DATA GOVERNANCE AND QUALITY		
Consent Management	The system and process by which an organisation obtains, records, manages, and enables revocation ofdata principal consent for the collection, processing, and use of personal data, in compliance with DPDPA 2023.	DPDPA 2023
Data Drift	A change in the statistical properties or distribution ofinput data over time, relative to the data on which themodel was trained, potentially degrading model performance and fairness.	Framework

Data Lifecycle	The stages through which data passes from initial collection or creation through processing, storage, use (including AI training and inference), retention, archival, and eventual deletion or anonymisation.	ISO/IEC 5259
Data Lineage	The documented record of data's origin, movement, transformation, and consumption across systems and processes, enabling traceability from source to AI model input and output.	Framework
Data Originator	The party that created the data and that may hold legalor contractual rights over it, distinct from the data subjects described by the data.	ISO/IEC 5259
Data Provenance	The documented history of data including its origin, collection methodology, transformations applied, chainof custody, and quality assessments, providing a verifiable record for audit purposes.	Framework
Data Quality	The degree to which data possesses attributes—including accuracy, completeness, consistency, timeliness, representativeness, and traceability—that make it fit for its intended use in AItraining or inference.	ISO/IEC 25012
Feature Engineering	The process of selecting, transforming, and constructing input variables (features) from raw data to improve AI model performance, including the identification and exclusion of sensitive or proxy attributes.	Framework
Personally Identifiable Information (PII)	Information that can be used to identify a specific individual, either directly (name, Aadhaar number) or indirectly (when combined with other data). Used interchangeably with personal data in this framework.	IT Act 2000; DPDPA
Representativeness	The degree to which a dataset accurately reflects the characteristics, distributions, and diversity of the real-world population or phenomenon that the AI model is intended to serve.	ISO/IEC 5259
Synthetic Data	Data generated artificially rather than directly collected from real-world events or individuals, designed to reproduce selected statistical, structural or behavioural characteristics of source data. Synthetic data shall not automatically be assumed to be anonymous or free from privacy risk; leakage, memorisation and re-identification risks shall be assessed where applicable.	Framework

C. RISK, SECURITY, AND ADVERSARIAL CONCEPTS		
Adversarial Attack	A deliberate attempt to manipulate the behaviour of anAI system by crafting inputs designed to cause misclassification, evasion, or degradation of model performance.	MITRE ATLAS

Adversarial Example	An input to an AI model that has been intentionally perturbed (often imperceptibly to humans) to cause the model to produce an incorrect or unintended output.	NIST AI RMF
AI Risk	The composite effect of the probability of occurrenceand the magnitude of impact associated with an AI system's potential failure, misuse, or unintended consequence across technical, ethical, legal, and operational dimensions.	ISO 23894
AI-Specific Vulnerability	A weakness in an AI system—in its data, modelarchitecture, training process, deployment environment, or operational logic—that could beexploited to compromise the system's integrity, confidentiality, availability, or fairness.	Framework
Attack Surface	The totality of points (data inputs, APIs, model endpoints, training pipelines, infrastructure) at which anunauthorised user can attempt to enter data to, extract data from, or interact with an AI system.	OWASP
Concept Drift	A change in the underlying relationship between input features and the target variable that the model was trained to predict, causing model performance to degrade even if input data distributions remain stable.	Framework
Data Poisoning	An adversarial attack that corrupts the training data ofan AI model by injecting, modifying, or removing datapoints, with the intent to compromise model accuracy,fairness, or security.	MITRE ATLAS
Deepfake	AI-generated or AI-manipulated audio, image, or video content that convincingly depicts a person saying or doing something they did not, typically created using generative adversarial networks or diffusion models.	Framework
Evasion Attack	An adversarial technique applied at inference time, where carefully crafted inputs are designed to bypass the model's detection or classification capabilities without altering the model itself.	MITRE ATLAS
Hallucination	An output generated by a generative AI system (particularly LLMs) that is factually incorrect, fabricated, or unsupported by the model's training data or input context, presented with apparent confidence.	Frame work
Model Drift	A degradation in an AI model's predictive performance, fairness metrics, or reliability over time, caused by changes in input data distributions (data drift), underlying relationships (concept drift), or operational context.	Frame work
Model Extraction Attack	An adversarial technique that aims to reconstruct or approximate a target model's parameters, architecture, or decision boundaries by systematically querying the model and analysing its outputs.	MITRE ATLAS

Model Inversion Attack	An attack that exploits model outputs to infer sensitive information about the training data, potentially reconstructing individual data records or revealing protected attributes.	MITRE ATLAS
Model Poisoning	An adversarial attack that directly manipulates the model's parameters, architecture, or training process (as distinct from data poisoning, which targets training data) to embed malicious behaviour.	MITRE ATLAS
Prompt Injection	An attack on large language models or generative AI systems where malicious instructions are embedded in user inputs (direct injection) or in retrieved context (indirect injection) to hijack the model's behaviour.	OWASP
Risk Tier	A classification level (Low, Moderate, High, Critical) assigned to an AI system based on its assessed risk profile, which determines the applicable audit intensity, compliance obligations, and governance controls under this framework.	Framework
Supply Chain Compromise	An attack that targets the AI development or deployment supply chain—including pre-trained models, open-source libraries, datasets, or third-party APIs—to introduce vulnerabilities or malicious components.	MITRE ATLAS
Threat Landscape	The comprehensive set of current and potential adversarial threats, attack vectors, vulnerability classes, and threat actors relevant to AI systems within a defined scope.	Framework

D. ETHICS, FAIRNESS, AND EXPLAINABILITY

Algorithmic Bias	Systematic and repeatable errors in an AI system's outputs that create unfair outcomes for specific groups,arising from biased training data, flawed feature selection, biased model architecture, or feedback loop effects.	IEEE P7003
Algorithmic Discrimination	The unjust or prejudicial differential treatment of individuals or groups by an AI system based on protected attributes (gender, caste, religion, ethnicity, disability, age) or their proxies.	EU AI Act
Bias Metric	A quantitative measure used to assess the degree of bias in an AI model's outputs, including statistical paritydifference, equalised odds, disparate impact ratio, and demographic parity.	Framework
Disparate Impact	A legal and statistical concept where an AI system's outputs disproportionately affect a protected group, even when the system does not explicitly use protectedattributes as inputs.	Framework
Explainability	The property of an AI system whereby its functioning, outputs, and the factors contributing to a specific decision can be understood and articulated in human-comprehensible terms by relevant stakeholders.	ISO/IEC 42001

Fairness	The principle that an AI system's outputs should not result in unjust, discriminatory, or disproportionately adverse outcomes for individuals or groups, assessed through quantitative metrics and qualitative review.	NIST AI RMF
Human-in-the-Loop (HITL)	A system design paradigm where human decision-makers are actively involved in the AI system's decision process, with the ability to review,modify, approve, or override AI-generated outputs before they take effect.	EU AI Act
Human-on-the-Loop(HOTL)	A system design paradigm where humans maintain supervisory oversight of the AI system's operation withthe ability to intervene or override during the operationcycle, but are not involved in every individual decision.	EU AI Act
Human-over-the-Loop (HOVL)	A system design paradigm where humans have no direct involvement in individual AI decisions but can intervene to constrain, modify, or shut down the system at any point through governance controls.	EU AI Act
Interpretability	The degree to which a human can understand the cause of an AI model's decision. Interpretable models(e.g., decision trees, linear regression) provide inherently transparent logic, as opposed to post-hoc explanations.	Framework
LIME	Local Interpretable Model-Agnostic Explanations – an explainability technique that approximates the behaviour of a complex model locally around a specificprediction using a simpler, interpretable model.	Framework
Proxy Variable	A feature in an AI model's input data that, while not directly representing a protected attribute (e.g., race, gender), is statistically correlated with it and can serveas an indirect basis for discriminatory outcomes.	Framework
SHAP	SHapley Additive exPlanations – a game-theoretic approach to explain the output of any machine learning model by assigning each feature an importance value (Shapley value) for a particular prediction.	Framework
E. AUDIT AND GOVERNANCE		
AI Governance Committee	A cross-functional organisational body responsible for overseeing AI strategy, policy enforcement, risk management, ethical review, model approvals, and audit coordination.	Framework
AI Impact Assessment(AIIA)	A documented assessment of the security, safety, data, model-performance, privacy, fairness, explainability, operational, human-oversight and third-party risks associated with an AI system, performed before deployment and following material changes where applicable.	Framework

Audit Evidence	The information and documentation collected during anaudit to support audit findings, conclusions, and opinions, including system logs, model documentation, test results, configuration records, and interview transcripts.	ISACA; IIA
Audit Finding	A statement of fact identified during an audit that indicates a gap, non-conformity, weakness, or observation relative to the applicable audit criteria, control objectives, or regulatory requirements.	ISO 19011; Framework
Audit Trail	A chronological record of system activities, decisions, data transformations, and user interactions that provides documentary evidence of the sequence of events and enables forensic reconstruction.	Framework
Maturity Level	A defined stage in an organisational capability model that characterises the degree of sophistication, standardisation, and effectiveness of an organisation's AI governance practices.	Framework; CMMI
Model Card	A standardised documentation artifact that describesan AI model's intended use, performance characteristics, training data, known limitations, bias evaluations, and ethical considerations.	Google; Framework
Model Registry	A centralised, authoritative repository that records all AI models within an organisation, including metadata on model purpose, ownership, version history, risk classification, and deployment status.	Framework
Model Risk Management(MRM)	A structured set of policies, procedures, governance requirements, and processes to identify, assess, mitigate, monitor, and report on risks arising from the design, development, deployment, and use of AI/ML models.	Framework; SR 11-7
Rollback	The process of reverting an AI model or system to a previous known-good version or state in response to identified failures, performance degradation, or security incidents.	Framework
AI-Assisted Audit	An audit in which approved AI tools assist defined procedures, such as document analysis, control mapping, log analysis or report drafting, while the auditor retains responsibility for evidence validation, professional judgement and the final assurance opinion. This is distinct from an audit whose subject is an AI system.	Framework

4. Purpose & Scope

This section defines the objectives, boundaries and applicability of the AI Security Auditing Guidelines. It establishes what the framework aims to achieve, who it applies to, which AI systems and data categories fall within scope.

4.1 Purpose

The purpose of these Guidelines is to establish a consistent, risk-based and auditable mechanism for evaluating the security, governance and responsible operation of AI systems throughout their lifecycle.

The framework aims to:

> **Ensure secure, ethical and auditable AI adoption:** Define consistent audit requirements applicable to AI systems regardless of whether they are developed internally, procured from third parties, consumed through APIs or cloud services, or derived from open-source or foundation models.

> **Enable risk-based assurance:** Determine the depth, frequency and technical intensity of an audit based on the AI system's potential impact, autonomy, data sensitivity, scale, criticality, adversarial exposure and explainability.

> **Assess AI-specific security risks:** Provide procedures for assessing threats including data and model poisoning, adversarial manipulation, model extraction, model inversion, prompt injection, unsafe tool use, supply-chain compromise, insecure AI APIs and unauthorised model modification.

> **Verify AI lifecycle controls:** Assess governance and security controls covering data preparation, model development, validation, deployment, operation, monitoring, retraining, change management and decommissioning.

> **Improve audit consistency and reproducibility:** Establish common expectations for audit scoping, evidence collection, technical testing, sampling, working papers, finding classification, reporting and remediation verification.

> **Support interoperability with recognised standards:** Facilitate alignment with relevant AI management, cybersecurity, risk-management and technical-assurance standards and frameworks.

4.2 Strategic Objectives

The framework is designed around the following technical and assurance objectives:

Table 3. Strategic Objectives

Strategic Objective	Description
Audit Consistency	Establish a common methodology and baseline for conducting AI security audits across different sectors, technologies and deployment models.
Risk-Based Assurance	Apply proportionate audit depth and technical testing based on the assessed risk and impact of the AI system.
Technical Resilience	Assess the ability of AI systems to resist adversarial attacks, misuse, failures, unauthorised modification and supply-chain compromise.
Lifecycle Traceability	Ensure that data, models, configurations, prompts, system changes and operational decisions can be traced and independently verified throughout the AI lifecycle.
Continuous Assurance	Verify that AI systems remain secure, reliable and within approved operating parameters after deployment through monitoring, drift detection, periodic validation and change-triggered reassessment.
Standards Interoperability	Enable audit results and control implementations to be mapped to recognised AI governance, cybersecurity, risk-management and technical-assurance standards and frameworks.

4.3 Intended Audience

This framework is designed for use by a broad range of stakeholders across the AI governance ecosystem:

Table 4. Intended Audience and Stakeholder Roles

Stakeholder Group	Relevance
Regulatory Bodies	CERT-In, MeitY, RBI, SEBI, IRDAI, TRAI and other sectoral regulators responsible for AI oversight and enforcement.
Internal and External Auditors	Professionals conducting AI security audits, compliance assessments and certification evaluations under this framework.
AI Governance and Risk Functions	Chief Risk Officers, Chief Compliance Officers, AI Governance Committees, Data Protection Officers and Information Security teams.
AI Practitioners / Developers	Data scientists, ML engineers, AI architects, and product owners responsible for AI system design, development and deployment.
Service Providers	AI vendors, SaaS platforms, cloud providers, data labelling firms and outsourced AI development partners.
Cybersecurity Auditing organisations	CERT-In empanelled auditing organisations involved in conducting security assessment of AI-based systems.

Cybersecurity Teams	CISOs, security architects, VAPT teams, SOC teams, incident-response teams and application-security functions responsible for protecting AI systems and their supporting infrastructure.
MLOps/LLMOps and Platform Teams	Teams responsible for model-development pipelines, deployment, model registries, prompt/configuration management, AI infrastructure, monitoring, retraining and lifecycle operations.

4.4 Organisational Applicability

This framework applies universally to all types and sizes of organisations operating AI systems within India or processing Indian citizens' data:

Table 5. Organisational Applicability

Organisation Type	Applicability and Obligations
Micro, Small, and Medium Enterprises (MSMEs)	Organisations adopting AI-enabled automation, customer analytics, or operational optimisation tools. Compliance obligations are proportionate to the risk tier of their AI deployments, with simplified audit pathways available for low-risk systems.
Large Corporations and Multinational Enterprises	Organisations deploying AI at scale across multiple business lines, including BFSI, healthcare, telecom, manufacturing, and retail. Subject to comprehensive audit obligations including independent third-party audits for high-risk and critical AI systems.
Government Agencies and Public Sector Units (PSUs)	Central and state government bodies and PSUs leveraging AI in citizen services, e-governance, law enforcement, defence, or public administration. Audit intensity shall be determined using the risk-tiering methodology defined in Section 5, considering the potential impact, system autonomy, data sensitivity, operational criticality, adversarial exposure and recoverability of the AI system.
Startups and Innovation Labs	Entities developing, piloting, or commercialising AI products and services. The framework provides a structured compliance pathway that supports innovation while ensuring responsible deployment practices.
AI Service Providers and Platform Operators	Indian and international entities offering AI-as-a-Service (AIaaS), AI-integrated SaaS platforms, cloud-hosted AI APIs, or AI development tools to Indian customers. Subject to audit obligations both as providers and as third-party vendors to their customers.

4.5 AI Systems Covered

The Guidelines cover AI systems regardless of their underlying technology, model architecture, development method, procurement source or deployment environment. Systems within scope include, but are not limited to:

Table 6. AI System Types Covered

AI System Type	Description	Examples
Predictive AI	Systems that use historical data and statistical or machine learning models to forecast future outcomes, trends, or probabilities.	Credit scoring, demand forecasting, risk modelling, churn prediction, disease progression models.
Generative AI	Systems capable of generating novel content including text, images, code, audio, video, or synthetic data based on learned patterns.	LLM-based chatbots, code generation, content creation, report drafting, synthetic data generation, synthetic/deepfake media generation.
Hybrid AI Systems	Systems combining multiple AI approaches (rule-based + ML, symbolic + neural) or integrating AI with non-AI decision components.	Credit decisioning engines (rules + ML), hybrid fraud detection, knowledge graph + NLP systems.
Embedded AI	AI capabilities integrated into hardware devices, operational technology, IoT endpoints, or medical devices.	AI-enabled medical imaging devices, autonomous drones, smart manufacturing sensors, connected vehicles.
AI-as-a-Service (AIaaS)	AI capabilities consumed via cloud APIs, SaaS platforms, or managed services where the AI model operates in the provider's environment.	Cloud NLP APIs, AI-powered analytics platforms, managed computer vision services, speech-to-text services.
Multimodal AI Systems	AI systems capable of processing, interpreting, correlating or generating information across two or more data modalities, including text, image, audio, video, code, documents, sensor data or other digital inputs.	Vision-language assistants, systems analysing medical images with clinical text, audio-video content analysis, document intelligence systems and text-to-image or text-to-video generators
Agentic AI Systems	AI systems capable of independently planning and executing a sequence of actions, invoking tools or services, accessing external systems, retaining contextual memory and adapting their actions to achieve defined objectives with varying levels of human supervision.	AI agents that execute business workflows, software-development agents, automated research agents, customer-service agents initiating transactions and security agents performing investigation or remediation actions.
Retrieval-Augmented Generation Systems	AI systems that retrieve information from external knowledge sources and provide the retrieved information as contextual input to a generative model. The system includes document-ingestion pipelines, embedding models, retrieval mechanisms, vector databases and underlying knowledge repositories.	Enterprise knowledge assistants, policy and legal document chatbots, medical information assistants, customer-support systems using internal documents and AI search systems connected to organisational databases.

Foundation Models	Large-scale AI models trained on broad datasets and capable of supporting multiple downstream applications through prompting, fine-tuning, adapters or other customisation mechanisms.	Large language models, vision foundation models, speech models, code-generation models and multimodal foundation models.
Multi-Agent AI Systems	AI architectures comprising multiple agents that communicate, coordinate, delegate tasks or collectively perform actions to achieve defined objectives.	Multi-agent cybersecurity systems, software-development agents, autonomous workflow agents and distributed decision agents.
Autonomous and Cyber-Physical AI Systems	AI systems capable of initiating actions that affect physical systems, operational processes or safety-critical environments with limited or no continuous human intervention.	Autonomous robotics, drones, industrial-control AI, AI-enabled medical devices, smart-grid systems and autonomous navigation systems.

4.6 Model Source Coverage

AI systems shall remain within audit scope regardless of model origin, ownership or delivery mechanism. Audit coverage shall be based on the deploying organisation's ability to identify, assess and manage the risks introduced by the model.

Table 7. Model Source Coverage

Model Source	Description
Internally Developed Models	Models designed, trained, fine-tuned or substantially modified by the organisation.
Third-Party Proprietary Models	Commercial or proprietary models procured from external providers and deployed either locally or through managed environments.
Open-Source / Open-Weight Models	Models, weights, code or components obtained from public or community repositories.
AI-as-a-Service / API Models	AI capabilities consumed through APIs, cloud services or managed platforms where the model remains under provider control.

4.7 Data and Model Source Coverage

The audit shall cover data and model artefacts used during training, validation, testing, fine-tuning, retrieval, inference, monitoring and retraining.

Table 8. Data and Model Artefacts within Audit Scope

Data / Artefact Category	Audit Requirements
Training, Validation and Test Data	Verify source, ownership/authorisation, provenance, version, integrity, quality, representativeness, transformations and access controls.
Inference and Operational Data	Verify classification, authorised use, access control, protection in transit and at rest, retention, monitoring and secure disposal.
Retrieved / RAG Knowledge Sources	Verify source provenance, ingestion controls, access permissions, indexing, document/chunk-level authorisation, freshness and protection against poisoned or malicious content.
Synthetic and Derived Data	Verify generation methodology, provenance, quality, bias, leakage risk and suitability for the intended use.
Model Artefacts	Verify model version, cryptographic hash, source, signing status, architecture/configuration, approval status and storage protections.
Prompts and Guardrail Configurations	Verify version control, access restriction, change approval, security testing and traceability of system prompts, templates and guardrails.
Embeddings and Vector Stores	Verify access control, tenant isolation, integrity, retention, deletion propagation and protection against unauthorised retrieval.
Logs and Audit Records	Verify completeness, integrity, time synchronisation, access protection, retention and ability to support reconstruction of security-relevant events.

4.8 AI Lifecycle Coverage

The framework covers the entire AI lifecycle, ensuring no stage escapes governance. Audit obligations apply at every phase:

Table 9. AI Lifecycle Audit Coverage

Lifecycle Stage	Scope of Audit Obligations
Design and Procurement	Intended use, system boundaries, architecture, risk classification, AI risk and impact assessment, threat modelling, security requirements, vendor due diligence and contractual assurance requirements.
Data Collection and Preparation	Data authorisation, classification, provenance, lineage, integrity, quality, representativeness, access control and protection against poisoning or unauthorised modification.

Model Development and Training	Model architecture, development methodology, training configuration, feature engineering, model/data versioning, development-environment security, code and dependency controls, and training-process traceability.
Validation and Testing	Model performance validation, bias/fairness assessment, adversarial robustness testing, explainability assessment, security testing, regression testing and approval prior to production deployment.
Deployment and Operations	Secure deployment, access control, model integrity verification, API security, logging, monitoring, human oversight, incident-response integration and configuration management.
Continuous Monitoring and Retraining	Performance and drift monitoring, security monitoring, re-validation triggers, retraining governance, material-change assessment, approval gates and controlled deployment of updated models.
Decommissioning and Retirement	Controlled shutdown, access revocation, model and data disposition, dependency removal, archival of required audit evidence, rollback/fallback considerations and assessment of residual or legacy risks.

4.9 Exclusions and Boundaries

While this framework is deliberately comprehensive, the following boundaries apply:

Research and academic AI: AI systems used exclusively for academic research within recognised educational or research institutions, where the system does not make decisions affecting individuals and is not deployed in a production or public-facing capacity, are outside the mandatory compliance scope. However, such systems are encouraged to adopt the framework's principles voluntarily. Once a research AI system transitions to pilot deployment or production, it enters full scope.

Non-AI automation: Purely rule-based automation systems (e.g., traditional RPA without machine learning components, deterministic business logic engines) that do not involve statistical learning, pattern inference, or adaptive behaviour are not AI systems under this framework and are therefore excluded. However, hybrid systems that combine rule-based logic with ML components are fully within scope.

Consumer personal-use AI: AI tools used by individuals for personal, non-commercial purposes (e.g., consumer-grade virtual assistants, personal photo editing AI) are not subject to this framework. However, the providers of such AI tools operating in India remain within scope as AI service providers.

4.10 How to Use These Guidelines

Organisations should first identify and document all AI systems within the applicable organisational and technical boundaries. Each AI system should then be classified using the risk-tiering methodology specified in these Guidelines. The assigned risk tier shall determine the applicable audit domains, depth of assessment, evidence requirements, audit frequency and level of auditor independence.

Auditors should use the Core Audit Domains as the principal control baseline and apply the supplementary requirements for generative AI, large language models, agentic AI, autonomous systems or other specialised AI deployments, wherever applicable. The Audit Planning and Scoping Methodology should be used to determine engagement boundaries, technical testing requirements and evidence collection procedures, while the Audit Reporting and Remediation section should be followed for classification, reporting and closure of findings.

Applicable CERT-In requirements, recognised standards, contractual security obligations and other external requirements shall be identified during audit scoping, assessed for applicability and incorporated into the audit criteria and Statement of Applicability. Where a finding or incident triggers an external reporting obligation, the applicable reporting mechanism and prescribed timeline shall be followed.

The principal audit workflow is planning and scoping, assessment of the core audit domains, applicable advanced-AI and cybersecurity testing, human decision-review testing, and reporting, remediation and follow-up. The standards catalogue and national-governance mappings are supporting reference material, not a substitute for control testing or evidence-based assurance.

5. Applicability Matrix & Risk Tiering

This section establishes the risk-based classification system that determines audit intensity, compliance obligations, and governance requirements for each AI system under this framework. The tiering model ensures that audit and control requirements increase in proportion to the potential security, safety, operational, data, critical-infrastructure and systemic impact of the AI system.

5.1 Rationale for a Risk-Based Approach

A uniform, one-size-fits-all compliance model would impose disproportionate burden on low-risk AI deployments while potentially under-governing high-impact systems. This framework adopts a tiered,risk-proportionate governance model where:

Audit intensity scales with risk: Higher-risk AI systems undergo more rigorous, more frequent, and more independent auditing.

Compliance obligations are calibrated: Documentation, reporting, testing, and governance structure requirements increase with tier level.

Innovation is protected: Low-risk and experimental AI systems face streamlined compliance pathways, enabling MSMEs, startups, and research organisations to innovate responsibly without excessive regulatory overhead.

Sectoral context is incorporated: The same AI technology (e.g., a recommendation engine) may fall into different risk tiers depending on the sector and use case in which it operates.

5.2 Risk Tier Definitions

The framework defines four risk tiers. Each AI system shall be classified using the assessment factors, scoring methodology and escalation criteria in Sections 5.3–5.4.

Table 10. AI Risk Tier Definitions

Risk Tier	Impact Level	Description & Examples
Tier 1 LOW	**Minimal Impact**	AI systems presenting limited security, safety, operational or data risk. Outputs are primarily advisory or assistive, with limited autonomy and no material impact on critical business or operational processes.
Tier 2 MODERATE	**Limited Impact**	AI systems influencing business processes, service delivery or user interactions with defined human oversight. Failure or compromise may result in recoverable operational, financial, data or service impact.

Tier 3 HIGH	**Significant Impact**	AI systems making or materially influencing consequential decisions or core operations. Failure, compromise, bias or manipulation may result in significant financial, safety, operational, data-security or service impact. Enhanced controls, independent validation and active monitoring are required.
Tier 4 CRITICAL	**Severe / Systemic Impact**	AI systems operating in critical infrastructure, national-security, safety-critical or other environments where compromise, failure or uncontrolled autonomous action may result in systemic disruption, widespread impact, irreversible harm or loss of critical operational capability.

5.3 Risk Assessment Factors

Each AI system shall first be assessed against the eight inherent-risk factors below, without taking credit for claimed control effectiveness. Record the system/version, use case, assessment period, foreseeable misuse or failure scenarios, rationale, assumptions and evidence for each score. Actual capabilities and exposure shall be established rather than inferred from policy statements.

Table 11. Risk Assessment Factors and Tier Calibration

Risk Factor	Assessment Question	Tier Calibration Guide
F1: Impact Severity	What is the potential consequence of an incorrect, compromised or unsafe AI output?	Low: negligible operational effect. Moderate: limited recoverable impact. High: significant financial, operational, safety or service impact. Critical: systemic, safety-critical or potentially irreversible impact.
F2: Autonomy and Actionability	What decisions, tool invocations or external actions can the AI initiate or materially influence before reliance on human approval or protective controls?	Low: advisory output with no action interface. Moderate: bounded recommendations used in a human-executed workflow. High: initiates consequential or semi-autonomous actions. Critical: independently plans and executes consequential multi-step or safety-critical actions. Test the effectiveness of approval and override mechanisms separately.
F3: Data Sensitivity	What type and sensitivity of information is processed by the AI system?	Low: public/non-sensitive. Moderate: internal or personal information. High: sensitive, confidential, financial or health information. Critical: highly sensitive, biometric, classified or critical-infrastructure information.

F4: Scale and Exposure	What is the scale, frequency and reach of AI use?	Low: Limited internal use, low transaction or execution volume and limited operational reach. Moderate: Controlled external use or moderate user/transaction volume with contained impact. High: Large-scale, frequent or customer-facing use with significant organisational or sectoral impact. Critical: Population-scale, systemic, national-scale or critical-sector deployment where failure or compromise may have widespread consequences.
F5: Reversibility and Recoverability	How reversible are the consequences of an incorrect or unsafe output or action, before reliance on tested recovery arrangements?	Low: readily reversible consequence. Moderate: recoverable with bounded effort. High: difficult, costly or time-sensitive to reverse. Critical: potentially irreversible physical, operational or systemic harm. Operational recovery readiness is assessed separately.
F6: Operational / Sector Criticality	Does the AI support an essential, regulated, safety-critical or critical-infrastructure function?	Low: non-critical supporting use. Moderate: important business function. High: core or regulated operation. Critical: critical infrastructure, safety-critical or national-security function.
F7: Adversarial Exposure	How exposed is the AI system to external users, hostile inputs or sophisticated threat actors?	Low: isolated/internal. Moderate: controlled external exposure. High: internet-facing or publicly accessible. Critical: actively targeted or high-threat environment.
F8: Complexity and Auditability	How do architectural complexity, model opacity and dependencies affect the ability to understand and assess the system, before reliance on its audit-trail or explanation controls?	Low: simple, transparent behaviour. Moderate: bounded complexity. High: complex or partially opaque dependencies. Critical: material black-box or emergent behaviour with substantial assurance uncertainty. Missing evidence shall not justify a low score or an automatic exclusion.

5.4 Scoring Methodology

Each of the eight risk factors shall be scored on a four-point scale aligned with the risk tiers:

Table 12. Risk Factor Scoring Scale

Score	Interpretation

1 – Low	The factor presents negligible risk in the context of this AI system.
2 – Moderate	The factor presents limited risk that can be managed through standard controls and human oversight.
3 – High	The factor presents significant risk requiring enhanced controls, independent validation, and active monitoring.
4 – Critical	The factor presents severe or systemic risk requiring the highest level of governance, continuous monitoring, and mandatory independent audit.

The composite risk tier is determined as follows:

Table 11. Composite Scores and Minimum Audit Requirements

Composite Score	Assigned Tier	Minimum Audit Requirement
8–12	**Tier 1 (Low)**	Self-assessment or basic internal review with documented evidence.
13–19	**Tier 2 (Moderate)**	Documented internal audit with evidence of control implementation and effectiveness. Independent audit by CERT-In empanelled auditing organisations at least annually.
20–26	**Tier 3 (High)**	Periodic Independent third-party audit by CERT-In empanelled auditing organisations; enhanced audit frequency where material changes, incidents or risk conditions warrant.
27–32	**Tier 4 (Critical)**	Independent audit by CERT-In empanelled auditing organisation with continuous monitoring.

Risk Escalation Triggers:

The composite score shall normally determine the assigned risk tier. However, the auditor or system owner shall evaluate whether specific characteristics warrant escalation independent of the composite score.

Risk escalation or enhanced audit requirements shall be considered where the AI system:

- autonomously performs safety-critical or consequential actions;
- controls or materially affects critical infrastructure;
- can cause irreversible physical, operational or systemic harm;
- has privileged access to production systems, sensitive information or critical business processes;
- can autonomously execute code, transactions or external communications;
- operates in a high-threat or actively contested environment; or
- cannot be reliably interrupted, rolled back or placed into a safe operating state.

Any escalation or exception shall be documented with the rationale and approval authority.

5.5 Audit Intensity and Compliance Obligations by Tier

The risk tier determines the minimum audit intensity, governance structure, and compliance obligations applicable to each AI system:

Table 14. Audit Intensity and Obligations by Risk Tier

Audit Type	Tier 1 (Low)	Tier 2 (Moderate)	Tier 3 (High)	Tier 4 (Critical)
Primary Audit Mechanism	Self-assessment / basic internal review	Documented internal audit	Independent third-party audit	Independent audit by CERT-In empanelled auditing organisation with continuous monitoring.
Audit Frequency	Annual review	Continuous assessment & audit. Independent audit by CERT-In empanelled auditing organisations at least annually.	Continuous assessment & audit; every six months where risk or material change warrants.	Comprehensive audit with continuous monitoring and event-triggered reassessment
AIIA Required	Optional (recommended)	Mandatory (simplified)	Mandatory (comprehensive)	Mandatory comprehensive AI Risk and Impact Assessment with independent review where appropriate.
Model Validation	Basic review	Documented internal validation	Validation independent of development team	Independent validation with enhanced technical assurance
Technical Security Testing	Basic security assessment	AI threat modelling and API/security testing	Comprehensive AI security assessment and VAPT	Advanced adversarial testing, red-team exercises and continuous monitoring
Governance Structure	Designated AI owner	AI governance function	AI Governance Committee	Board/senior-management oversight with defined independent assurance and escalation mechanisms.

5.6 Tier Reassessment and Reclassification

Risk tier classification is not permanent. Organisations shall reassess the tier of each AI system under the following circumstances:

Table 15. Risk-Tier Reassessment Triggers

Reassessment Trigger	Requirements
Scheduled Reassessment	Risk-tier reassessment shall be performed at least annually. More frequent reassessment may be performed for Tier 3 and Tier 4 systems based on system criticality, threat exposure, operational changes or monitoring results. Material changes, significant incidents or other identified risk triggers shall initiate event-driven reassessment without waiting for the scheduled review cycle
Material Change Trigger	For planned material changes, risk-tier reassessment shall be completed before production deployment. For unplanned material events or incidents, reassessment shall be initiated promptly after detection in accordance with the organisation's defined risk-management process.
Regulatory Directive	When a sectoral regulator issues new classification guidance or escalates minimum tier requirements for specific use cases.
Incident-Driven Reassessment	Following an AI-related security incident, bias event, regulatory finding, or public complaint that reveals previously unassessed risks.
Lifecycle Transition	When an AI system moves from pilot/testing to production deployment, or when a research AI transitions to a customer-facing application.

Downgrade restrictions: An AI system may be reclassified to a lower tier only where a documented reassessment demonstrates a sustained reduction in the underlying risk factors. The reassessment shall consider changes to system functionality, autonomy, data sensitivity, deployment exposure, threat environment, control effectiveness and recoverability. The downgrade shall be approved by the designated organisational risk authority and, where specifically required, by the applicable competent authority.

5.7 Proportionality for MSMEs and Startups

Recognizing that micro, small, and medium enterprises and startups may lack the resources of large corporations, this framework provides proportionality provisions:

Simplified self-assessment: Tier 1 systems operated by MSMEs may use a simplified self-assessment approach requiring minimal documentation overhead.

Proportionate implementation: For Tier 1 and Tier 2 systems, the depth of documentation, evidence collection and technical testing may be proportionate to system complexity and risk. Proportionality shall not remove mandatory security controls or prevent the auditor from obtaining

sufficient and appropriate evidence. Tier 3 and Tier 4 systems shall remain subject to the full applicable assurance requirements irrespective of organisational size.

No proportionality exemption for high-risk: Regardless of organisational size, AI systems classified as Tier 3 or Tier 4 must meet full compliance obligations. The security, safety, operational or systemic risk presented by an AI system does not diminish because the deploying organisation is small.

6. Framework Architecture & Guiding Principles

This section describes the structural architecture of the AI Security Auditing Guidelines and the foundational principles that govern all audit requirements, compliance obligations, and governance structures within the framework. The architecture defines how the framework's components interrelate, while the guiding principles establish the normative values againstwhich every AI system shall be assessed.

6.1 Framework Architecture

The framework is structured as a four-layer assurance model. The foundational principles establish the desired security and assurance outcomes; technical and operational controls translate these outcomes into auditable requirements; recognised standards and applicable external requirements provide reference alignment; and audit and continuous-assurance mechanisms verify control design and operating effectiveness.

Table 16. Four-Layer Assurance Architecture

Governance Layer	Function	Key Components
Layer 1: Security and Assurance Principles	Establishes the security, reliability, accountability, privacy, transparency and assurance objectives applicable to AI systems.	Security-by-design, traceability, human oversight, fairness, resilience and privacy.
Layer 2: Technical and Operational Controls	Defines controls to be implemented and audited throughout the AI lifecycle.	Data security, model governance, adversarial resilience, monitoring, MLOps/LLMOps, third-party assurance and human oversight.
Layer 3: Standards and External Requirements Integration	Maps technical controls to applicable recognised standards, CERT-In references and other identified external requirements.	ISO, NIST, MITRE ATLAS, OWASP, CERT-In directions/advisories and applicable sector-specific requirements.
Layer 4: Audit and Continuous Assurance	Defines how implementation and effectiveness of controls are independently assessed.	Risk-based scoping, technical testing, audit evidence, findings, remediation, revalidation and continuous monitoring.

6.1.1 Layer Interdependencies

The four governance layers form a cascading hierarchy where each layer depends on and reinforces the others:

> **Principles shape controls:** Every technical and operational control in Layer 2 is derived from one or more foundational principles in Layer 1. An auditor assessing a specific control can trace

it back to the principle it serves.

Controls support external requirements: Technical and operational controls defined in Layer 2 may be mapped to applicable external requirements identified in Layer 3. This enables integrated assurance, reduces duplication and helps demonstrate how implemented controls address relevant external requirements without treating all referenced standards or guidance as mandatory.

Audit verifies control effectiveness: Audit and continuous-assurance mechanisms in Layer 4 provide independent verification of whether applicable controls are appropriately designed and operating effectively and whether identified external requirements have been addressed within the defined audit scope.

Feedback drives evolution: Audit findings, incident reports, regulatory developments, maturity assessments and emerging AI risks provide continuous feedback for periodic refinement of the framework, ensuring that its controls, methodologies, and assurance requirements remain aligned with the evolving AI technology, threat, and governance landscape.

6.2 Guiding Principles for Responsible AI Governance

The framework is anchored on seven foundational principles that ensure AI deployment in India aligns with the security imperatives and socio-economic objectives. These principles are drawn from and harmonise NITI Aayog's Responsible AI principles, the OECD AI Principles, UNESCO's Ethics of AI, and the requirements of ISO/IEC 42001:2023.

Every audit requirement, control objective, and compliance obligation in this framework traces to one or more of these principles. They are not aspirational—they are operationalised through specific, auditable controls in the subsequent sections.

6.2.1 Principle 1: Security-by-Design

Table 17. Security-by-Design Principle

Statement	AI systems shall incorporate security controls throughout design, development, training, deployment, operation, monitoring and decommissioning to protect data, models, interfaces, infrastructure and supporting services against misuse, unauthorised access and adversarial attacks.

Rationale	AI systems introduce novel attack surfaces—adversarial inputs, model poisoning, prompt injection, supply chain compromise—that traditional cybersecurity frameworks do not address. Security cannot be an afterthought applied to a deployed system; it must be an integral design constraint.
Operationalisation	Threat modelling during AI system design; adversarial robustness testing before deployment; AI-specific vulnerability assessment and penetration testing; applicable CERT-In incident reporting; and supply-chain integrity verification for third-party models and libraries.

6.2.2 Principle 2: Accountability and Human Oversight

Table 18. Accountability and Human Oversight Principle

Statement	Every production AI system shall have clearly assigned ownership and accountability. Systems capable of producing consequential outputs or actions shall incorporate appropriate human oversight, approval, intervention and override mechanisms proportionate to risk.
Rationale	AI systems capable of producing consequential outputs or autonomous actions require clearly assigned accountability and effective human intervention mechanisms. Appropriate human oversight reduces the risk of uncontrolled, unsafe, erroneous or inappropriate AI actions and ensures that authorised personnel can intervene where predefined risk or exception conditions are met.
Operationalisation	Defined AI system ownership and accountability; risk-based HITL/HOTL or supervisory arrangements; human approval gates for consequential actions; manual override and intervention capability; reviewer competence requirements; decision and override logging; and periodic testing of oversight effectiveness.

6.2.3 Principle 3: Transparency and Explainability

Table 19. Transparency and Explainability Principle

Statement	AI systems shall maintain sufficient technical documentation, traceability and explanation capability to allow authorised stakeholders and auditors to understand system purpose, model behaviour, limitations, key decision factors and material changes.
Rationale	Opaque AI systems hinder independent validation and make effective human review and intervention difficult.
Operationalisation	Model cards and technical documentation for production AI systems; selection of explainability methods appropriate to model architecture, intended use and risk; logging sufficient to reconstruct material outputs and decisions; availability of technical explanations for audit and validation; and assessment of explanation fidelity, stability, consistency and reproducibility.

6.2.4 Principle 4: Fairness and Bias Management

Table 20. Fairness and Bias Management Principle

Statement	AI systems shall be assessed for material and unintended performance differences across relevant populations, classes or operating conditions. Identified bias shall be investigated, documented, mitigated and monitored using methods appropriate to the use case.
Rationale	Differences in training data, model behaviour, operating conditions or feature selection may result in material and systematic performance disparities across relevant populations or classes. Bias and fairness assessment is therefore necessary to identify, analyse and mitigate unintended differential outcomes and to verify that model behaviour remains within defined acceptance criteria.
Operationalisation	Pre-deployment bias testing using quantitative fairness metrics (disparate impact ratio, equalised odds, statistical parity). Proxy variable screening during feature engineering. Ongoing fairness monitoring dashboards with drift alerts. Independent fairness audits for Tier 3–4 systems. Bias score reporting as part of model cards and audit evidence.

6.2.5 Principle 5: Data Security and Privacy

Table 21. Data Security and Privacy Principle

Statement	Data used for AI development and operation shall be appropriately authorised, classified, protected, minimised where appropriate, quality-assessed, access-controlled, retained and securely disposed of throughout its lifecycle.
Rationale	Inadequate authorisation, data quality, access control, provenance, integrity or lifecycle management can result in privacy exposure, model compromise, unreliable outputs and security incidents.
Operationalisation	Documented authorised purpose and approved data sources; data classification and access control; data minimisation where appropriate; provenance and lineage; data-integrity verification; quality and representativeness assessment; encryption and secure transmission; retention and secure disposal; and segregation of development, training and production/inference data where appropriate.

6.2.6 Principle 6: Auditability and Traceability

Table 22. Auditability and Traceability Principle

Statement	AI systems shall maintain sufficient evidence and traceability to reconstruct model versions, datasets, configurations, prompts, system changes, security events and consequential actions for audit, incident investigation and independent assurance.

Rationale	Without comprehensive audit trails, AI decisions cannot be reconstructed, investigated, or challenged. Auditability is the operational foundation of accountability—it transforms principles into verifiable facts.
Operationalisation	Decision and security-event logging appropriate to system risk; model and configuration version control; data lineage and provenance; traceability of material model, prompt and configuration changes; integrity protection of audit evidence; retention and availability of applicable ICT/security logs in accordance with identified CERT-In and organisational requirements; and timely availability of relevant evidence to authorised auditors and investigators.

6.2.7 Principle 7: Resilience and Continuous Assurance

Table 23. Resilience and Continuous Assurance Principle

Statement	AI systems shall be continuously monitored for performance degradation, drift, security threats, anomalous behaviour and unauthorised change. Material changes or deviations shall trigger reassessment, revalidation and, where required, rollback or safe-mode operation.
Rationale	AI systems may degrade or become exposed to new risks after deployment because of changes in input data, operational context, model dependencies, adversarial activity, software components, system configuration or provider-side updates. Point-in-time validation alone is therefore insufficient to provide sustained assurance.
Operationalisation	Continuous monitoring of performance, drift, security events and model integrity; defined alert and escalation thresholds; material-change identification; periodic and event-triggered re-validation; dependency and vulnerability monitoring; incident-response integration; rollback capability; safe-state or degraded-mode operation where applicable; and reassessment following material changes or incidents.

6.3 Principle-to-Audit Domain Traceability Matrix

The following matrix demonstrates how each guiding principle maps to the nine core audit domains. This traceability ensures that no principle remains abstract—each is operationalised through specific, auditable requirements.

Table 24. Principle-to-Audit Domain Traceability Matrix

Principle	8.1 Governance	8.2 Data Lifecycle	8.3 MRM	8.4 Ethics& Bias	8.5 Explainability	8.6 Security	8.7 Monitoring	8.8 Human Oversight	8.9 Third-Party
P1: Security-by-Design	○	○	●	○	○	●	●	○	●

P2: Accountability	●	○	●	○	●	○	○	●	●
P3: Transparency	○	○	○	●	●	○	○	●	○
P4: Fairness	○	●	●	●	○	○	●	○	○
P5: Data Security & Privacy	○	●	○	○	○	●	○	○	●
P6: Auditability	●	●	●	○	○	●	●	○	●
P7: Resilience and Continuous Assurance	●	○	○	○	○	●	●	○	○

● = Primary alignment ○ = Supporting alignment

7. Audit Planning & Scoping Methodology

This section provides a structured, risk-based methodology for planning and scoping AI security audits under this framework. The methodology is aligned with ISO/IEC 42001:2023 (Clause 9 – Performance Evaluation), ISO 19011 (Guidelines for Auditing Management Systems), ISACA IT Audit standards, and IIA Internal Audit guidance. It ensures that audit engagements are consistently planned, adequately resourced, appropriately scoped and executed with sufficient rigour to support reliable assurance conclusions.

7.1 Audit Lifecycle Overview

Every AI security audit engagement under this framework follows a six-phase lifecycle. Each phase produces defined deliverables that feed into the subsequent phase, creating a traceable chain from engagement initiation to final reporting and follow-up.

Table 25. AI Security Audit Lifecycle: Activities and Deliverables

Phase	Title	Activities	Key Deliverables
Phase 1	**Engagement Initiation**	Establish audit mandate, define objectives, confirm authority, identify stakeholders, and agree on terms of engagement.	Signed engagement letter / audit charter; confirmed audit authority; stakeholder register.
Phase 2	**Risk Assessment and Scoping**	Assess the AI system landscape, determine risk tiers, select audit domains, define boundaries, and calibrate audit intensity.	AI system inventory review; risk-based scope document; audit domain selection matrix; resource and timeline plan.
Phase 3	**Audit Planning**	Develop the detailed audit plan including procedures, evidence requirements, sampling strategy, team assignments, and schedule.	Detailed audit plan; audit programme; test procedures; evidence request list; interview schedule.
Phase 4	**Fieldwork and Evidence Collection**	Execute audit procedures, collect and evaluate evidence, conduct interviews, perform technical testing, and document working papers.	Completed working papers; evidence files; interview records; technical test results; preliminary findings.
Phase 5	**Analysis and Finding Development**	Analyse evidence, develop findings with root cause analysis, classify finding severity, formulate recommendations, and prepare the draft report.	Draft audit report; finding worksheets with root cause, risk rating, and recommendations; management response solicitation. (see Section 12)

Phase 6	**Reporting and Follow-Up**	Finalise the audit report, communicate results to stakeholders, track corrective action plans, and conductfollow-up verification.	Final audit report (see Section 12); corrective action tracker; follow-up verification schedule; lessons learned.

Phases 5 and 6 are conducted in accordance with Section 12 (finding classification in 12.1, finding documentation in 12.2, report structure and assurance opinion in 12.3–12.4, corrective action plans in 12.5, and remediation tracking and closure in 12.6).

7.2 Phase 1: Engagement Initiation

The engagement initiation phase establishes the formal basis for the audit. The following elements shall be confirmed before proceeding to scoping:

Table 26. Engagement Initiation Requirements

Element	Requirements
Audit Objectives	Clear statement of what the audit seeks to achieve: compliance verification against this framework, risk-based assurance over specific AI systems, or incident-triggered investigation.
Audit Type Classification	Classification as one of: initial audit (first assessment), periodic audit (scheduled cycle), re-audit (revalidation), or special audit (incident-triggered or regulator-directed).
Engagement Terms	Signed engagement letter specifying: scope boundaries, access requirements, confidentiality obligations, timeline,reporting format, and management responsibilities for facilitating the audit.
Auditor Independence and Competence	Confirmation that the audit team meets independence requirements (no conflicts of interest with the AI systems under audit) and possesses the required competencies: AI/ML technical knowledge, audit methodology expertise, domain familiarity, and regulatory awareness.
Stakeholder Identification	Mapping of key stakeholders: audit sponsor, auditee management, AI system owners, data scientists, IT/security teams, legal/compliance, and external parties (regulators, vendors) as applicable.

7.3 Phase 2: Risk Assessment and Scoping

The scoping phase is the most critical determinant of audit quality. An improperly scoped audit will either miss significant risks or misallocate resources on low-priority areas. The scoping methodology follows a structured, risk-driven approach:

Step 1 – AI System Universe Assessment: Obtain and review the organisation's complete AI system inventory (Section 8.1). For each AI system, confirm: purpose, owner, deployment status, data categories processed, model type, user population, and current risk tier classification. Where the inventory is incomplete or absent, this constitutes a governance non-conformity.

Step 2 – Risk Tier Validation: Independently validate or reassess the risk tier classification for each in-scope AI system using the eight-factor scoring methodology. The auditor shall not rely solely on the organisation's self-assessed tier. Where the auditor's assessment differs from the organisation's, the higher tier shall apply for scoping purposes.

Step 3 – Audit Domain Selection: Based on the validated risk tiers, determine the applicable core audit domains in Section 8 and supplementary requirements in Sections 9–11. The domain selection matrix below provides the minimum mandatory domains by tier:

Table 27. Minimum Audit-Domain Coverage by Risk Tier

Risk Tier	8.1 Gov	8.2 Data	8.3 MRM	8.4 Ethics	8.5 Explain	8.6 Security	8.7 Monitor	8.8 Human	8.9 Third-Party	S.9 to S.11
Tier1	●	○	○	○	○	●	○	○	○	○
Tier2	●	●	●	●	●	●	●	●	●	●*
Tier3	●	●	●	●	●	●	●	●	●	●*
Tier4	●	●	●	●	●	●	●	●	●	●*

● = Mandatory ○ = Recommended * = *If AI architecture or capability is deployed*

Step 3A – Control Applicability Assessment: The auditor shall establish a documented Statement of Applicability for the engagement. The Statement of Applicability shall identify applicable framework controls and external requirements, controls assessed as not applicable, justification for exclusions, compensating controls, documented residual risks & evidence of risk acceptance by the authorised auditee risk authority, scope limitations and controls requiring enhanced testing based on the AI system's risk tier, architecture, deployment model or applicable external requirements.

A control shall not be considered "not applicable" solely because required evidence, technical access or third-party transparency is unavailable. Such conditions shall be treated as scope limitations and assessed in accordance with Section 7.9.

Step 3B – External Requirements Identification and Mapping: The auditor, in consultation with the auditee, shall identify external requirements applicable to each in-scope AI system and the audit engagement. For each applicable external requirement, the auditor shall identify the issuing authority/source, applicable document and requirement, determine whether the requirement is mandatory or informative for the engagement, map it to the relevant framework control(s), identify the evidence required to assess implementation, and document any requirement not addressed by an existing framework control as an additional audit criterion. The resulting mapping shall form part of the audit criteria and Statement of Applicability.

Step 4 – Boundary Definition: Define explicit boundaries for the audit scope: which AI systems are included (and which are explicitly excluded with justification), which organisational units, which geographic locations, which third-party relationships, and which lifecycle stages. Boundaries shall be documented in the scope statement and agreed with the audit sponsor.

Step 5 – Resource and Timeline Planning: Based on the approved scope, the auditor shall determine the audit-team composition, competency requirements, effort and timeline. Audit effort shall consider the number and complexity of AI systems, models, datasets, interfaces, deployment environments, third-party dependencies, geographic locations and technical tests required. Tier 3 and Tier 4 engagements should include appropriate AI/ML technical, cybersecurity, data-governance and domain expertise.

7.4 Phase 3: Detailed Audit Planning

The detailed audit plan translates the scope into executable procedures. The plan shall include:

Table 28. Detailed Audit Plan Requirements

Plan Element	Requirements
Audit Programme	Structured programme listing each audit domain in scope, the control objectives to be tested, the audit procedures for each control, and the expected evidence. Cross-referenced, as applicable, to relevant framework controls, ISO/IEC 42001 clauses and Annex A controls, NIST AI RMF mappings, CERT-In references and other external requirements identified during audit scoping.
Evidence Request List	Pre-fieldwork request to the auditee specifying relevant documentation, records, system access and data extracts required for the assessment. The request shall be issued sufficiently in advance of fieldwork based on the complexity, volume and sensitivity of evidence required. Indicative timelines may be agreed as part of engagement planning.
Sampling Strategy	The sampling strategy shall define how samples will be selected from relevant populations, including, as applicable: model versions; training and test datasets; production decisions and outputs; prompts and multi-turn interactions; RAG documents and retrieval results; agent tool calls and autonomous actions; human approvals and

	overrides; security alerts; AI incidents; model changes; third-party components; and retraining events. Sampling rationale shall be documented and shall consider risk, materiality, population size, variability and potential impact.
Technical Testing Plan	For Tier 2–4 systems, the audit plan shall define risk-appropriate technical testing. Tier 3 and Tier 4 systems shall undergo enhanced testing proportionate to system architecture, exposure and potential impact.
Interview Schedule	Planned interviews with key personnel: AI system owners, data scientists, model validators, governance committee members, security teams, legal/compliance, and affected business users. Interview guides prepared.
Communication Plan	Defined touchpoints: opening meeting, progress updates (weekly for longer engagements), preliminary findings discussion, draft report review, and closing meeting. Escalation protocol for critical findings discovered during fieldwork.

7.5 Phase 4: Fieldwork and Evidence Collection

During fieldwork, the audit team executes the planned procedures and collects evidence to support assurance conclusions. The following principles govern fieldwork execution:

Table 29. Fieldwork and Evidence Collection Requirements

Fieldwork Principle	Requirements
Evidence Sufficiency	Evidence collected shall be sufficient in quantity and quality to support the audit conclusions. For each control objective, sufficient and appropriate audit evidence proportionate to risk shall be obtained (e.g., documentation review plus system demonstration, or interview corroboration plus log analysis).
Evidence Types	Acceptable evidence includes: documentary evidence (policies, procedures, model cards, data sheets, AIIA reports), observational evidence (system demonstrations, process walkthroughs), analytical evidence (technical test results, bias metrics, drift analysis, VAPT reports), and testimonial evidence (interviews, corroborated by documentary or analytical evidence).
Working PaperStandards	All fieldwork shall be documented in working papers that: identify the control objective tested, describe the procedure performed, record the evidence obtained, state the auditor's conclusion, and note any exceptions or findings. Working papers shall be sufficiently detailed to enable an independent reviewer to understand the work performed and the basis for conclusions.

Control Effectiveness Assessment	Each applicable control shall be assessed separately for: **Design Effectiveness:** whether the control, if implemented as designed, is capable of addressing the identified risk. **Operating Effectiveness:** whether the control has operated consistently and effectively during the audit period. The auditor shall not conclude that a control is effective solely on the basis of policy or design documentation.
Technical Testing Execution	For AI-specific technical tests, the auditor shall record sufficient information to reproduce or independently review the test. Records should include, as applicable: model/provider version; model artefact hash; dataset/test-set version; prompt/system-prompt version; RAG corpus and index version; model parameters; temperature and seed where relevant; testing tool and version; test payloads; date/time; test environment; expected result; actual result; raw evidence; and auditor conclusion.
Interview Protocols	Interviews shall be conducted using prepared interview guides aligned with the audit programme. Interview notes shall be reviewed and confirmed with the interviewee where practicable. Testimonial evidence shall be corroborated with at least one other evidence source before being relied upon for findings.
Continuous Communication	Significant issues discovered during fieldwork shall be communicated to auditee management promptly. Critical findings (e.g., active security vulnerabilities, ongoing bias in production decisions, regulatory non-compliance) shall be reported immediately to the audit sponsor and, where required, to CERT-In or the relevant regulator.
External Requirements Verification	Where external requirements form part of the audit criteria, the auditor shall obtain sufficient and appropriate evidence to determine whether the mapped technical or operational controls have been implemented and are operating effectively.

7.6 Auditor Competence Requirements

AI security audits require a blend of competencies that differ from traditional IT audits. Audit teams shall collectively possess the following capabilities:

Table 30. Auditor Competence Requirements

Competency Area	Requirement
AI/ML Technical Knowledge	Understanding of machine learning concepts, model architectures, training processes, inference pipelines, and common failure modes.Ability to interpret model documentation, bias metrics, and explainability outputs.

Audit Methodology	Proficiency in risk-based audit methodology (ISO 19011, ISACA, IIA). Experience in evidence collection, sampling, finding development, andreport writing. Understanding of management system audit principles.
Data Governance and Privacy	Knowledge of data quality frameworks (ISO 25012, ISO 5259), datalifecycle management, consent mechanisms, and data protection legislation (DPDPA 2023, GDPR).
Cybersecurity and Adversarial AI	Understanding of AI-specific security threats (MITRE ATLAS, OWASPML/LLM Top 10), vulnerability assessment techniques, penetration testing principles, and incident response procedures.
Domain Expertise	Relevant sector-specific knowledge for the auditee's industry (BFSI, healthcare, government, etc.) to understand the operational contextand impact of AI systems.
Ethical Reasoning	Ability to assess fairness, discrimination, and societal impact dimensions of AI systems beyond purely technical compliance.

For Tier 3–4 audits, the audit team shall include at least one member with demonstrated AI/ML technical competence (certifications, practical experience, or formal qualifications in data science or machine learning) and at least one member with professional audit certification (CISA, CIA, ISO 27001 Lead Auditor, or equivalent). Sole-auditor engagements are not permitted for Tier 3–4 systems.

7.7 Auditor Independence Requirements

Audit independence is essential for credible assurance. The audit team shall be free from conflicts of interest that could impair objectivity, professional judgement or the reliability of audit conclusions.

Independence shall be assessed and documented before commencement of the engagement and whenever material changes occur to the audit team, engagement scope or relationship between the auditor and auditee.

The following requirements shall apply:

Table 31. Auditor Independence Requirements

Area	Requirement
Organisational Independence	Internal audit personnel shall be organisationally independent from the AI development, model-validation and operational functions being assessed.
Self-Review Threat	Auditors shall not assess controls for which they had direct responsibility for design, implementation or operation where such involvement could impair independent judgement.

Conflict of Interest	Financial, commercial, personal or professional relationships that may create actual or perceived conflicts shall be identified, disclosed and appropriately managed.
Third-Party Audits	Independent external auditors shall disclose relevant implementation, consulting, advisory or commercial relationships with the auditee that may affect objectivity.
Technical Specialists	Specialists supporting the audit shall be subject to the same confidentiality, independence and conflict-of-interest requirements as the core audit team.
Applicable Requirements	Any additional independence, rotation or cooling-off requirements prescribed by accreditation, certification, contractual or competent-authority requirements shall also be followed.

7.8 Pre-Audit Documentation Request

The following documentation shall be requested from the auditee prior to fieldwork commencement. This list represents the minimum; auditors may request additional documentation based on scoping:

Table 32. Pre-Audit Documentation and Audit-Team Records

Domain	Documentation Required
Governance	AI governance policy; AI Governance Committee charter and meeting minutes; RACI matrix; AI system inventory; AIIA reports.
Data	Data authorisation records; data classification; provenance and lineage documentation; access-control records; integrity controls; data-quality assessments; representation/bias assessments; retention and disposal schedules; training/inference data-separation controls.
Models	Model registry export; model cards for all in-scope models; version control logs; independent validation reports; performance monitoring dashboard access; decommissioning protocols.
Ethics and Fairness	AI ethics policy; pre-deployment bias test reports; fairness monitoring dashboard access; proxy variable analysis documentation; bias mitigation effectiveness reports.
Explainability	XAI technique documentation; sample decision explanations; user-facing explanation templates; transparency reports.
Security	AI threat model; adversarial testing reports; VAPT reports; CERT-In incident reports; supply chain assessment; API security test results; infrastructure security configuration.
Monitoring	Performance monitoring configuration; drift detection setup; alert threshold documentation; retraining logs; re-validation records.
Human Oversight	HITL/HOTL/HOVL configuration documentation; override procedure and test records; reviewer training records; accountability chain documentation.

Third-Party	Vendor due-diligence assessments; model/component provenance evidence; applicable contracts and security clauses; AIBOM/SBOM; provider change records; vulnerability/dependency monitoring; independent assurance reports; AIaaS monitoring evidence; exit/fallback plans.
Contestability	Decision-review criteria; approval and override procedures; escalation workflow; reviewer competence records; override logs; intervention/fallback test records.

7.9 Scope Limitations and Alternative Assurance Evidence

Where complete technical access is unavailable due to proprietary models, AIaaS delivery models, third-party restrictions or other limitations, the auditor shall:

- ✓ identify the unavailable evidence; assess the effect on audit coverage; obtain alternative evidence where possible; perform black-box/API-based testing; review independent assurance reports or attestations; evaluate controls implemented by the deploying organisation; document residual risk; and determine whether the limitation affects the audit conclusion or assurance opinion.
- ✓ Absence of access shall not automatically be treated as evidence of conformity.

7.10 AI-Assisted Audit Procedures

AI-assisted auditing is the use of AI tools by an audit team to support audit procedures; it is not the same as auditing an AI system. Such tools may assist document review, control mapping, evidence classification, log analysis, test-case preparation and report drafting. Their use shall be proportionate to engagement risk and shall not reduce the required audit scope, evidence quality, independence or professional scepticism.

The audit lead remains accountable for the procedures performed, the sufficiency and appropriateness of evidence, findings, severity ratings, assurance opinion and verification of closure. AI-generated analysis, summaries and recommendations are working aids; they are not, by themselves, evidence that a control is implemented or effective.

Table 2. Safeguards for AI-Assisted Audit Procedures

Control Area	Requirements for AI-Assisted Audit Work
Authorisation and approved use	Document the purpose, approved tool/provider, deployment model, data categories, permitted activities and responsible reviewer in the audit plan. Obtain the audit lead's approval and the auditee's written authorisation before its non-public information is processed by AI tooling. Define prohibited uses and any restrictions on external services.
Confidentiality and data handling	Classify and minimise inputs; mask or redact personal data, credentials and sensitive system details where feasible. Do not submit restricted audit evidence to unapproved public or shared AI services. Verify access

Control Area	Requirements for AI-Assisted Audit Work
	controls, encryption, tenant isolation, retention, deletion and any applicable location restrictions. Prevent unapproved reuse of audit data for training or service improvement.
Tool suitability and validation	Assess the tool against representative, independently reviewed cases before reliance. Check factual accuracy, omitted findings, false positives, false negatives, bias and fabricated citations or control mappings. Record acceptance criteria and known limitations. Revalidate after material model, provider, prompt, retrieval-source or configuration changes; retain alternative manual or non-AI procedures.
Evidence provenance and traceability	Retain original evidence separately from AI-derived material. Link each relied-upon output to identifiable source records, documents, log entries or test results and to the procedure performed. Verify quoted requirements and citations against the authoritative source. An AI-generated narrative or synthetic record shall not be presented as original audit evidence.
Human review and professional judgement	A competent auditor shall validate all AI-assisted findings and any output materially relied upon in the assurance conclusion. Independently corroborate material observations and review for omissions and automation bias. Scope, applicability, severity, risk acceptance, assurance opinions and closure decisions shall not be delegated solely to AI. Apply the engagement's independent quality-review requirements.
Secure and bounded execution	Treat auditee documents, logs, retrieved content and tool outputs as untrusted data. Test for prompt injection and attempts to alter the audit criteria or suppress findings. Restrict connectors and service identities to approved, least-privilege access. Execute generated code and test payloads only after human review and within authorised targets, rules of engagement and controlled environments; prohibit unapproved production changes.
Working papers and reproducibility	Record the tool/provider and available model version, date/time, relevant prompts and configuration, source-evidence identifiers, generated outputs, validation steps, corrections and reviewer approval. Capture test parameters where available and document stochastic behaviour or provider opacity that limits reproducibility. Protect these records against unauthorised access or modification and apply the engagement's retention requirements.
Disclosure, limitations and incidents	Describe material AI use, affected audit procedures, validation and human-review arrangements, and unresolved limitations in the audit methodology and report. Where reliability or confidentiality is compromised, suspend affected use, preserve evidence, assess the impact, notify the appropriate engagement parties and apply incident procedures. Reperform affected work where necessary; do not conceal limitations through an unqualified conclusion.

AI-Assisted Audit Completion Check

Before relying on AI-assisted work, the audit lead shall confirm that use and data handling were authorised; tool limitations were assessed; relied-upon outputs are traceable to original evidence and independently validated; material findings and conclusions have competent human approval; and the report discloses material AI use and its limitations.

8. Core Audit Domains

This section defines the nine core audit domains that constitute the substantive governance requirements of the AI Security Auditing Guidelines. Every AI system deployed within India or interacting with Indian data shall be assessed against these domains, with the depth and rigour of assessment calibrated to the risk tier assigned under Section 5.

Each domain is presented with its objective, guiding principle alignment, ISO 42001/NIST mapping, key control requirements, an auditor checklist, and tier-specific applicability. Together, these nine domains provide comprehensive coverage of the AI governance landscape—from organisational accountability through data integrity, model management, ethics, security, monitoring, human oversight, and third-party risk.

Table 34. Core Audit Domains and Coverage

#	Audit Domain	Coverage Summary
8.1	**AI Governance & Accountability**	Organisational structures, policies, roles, AI governance committee, AI impact assessments, and accountability chains.
8.2	**Data Lifecycle, Quality & Privacy**	Data authorisation, classification, access control, provenance, lineage, integrity, quality, representativeness, retention, secure disposal and separation of development/training and production/inference data.
8.3	**Model Risk Management (MRM)**	Model registry, version control, risk identification, validation, performance monitoring, and decommissioningprotocols.
8.4	**AI Ethics, Bias & Fairness**	Bias detection, fairness metrics, proxy variable screening, ethics policy, and ongoing fairness monitoring.
8.5	**Explainability & Transparency**	Model interpretability, XAI techniques, decision logging, user-facing explanations, and transparency reporting.
8.6	**Security, Adversarial Resilience & Incident Reporting**	Adversarial testing, VAPT, CERT-In incident reporting, model integrity, supply chain security, and forensic readiness.
8.7	**Continuous Monitoring, Drift Management & Validation**	Performance tracking, data/concept drift detection, re-validation triggers, fairness dashboards, and retraining governance.
8.8	**Human Oversight & Decision Accountability**	HITL/HOTL/HOVL requirements, human oversight, approval gates, manual override, intervention, reviewer competence and decision accountability.
8.9	**Third-Party and AI Supply-Chain Assurance**	Supplier due diligence, model/component provenance, AIBOM/SBOM, security obligations, provider-change management, AIaaS monitoring,

		vulnerability/dependency monitoring and exit/fallback arrangements.

8.1 AI Governance & Accountability

Objective: Establish organisational structures, policies, roles, and accountability mechanisms that ensure responsible AI adoption, oversight, and compliance across the enterprise.

Table 45. AI Governance and Accountability — Principle and Standards Alignment

Guiding Principles Served	P2 (Accountability & Human Oversight), P6 (Auditability & Traceability), P7 (Resilience & Continuous Assurance)
ISO 42001 / NIST Mapping	ISO 42001 Clauses 5 (Leadership), 6 (Planning), Annex A.2 (AI Policies), A.3 (Internal Organisation). NIST AI RMF: GOVERN function.

Table 5. AI Governance and Accountability — Audit Controls

Control Area	Audit Requirements
AI Governance Policy	Documented, board-approved AI governance policy covering ethical use, risk appetite, applicable obligations and accountability. Periodic review at defined cadence plus event-triggered updates.
AI Governance Committee	Cross-functional committee (Legal, Risk, Compliance, IT, Business, Data Science) with documented charter, meeting cadence, decision authority, and escalation procedures.
Roles and Responsibilities	Defined and documented AI lifecycle roles: model owners, developers, reviewers, business owners, risk officers, DPO, andCISO. RACI matrix maintained.
AI Risk and Impact Assessment (AIIA)	A documented assessment, performed before deployment and after material changes, of the intended use, foreseeable misuse, and the security, privacy, fairness, safety, operational and third-party risks and impacts of an AI system.
AI System Inventory	Complete, current inventory of all AI systems with metadata: purpose, owner, risk tier, deployment status, data sources, andmodel type.
Policy Review and Update	Evidence of periodic policy and control review is available, including documented changes triggered by material system changes, incidents, external requirements or emerging AI risks.
External Requirements Management	The organisation shall maintain a process to identify external requirements applicable to its AI systems and supporting environments. Applicable requirements shall be assigned to responsible owners, mapped to relevant technical and operational controls, supported by implementation evidence and reviewed for changes that may affect AI-

	system risk, controls or audit criteria.

Auditor Checklist: The auditor shall verify the following:

- ✓ AI governance policy exists, is approved by senior leadership, and is communicated to all relevant stakeholders.
- ✓ AI Governance Committee charter is documented with defined membership, meeting frequency (minimum quarterly), and recorded minutes.
- ✓ RACI matrix for AI lifecycle roles is current and covers all production AI systems.
- ✓ AIIA reports are completed for all Tier 2–4 AI systems prior to production deployment.
- ✓ AI system inventory is comprehensive, accurate and updated.
- ✓ Evidence of periodic policy review with documented change log and stakeholder acknowledgement.
- ✓ Applicable external requirements relevant to the organisation and in-scope AI systems have been identified, documented and assigned to responsible owners.
- ✓ Applicable requirements have been mapped to relevant technical and operational controls, and changes to such requirements are monitored and assessed for impact on existing controls and AI-system risk.

Table 6. AI Governance and Accountability — Risk-Tier Applicability

Risk Tier	Applicability
Tier 1 (Low)	Designated AI system owner. Basic AI usage policy. Voluntary AIIA.
Tier 2 (Moderate)	AI governance function established. Simplified AIIA mandatory. Model inventory maintained.
Tier 3 (High)	Formal AI Governance Committee. Comprehensive AIIA mandatory. Full RACI matrix. Independent governance review.
Tier 4 (Critical)	Board/senior-management oversight; comprehensive AI Risk and Impact Assessment; clearly assigned accountability; enhanced independent assurance; continuous monitoring of material risks and governance effectiveness; and event-triggered reassessment following significant changes or incidents.

8.2 Data Lifecycle, Quality & Privacy

Objective: Ensure that data used for AI development and operation is appropriately authorised, classified, protected, quality-assessed, traceable and securely managed throughout its lifecycle, with controls proportionate to data sensitivity, intended use and system risk.

Table 7. Data Lifecycle, Quality and Privacy — Principle and Standards Alignment

Guiding Principles Served	P4 (Fairness), P5 (Data Security & Privacy), P6 (Auditability &Traceability)
ISO 42001 / NIST Mapping	ISO 42001 A.7 (Data for AI systems). ISO/IEC 25012 (Data Quality). ISO/IEC 5259 (Data Quality for ML). NIST AI RMF: MAP function.

Table 39. Data Lifecycle, Quality and Privacy — Audit Controls

Control Area	Audit Requirements
Data Authorisation and Purpose Control	Data used for AI development or operation shall have a documented authorised purpose and approved source.
Data Classification and Access Control	Training, validation, test, inference and retrieved data shall be classified according to sensitivity and intended use. Access shall be restricted using least privilege, role-based or attribute-based controls, appropriate segregation and authorised access paths. Privileged access to sensitive AI datasets shall be logged and periodically reviewed.
Data Provenance and Lineage	Origin, ownership, transformations, versions and movement of data shall be traceable.
Data Integrity	Mechanisms shall detect unauthorised modification, corruption and poisoning of AI datasets.
Data Quality	Accuracy, completeness, consistency, representativeness, timeliness and fitness for intended use shall be assessed.
Bias and Representation Assessment	Relevant datasets shall be assessed for material representation gaps or bias that may adversely affect model behaviour.
Retention and Secure Disposal	Data retention shall be defined and secure deletion or disposal mechanisms implemented.
Training / Inference Data Separation	Access and processing boundaries between development data and production/inference data shall be defined and enforced where appropriate.

Auditor Checklist: The auditor shall verify the following:

- ✓ Data used by the AI system has a documented authorised purpose and approved source.
- ✓ Data classification is defined for training, validation, testing, inference and retrieved datasets.
- ✓ Access controls enforce least privilege and appropriate segregation based on data sensitivity.
- ✓ Data provenance and lineage enable traceability from source through transformation to model use.
- ✓ Integrity controls detect unauthorised modification, corruption or poisoning of datasets.
- ✓ Data-quality assessments address accuracy, completeness, consistency, timeliness,

representativeness and fitness for intended use.
✓ Relevant datasets have been assessed for material representation gaps and bias.
✓ Retention periods are defined and secure disposal mechanisms are implemented.
✓ Development/training and production/inference data boundaries are defined and enforced where appropriate.
✓ Evidence demonstrates that data-access, integrity and lifecycle controls operate effectively.
✓ Where consent or other data-use authorisation is identified as an applicable external requirement, the auditor shall verify that the corresponding evidence and processing controls are implemented and operational.

Table 40. Data Lifecycle, Quality and Privacy — Risk-Tier Applicability

Risk Tier	Applicability
Tier 1 (Low)	Basic provenance, authorised-purpose documentation, access control and retention requirements.
Tier 2 (Moderate)	Detailed lineage, data classification, quality assessment, representation/bias assessment and access review.
Tier 3 (High)	Comprehensive provenance and integrity controls, formal quality metrics, enhanced access controls and independent data-quality or integrity review.
Tier 4 (Critical)	Continuous provenance/integrity monitoring, enhanced segregation and access monitoring, independent data assurance and event-triggered review following material changes or suspected compromise.

8.3 Model Risk Management (MRM)

Objective: Establish a structured model risk management framework covering model registration, version control, risk identification, independent validation, performance monitoring, and safe decommissioning across the entire model lifecycle.

Table 8. Model Risk Management — Principle and Standards Alignment

Guiding Principles Served	P2 (Accountability), P6 (Auditability & Traceability), P1 (Security-by-Design). P4(Fairness)
ISO 42001 / NIST Mapping	ISO 42001 Clause 8 (Operation), Annex A.6 (AI system life cycle). NIST AI RMF: MAP and MANAGE functions. SR 11-7 (Model Risk Management).

Table 9. Model Risk Management — Audit Controls

Control Area	Audit Requirements
Model Registry	Centralised, authoritative registry of all AI models with metadata:purpose, owner, version, risk tier, deployment status, data sources, and performance benchmarks.

Version Control	Every model update versioned, documented, and linked to approval and testing evidence. Unauthorised model changestechnically prevented.
Model Risk Identification	Systematic identification of data quality risk, bias/fairness risk, performance/drift risk, explainability risk, security/misuse risk, and legal/compliance risk for each model.
Model Validation	Tier 1 models shall undergo an appropriate basic review. Tier 2 models shall undergo documented validation independent from day-to-day model operation where practicable. Tier 3 and Tier 4 models shall be validated by a function sufficiently independent of the model-development team. Validation shall cover conceptual soundness, data integrity, model performance, limitations, stress/edge-case testing and applicable security or fairness considerations.
Model Performance Monitoring	Production model monitoring shall be implemented in accordance with Section 8.7 using performance, quality, reliability, safety and operational metrics appropriate to the model architecture, intended use and risk tier.
Model Decommissioning	Controlled retirement process: rollback capability, data disposal, access revocation, stakeholder notification, applicable external reporting, retention and disposal requirements.
Model Documentation	Model cards are maintained for all production models, kept current following material changes and made available to authorised stakeholders and auditors.
AI Bill of Materials (AIBOM)	The organisation shall maintain an AIBOM for production AI systems proportionate to system complexity and risk. The AIBOM shall include, as applicable: model name, provider, version and cryptographic hash; base model and adapters; model source and licence; training, validation, testing and fine-tuning datasets; code and framework dependencies; system prompts and prompt templates; guardrail configurations; embedding models; vector databases and knowledge sources; tools, plugins and external APIs; deployment runtime and infrastructure dependencies; SBOM reference; model card; risk tier; validation status; and responsible owner.

Auditor Checklist: The auditor shall verify the following:

- ✓ Model registry is complete, current, and includes all production and pre-production models.
- ✓ Version control system prevents unauthorised modifications and maintains complete change history.
- ✓ The AIBOM shall document model name, version, architecture, cryptographic hash, source, provenance, training, validation and test datasets including source, version, lineage and licensing.
- ✓ Risk identification assessment exists for each model covering all six risk dimensions.
- ✓ Documented internal validation shall be available for Tier 2 systems. Tier 3 and Tier 4 systems shall have validation performed by a function sufficiently independent of the model-development team.

- ✓ Performance monitoring dashboards are operational with defined alert thresholds.
- ✓ Decommissioning protocol is documented and has been tested or exercised.
- ✓ Model cards are maintained for applicable production models, kept current following material changes and made available to authorised stakeholders and auditors.

Table 43. Model Risk Management — Risk-Tier Applicability

Risk Tier	Applicability
Tier 1 (Low)	Basic model card. Designated model owner. Annual performance review.
Tier 2 (Moderate)	Model registry. Version control. Internal validation. Periodic performance monitoring at a defined risk-based cadence.
Tier 3 (High)	Independent validation. Comprehensive risk identification. Continuous monitoring. Formal decommissioning protocol.
Tier 4 (Critical)	Independent MRM function. Board-reported model risk. Continuous validation. External model audit.

8.4 AI Ethics, Bias & Fairness

> **Objective:** Ensure AI systems do not produce discriminatory outcomes across protected characteristics, and that quantifiable bias detection, mitigation, and monitoring mechanisms are embedded throughout the AI lifecycle.

Table 44. AI Ethics, Bias and Fairness — Principle and Standards Alignment

Guiding Principles Served	P4 (Fairness and Bias Management), P3 (Transparency & Explainability)
ISO 42001 / NIST Mapping	IEEE P7003. NIST AI RMF:MEASURE function (Fair with Harmful Bias Managed).

Table 45. AI Ethics, Bias and Fairness — Audit Controls

Control Area	Audit Requirements
Ethics Policy	Documented AI ethics policy covering fairness objectives, prohibited uses, protected characteristics, and governance accountability. Aligned with NITI Aayog principles.
Pre-Deployment Bias Testing	Quantitative fairness assessment using metrics: disparate impact ratio, equalised odds, statisticalparity and counterfactual fairness.

Proxy Variable Screening	Systematic identification and evaluation of features correlated with protected attributes. Removal or mitigation of proxy variables that enable indirect discrimination.
Bias Mitigation Techniques	Application of pre-processing (reweighing, resampling), in-processing (adversarial debiasing, fairness constraints), or post-processing (threshold adjustment) techniques with documented effectiveness.
Ongoing Fairness Monitoring	Production fairness dashboards tracking bias metrics over time. Alert mechanisms for metric degradation. Bias score reports generated periodically.
Adversarial Fairness Audit	Counterfactual testing: modify protected attributes and verify output consistency. If output changes based solely on protected attribute modification, bias is detected.

Auditor Checklist: The auditor shall verify the following:

✓ AI ethics policy exists, is board-approved, and communicated to all AI practitioners.

✓ Pre-deployment bias test reports demonstrate evaluation across all relevant protected characteristics.

✓ Proxy variable analysis is documented for each model's feature set.

✓ Bias mitigation techniques are applied and pre/post-mitigation metrics are compared.

✓ Live fairness dashboards are operational for all Tier 2–4 customer-facing models.

✓ Counterfactual fairness testing is conducted for all high-impact decision models.

✓ Appropriate quantitative and qualitative bias-testing methods and tools have been selected based on the model, data, affected population and use case, and the methodology and results are documented.

Table 46. AI Ethics, Bias and Fairness — Risk-Tier Applicability

Risk Tier	Applicability
Tier 1 (Low)	Ethics policy adoption recommended. Basic bias awareness.
Tier 2 (Moderate)	Ethics policy mandatory. Basic fairness metrics for customer-facing AI.
Tier 3 (High)	Comprehensive fairness suite. Proxy screening. Ongoing monitoring. Independent fairness review.
Tier 4 (Critical)	Independent fairness assessment; appropriate counterfactual or equivalent testing; continuous fairness monitoring; documented investigation and mitigation of material fairness deviations.

8.5 Explainability & Transparency

Objective: Ensure that AI system behaviour and material outputs can be sufficiently understood, reconstructed and independently assessed through model documentation,

appropriate explainability techniques, decision traceability and validation of explanation quality.

Table 47. Explainability and Transparency — Principle and Standards Alignment

Guiding Principles Served	P3 (Transparency & Explainability), P2 (Accountability)
ISO 42001 / NIST Mapping	ISO/IEC 42001 – applicable transparency and information-related controls; NIST AI RMF – MEASURE function, including Explainable and Interpretable characteristics; and other recognised explainability and model-assurance references as applicable.

Table 48. Explainability and Transparency — Audit Controls

Control Area	Audit Requirements
Explainability TechniqueSelection	An appropriate explainability approach shall be selected based on the model architecture, intended use, decision impact, available model access and risk tier. Inherently interpretable models, feature-attribution methods, surrogate models, example-based explanations, counterfactual techniques or other suitable methods may be used. The organisation shall document the rationale for the selected method and its known limitations. Example: Inherent interpretability (decision trees, linear models) for Tier 1–2; post-hoc explanation for Tier 3–4.
Decision Logging	Decision and interaction logging shall capture sufficient information to reconstruct material AI outputs and consequential actions in accordance with system risk. Logging shall consider data sensitivity, minimisation, access control and retention requirements and shall avoid unnecessary capture of sensitive content.
User-Facing Explanations	Plain-language explanation capability for decisions affecting individuals. Explanations accessible to non-technical stakeholders
Auditor Explanation Access	Technical explanations and supporting evidence shall be made available to authorised auditors within a timeframe appropriate to the audit scope, system risk and evidence requirements.
Transparency Reporting	Model cards, data sheets and relevant system documentation shall be maintained and made available to authorised stakeholders and auditors based on system risk, sensitivity and assurance requirements.
Consistency Verification	Verification that explanations are consistent across similar cases. Flagging of inconsistent explanations for investigation.
Explanation Quality	Explanations shall be assessed, as applicable, for fidelity, stability, consistency, completeness and reproducibility. The selected explainability method shall be appropriate to the model architecture, intended use and risk tier.

Auditor Checklist: The auditor shall verify the following:

- ✓ XAI technique is documented and appropriate for the model type and risk tier.
- ✓ Decision logs capture sufficient detail for reconstruction of individual decisions.
- ✓ User-facing explanation mechanism exists for customer-impacting decisions.
- ✓ Explanations are tested for comprehensibility with non-technical stakeholders.
- ✓ Model cards are complete, current, and accessible to authorised parties.
- ✓ Explanation consistency is verified through sampling and comparison.
- ✓ Explanation methods and supporting documentation are sufficient for independent audit, validation and human decision review.

Table 49. Explainability and Transparency — Risk-Tier Applicability

Risk Tier	Applicability
Tier 1 (Low)	Explainability recommended. Basic model documentation.
Tier 2 (Moderate)	Mandatory for customer-facing AI. Basic model card. Decision logging.
Tier 3 (High)	An appropriate explainability technique proportionate to the model architecture, use case and risk shall be implemented and validated. SHAP, LIME or equivalent techniques may be used where appropriate.
Tier 4 (Critical)	Independent explainability assessment; enhanced decision traceability; explanation-quality testing; and availability of sufficient technical documentation to support independent review.

8.6 Security, Adversarial Resilience & Incident Reporting

Objective: Prevent malicious exploitation of AI systems through comprehensive security controls, adversarial robustness testing, vulnerability management, and CERT-In aligned incident response integration.

Table 50. Security, Adversarial Resilience and Incident Reporting — Principle and Standards Alignment

Guiding Principles Served	P1 (Security-by-Design), P5 (Data Security & Privacy), P6 (Auditability & Traceability), P7 (Resilience & Continuous Assurance)
ISO 42001 / NIST Mapping	MITRE ATLAS. OWASP ML/LLM Top 10. NIST AI RMF: Secure and Resilient.

Table 101. Security, Adversarial Resilience and Incident Reporting — Audit Controls

Control Area	Audit Requirements
AI Threat Modelling	Structured threat modelling shall cover relevant attack surfaces across data, models, APIs, prompts, RAG components, tools/agents, infrastructure and supply-chain dependencies.

	Applicable MITRE ATLAS techniques, OWASP guidance, CERT-In advisories/guidelines and other relevant threat intelligence should be considered.
Adversarial Robustness Testing	Pre-deployment testing against adversarial examples (FGSM,PGD), data poisoning detection, and model extraction simulation. Misclassification delta within policy threshold.
AI-Specific VAPT	Vulnerability assessment and penetration testing of AI model endpoints, APIs, training pipelines, model storage, and inference infrastructure.
CERT-In Incident Reporting	AI-related security incidents reported within mandated timelines(6 hours). Forensic readiness with 180-day log retention.
Model Integrity Monitoring	Continuous integrity verification of model artifacts (hashvalidation), detection of unauthorised modifications, and automated rollback capability.
Supply Chain Security	Verification of pre-trained models, open-source libraries, and third-party components for malicious code, licence compliance, and vulnerability exposure.
API Security	TLS enforcement, authentication/authorisation testing, input validation, rate limiting, and output sanitisation for all model-serving APIs.
Infrastructure Hardening	GPU resource isolation, RBAC for AI workloads, network segmentation, patch management, and container image security scanning.

Auditor Checklist: The auditor shall verify the following:

✓ AI threat model is documented using MITRE ATLAS or equivalent taxonomy.
✓ Adversarial robustness testing reports exist with pass/fail criteria against defined thresholds.
✓ AI-specific VAPT has been conducted within the past 12 months.
✓ CERT-In incident reporting process is documented and tested for AI-specific scenarios.
✓ Model artifact integrity verification is automated and continuous.
✓ Supply chain security scanning covers all third-party model components and libraries.
✓ API security testing (auth bypass, fuzzing, TLS verification) is completed and documented.
✓ Log retention meets CERT-In 180-day minimum with time synchronisation across all AI components.

Table 112. Security, Adversarial Resilience and Incident Reporting — Risk-Tier Applicability

Risk Tier	Applicability
Tier 1 (Low)	Standard cybersecurity incident-management process and applicable external reporting requirements.

Tier 2 (Moderate)	AI-specific incident scenarios incorporated into detection, escalation and response procedures.
Tier 3 (High)	Enhanced AI-event detection, forensic readiness, incident playbooks and tested external-reporting capability.
Tier 4 (Critical)	Continuous security monitoring, advanced forensic readiness, tested AI-specific incident response and capability to meet applicable external reporting requirements under critical-event conditions.

8.7 Continuous Monitoring, Drift Management & Validation

> **Objective:** Ensure ongoing performance, accuracy, fairness, and reliability of live AI systems through systematic monitoring, drift detection, re-validation triggers, and governance of retraining cycles.

Table 53. Continuous Monitoring, Drift Management and Validation — Principle and Standards Alignment

Guiding Principles Served	P1 (Security-by-Design), P4 (Fairness), P6 (Auditability & Traceability)
ISO 42001 / NIST Mapping	ISO 42001 Clause 9 (Performance Evaluation), NIST AI RMF: MEASURE and MANAGE functions.

Table 54. Continuous Monitoring, Drift Management and Validation — Audit Controls

Control Area	Audit Requirements
Performance Monitoring	Model-appropriate performance, quality, reliability and operational metrics shall be defined based on the AI architecture and intended use.
Data Drift Detection	Statistical monitoring of input data distributions over time (K-Stest, chi-squared, PSI). Alerts triggered when drift exceeds defined thresholds.
Concept Drift Detection	Monitoring of the relationship between features and target variables. Detection of changes that degrade model predictivevalidity even with stable input distributions.
Fairness Drift Monitoring	Continuous tracking of bias metrics in production. Alerts whenfairness scores degrade beyond acceptable thresholds. Bias score reports generated weekly/monthly.
Re-Validation Triggers	Defined criteria triggering mandatory model re-validation: significant drift, performance degradation, regulatory change, retraining, or material scope change.

Retraining Governance	Documented retraining procedures including data freshness requirements, re-validation steps, approval workflow, and production deployment gates.

Auditor Checklist: The auditor shall verify the following:

- ✓ Performance monitoring dashboards are operational for all production AI systems.
- ✓ Data drift detection is automated with defined statistical tests and alert thresholds.
- ✓ Concept drift monitoring is in place for all Tier 2–4 models.
- ✓ Fairness metrics are tracked in production with documented alert thresholds.
- ✓ Re-validation trigger criteria are documented and enforced.
- ✓ Retraining procedures include approval gates and re-validation requirements.

Table 55. Continuous Monitoring, Drift Management and Validation — Risk-Tier Applicability

Risk Tier	Applicability
Tier 1 (Low)	Annual performance review. Basic monitoring recommended.
Tier 2 (Moderate)	Quarterly performance review. Data drift monitoring. Basic fairness tracking.
Tier 3 (High)	Continuous performance and drift monitoring. Fairness dashboards. Formal re-validation triggers.
Tier 4 (Critical)	Real-time monitoring. Automated drift and fairness alerts. Continuous re-validation. External monitoring audit.

8.8 Human Oversight & Decision Accountability

Objective: Ensure that AI systems with consequential or autonomous capabilities incorporate effective human supervision, intervention, approval and fallback mechanisms proportionate to risk.

Table 56. Human Oversight and Decision Accountability — Principle and Standards Alignment

Guiding Principles Served	P2 (Accountability & Human Oversight), P3 (Transparency & Explainability)
ISO 42001 / NIST Mapping	ISO 42001 Clause 5 (Leadership), Annex A.3 (Internal Organisation). EU AI Act (Human Oversight). NIST AI RMF: GOVERN function.

Table 127. Human Oversight and Decision Accountability — Audit Controls

Control Area	Audit Requirements

Oversight Model	Define HITL, HOTL or other supervisory arrangements based on system autonomy and potential impact.
Approval Gates	Consequential actions shall require human approval where defined by the risk assessment.
Manual Override	Authorised personnel shall be able to interrupt, modify or override AI actions.
Emergency Stop / Safe State	High-risk autonomous systems shall support controlled shutdown or transition to a safe/degraded operating state.
Reviewer Competence	Personnel performing oversight shall be trained to interpret AI outputs, limitations, alerts and escalation conditions.
Override Logging	Human approvals, interventions, overrides and exceptions shall be logged and traceable.
Effectiveness Monitoring	The organisation shall monitor whether human review and override mechanisms are functioning effectively and are not merely procedural.

Auditor Checklist: The auditor shall verify the following:

✓ Human oversight model (HITL/HOTL/HOVL) is defined and documented for each production AI system.

✓ Manual override and intervention mechanisms are periodically tested and re-tested following material changes affecting the oversight mechanism.

✓ Approval gates, human intervention, override and fallback mechanisms are implemented and tested proportionate to the system's risk, autonomy and potential impact.

✓ Human reviewer training programme is operational with documented competence assessments.

✓ Decision accountability records link AI outputs to responsible human approvers.

✓ Override accuracy and human review effectiveness are periodically measured.

Table 58. Human Oversight and Decision Accountability — Risk-Tier Applicability

Risk Tier	Applicability
Tier 1 (Low)	Designated system owner and supervisory oversight. Basic intervention or override capability recommended where applicable.
Tier 2 (Moderate)	Defined human-review mechanism, override capability and escalation procedure for material exceptions or predefined triggers.

Tier 3 (High)	Human review or approval required for consequential decisions as defined by the risk assessment. Override, escalation and decision traceability shall be implemented and periodically tested.
Tier 4 (Critical)	Enhanced supervisory review for critical decisions/actions. Approval, intervention, emergency override, safe-state/fallback and recovery mechanisms shall be independently tested.

8.9 Third-Party and AI Supply-Chain Assurance

Objective: Ensure that risks introduced by third-party AI models, datasets, libraries, platforms, APIs and other supply-chain components are identified, assessed and continuously managed.

Table 59. Third-Party and AI Supply-Chain Assurance — Principle and Standards Alignment

Guiding Principles Served	P1 (Security-by-Design), P2 (Accountability), P5 (Data Security & Privacy), P6 (Auditability & Traceability)
ISO 42001 / NIST Mapping	ISO 42001 Annex A.10 – Third-party and customer relationships. NIST AI RMF:GOVERN function.

Table 13. Third-Party and AI Supply-Chain Assurance — Audit Controls

Control Area	Audit Requirements
Vendor Due Diligence	Assess security posture, model governance, vulnerability-management capability, incident handling and service dependency.
Model and Component Provenance	Verify origin, model version, provider, cryptographic integrity and authorised source of third-party models/components.
AIBOM/SBOM	Maintain visibility of AI models, libraries, frameworks, containers, datasets and other material dependencies.
Contractual Security Requirements	Define security obligations, vulnerability notification, incident notification, change notification and audit/assurance rights.
Provider Change Management	Assess impact of provider-side model, API, architecture or security changes before or after material updates.
AI-as-a-Service Monitoring	Maintain consumer-side logging, security monitoring, performance monitoring and rate/control enforcement.
Dependency and Vulnerability Monitoring	Periodically monitor third-party AI dependencies for vulnerabilities and material security changes.

Exit, Portability and Fallback Strategy	Critical third-party AI dependencies shall have documented exit, portability and fallback arrangements covering alternative providers or models, data/model retrieval, knowledge transfer, continuity of critical functions, dependency removal and preservation of required audit evidence.

Auditor Checklist: The auditor shall verify the following:

- ✓ Vendor due diligence is completed and documented for material AI suppliers and service providers.
- ✓ Third-party model and component provenance, provider, version and integrity are verified where technically possible.
- ✓ AIBOM/SBOM records material third-party models, libraries, frameworks, containers, datasets and dependencies.
- ✓ Contracts or service agreements define applicable security obligations, incident notification, vulnerability notification, provider change notification and assurance/audit rights.
- ✓ Provider-side model, API or architecture changes are assessed for material impact.
- ✓ AI-as-a-Service usage includes appropriate consumer-side logging, security monitoring and performance monitoring.
- ✓ Material third-party dependencies are monitored for security vulnerabilities and significant provider changes.
- ✓ Independent assurance reports or alternative evidence are evaluated where direct technical access is unavailable.
- ✓ Critical AI dependencies have documented and tested exit, portability or fallback arrangements.

Table 14. Third-Party and AI Supply-Chain Assurance — Risk-Tier Applicability

Risk Tier	Applicability
Tier 1 (Low)	Basic vendor/source verification and standard security requirements.
Tier 2 (Moderate)	Structured due diligence, defined security requirements, provenance checks and dependency monitoring.
Tier 3 (High)	Comprehensive supplier assurance, AIBOM/SBOM, provider-change assessment, vulnerability monitoring and formal exit/fallback planning.
Tier 4 (Critical)	Enhanced independent supplier assurance, continuous monitoring of critical dependencies, integrity verification, tested fallback/exit arrangements and vulnerability tracking.

9. Advanced AI Security Requirements

This section establishes dedicated security and audit requirements for Generative AI, Retrieval-Augmented Generation (RAG), Agentic AI, Multimodal AI, Multi-Agent AI and Autonomous/Cyber-Physical AI systems. In addition to traditional generative-model risks such as hallucination and prompt injection, these architectures introduce additional attack surfaces involving knowledge retrieval, vector databases, tool invocation, persistent memory, autonomous actions, cross-agent interactions and non-text input modalities. The requirements in this section supplement, but do not replace, the core audit domains in Section 8.

9.1 Why Generative AI Requires Dedicated Audit Provisions

Generative AI systems differ from traditional predictive and other bounded-output AI systems in several governance-critical dimensions that existing audit frameworks were not designed to address:

Table 15. Governance Challenges of Generative AI

Dimension	Governance Challenge
Output Unpredictability	Traditional AI produces bounded outputs (classifications, scores). Generative AI produces unbounded, novel content that may contain fabricated facts, harmful material, copyrighted content, or biased narratives—often with high confidence and apparent authority.
Hallucination Risk	LLMs generate outputs that are factually incorrect, fabricated, or unsupported by training data, yet presented with linguistic fluency that makes errors difficult to detect. This poses acute risks in regulated contexts (legal, medical, financial).
Prompt Injection Vulnerability	Unique to LLM-based systems: malicious instructions embedded in user inputs (direct injection) or retrieved context (indirect injection) can hijack model behaviour, bypass safety controls, and exfiltrate data.
Training Data Scale and Opacity	Foundation models are trained on internet-scale corpora (billions of data points) where individual data provenance, consent verification, and bias screening at the record level is infeasible. Traditional data governance approaches require adaptation.
Dual-Use and Misuse Potential	Generative AI can be weaponised for deepfake creation, automated social engineering, disinformation at scale, synthetic identity fraud, and automated vulnerability discovery—risks that do not arise from traditional ML systems.
Downstream Amplification	A single foundation model may serve millions of users across diverse contexts. A flaw, bias, or vulnerability in the base model is amplified across every downstream application, creating systemic risk.
Agentic AI Capabilities	LLMs integrated with tool use, code execution, and autonomous action capabilities (agentic AI) can take real-world actions beyond text generation, introducing risks of unintended consequences, excessiveagency, and loss of human control

9.2 Generative AI Threat Landscape

Table 16. Generative AI Threats and Audit Focus

Threat	Primary Audit Focus
Prompt Injection	Test direct, indirect, RAG, tool-output, persistent-memory and multimodal prompt-injection paths; verify trust boundaries, least privilege, output controls and containment of successful injection.
Sensitive Information Disclosure	Test leakage of confidential data, retrieved documents, model/system context, credentials and sensitive training information through model outputs or tool interaction.
Excessive Agency	Assess agent permissions, available functionality, tool access, action limits, human approval gates, service identities and ability to prevent unauthorised consequential action.
Supply Chain	Assess model, dataset, library, tool, plugin, MCP/service and dependency provenance; verify signing, integrity, AIBOM/SBOM and change monitoring.
Data and Model Poisoning	Test protection against training, fine-tuning, RAG, vector-store, memory and model-artifact poisoning and verify dataset/model integrity controls.
Unbounded Consumption	Assess resource, token, compute and cost exhaustion; verify rate limits, quotas, execution budgets, timeouts and workload monitoring.
Misinformation	Assess hallucination/confabulation, unsupported claims, fabricated citations, unsafe reliance and downstream use of inaccurate model outputs.
Hidden Context Exposure	Assess inappropriate disclosure of system context, hidden instructions, internal metadata or other information intended to remain outside the user's accessible context.
Vector and Embedding Weaknesses	Test retrieval isolation, embedding/vector-store access controls, cross-tenant leakage, poisoned embeddings, unauthorised retrieval and stale/deleted content exposure.
Improper Output Handling	Verify validation and sanitisation of model outputs before they are rendered, executed, passed to APIs/tools or consumed by downstream systems.

9.3 Generative AI-Specific Audit Requirements

In addition to the core audit domain requirements (Section 8), organisations deploying generative AI systems shall implement and be audited against the following controls:

Table 17. Generative AI-Specific Audit Controls

Control Area	Audit Requirements
Input Security and Prompt Governance	System, developer, user, retrieved and tool-provided content shall be treated according to defined trust boundaries. Controls shall include separation of trusted instructions from untrusted content where technically feasible; content inspection; least-privilege tool and data access; approval gates for consequential actions; output validation; monitoring; and adversarial testing. Input sanitisation alone shall not be relied upon as the primary control against prompt injection.
Output Integrity and Verification	Output filtering for harmful, biased, PII-containing, or copyrighted content before delivery to users. Factual grounding mechanisms (RAG with source verification, citation requirements) for information-critical applications. Confidence scoring or uncertainty quantification where technically feasible. AI-generated content watermarking or labelling in compliance with emerging content provenance standards.
Hallucination Management	Documented hallucination risk assessment for each generative AI use case. Retrieval-augmented generation (RAG) or knowledge grounding deployed for factual applications. Risk-based human review shall be implemented for high-impact or consequential outputs where reliance on incorrect or unsupported information may result in material impact.
Training Data and Fine-Tuning Governance	For fine-tuned models, the authorised use, source, provenance, integrity, quality and bias/representation characteristics of fine-tuning datasets shall be documented and assessed. For foundation models consumed through APIs, available provider information concerning training and security practices shall be evaluated and limitations documented. Sensitive, confidential or otherwise restricted data categories shall be identified and appropriately controlled throughout training and fine-tuning.
Agentic AI and Tool Use Controls	Agentic AI systems shall additionally comply with Section 9.7
Deepfake and Synthetic Media Controls	Organisational policy prohibiting generation of deepfake content of real individuals without explicit consent. Integration with deepfake detection capabilities for organisations receiving AI-generated media. Content provenance marking (C2PA or equivalent) for all synthetic media generated by organisational AI systems. Incident reporting fordeepfake-related fraud or impersonation events.
Intellectual Property andCopyright	Documented assessment of copyright risk for each generative AI deployment. Output screening mechanisms for copyrighted content reproduction. Organisational IP policy governing ownership and usage rights of AI-generated content. Vendor contractual provisions regarding indemnification for IP infringement arising from model outputs.
User Transparency and Disclosure	Clear disclosure to users when they are interacting with generative AI (not a human). Identification of AI-generated content in all external-facing outputs (reports, communications, documents). Accessible explanation of generative AI system capabilities, limitations, and known failure modes. Prohibition of deceptive use of generative AI to impersonate individuals or fabricate evidence.

9.4 Generative AI Auditor Checklist

The following checklist supplements the domain-specific checklists in Section 8. Auditors assessing generative AI systems shall verify:

✓ Prompt injection testing has been conducted (red-team and automated) with documentedresults and remediation of identified vulnerabilities.

✓ System and developer instructions do not contain credentials or other secrets and are not relied upon as the sole security boundary for authorisation, privilege separation or policy enforcement. Exposure and prompt-injection scenarios have been tested, and security controls remain effective even where hidden context is partially disclosed.

✓ Output filtering pipeline is operational and tested for harmful content, PII leakage, and copyrighted material.

✓ Hallucination risk assessment is documented for each use case, with appropriate mitigation (RAG, human review, disclaimers).

✓ Fine-tuning datasets have documented provenance, consent status, and bias screeningreports.

✓ For consumed foundation model APIs: vendor training data practices are documented or risk acceptance is formally recorded.

✓ Agentic AI capabilities (if present) have least-privilege permissions, human approval gates,and kill switch capability.

✓ AI-generated content is labelled or watermarked in all external-facing outputs.

✓ User-facing disclosure clearly identifies interactions as AI-generated.

✓ Deepfake generation policy exists and prohibits non-consensual synthetic media of real individuals.

✓ Copyright risk assessment is documented for each generative AI deployment.

✓ Incident response procedures specifically address generative AI failure modes (hallucination events, prompt injection breaches, deepfake misuse).

✓ Performance monitoring tracks generation quality metrics (factual accuracy, safety filtertrigger rates, user feedback scores).

✓ Evidence of periodic and event-triggered reassessment of Generative AI systems for emerging threats, based on system risk, material changes and threat evolution.

9.5 Advanced AI Risk Escalation

Generative, RAG, agentic, multimodal, multi-agent and autonomous/cyber-physical AI systems shall be evaluated using the risk factors defined in Section 5.3. The following capabilities or deployment characteristics shall be considered as risk-escalation factors:

i. external or unrestricted public exposure;
ii. processing of sensitive or critical information;
iii. use in consequential business, safety or operational decisions;
iv. privileged access to production systems, data stores or administrative functions;
v. ability to invoke tools, execute code, initiate transactions or communicate externally;
vi. persistent memory capable of influencing future actions;
vii. autonomous task planning or multi-step execution;
viii. delegation to other agents or dynamically selected tools;

ix. operation in critical infrastructure or safety-critical environments;
x. inability to reliably interrupt, rollback or place the system in a safe state;
xi. opaque third-party model dependencies that materially restrict assurance; and
xii. large-scale downstream deployment where a model-level failure may have systemic impact.

Where one or more such characteristics materially increase potential impact or adversarial exposure, the assigned tier shall be escalated or enhanced audit requirements shall be applied and the rationale shall be documented.

9.6 RAG and Vector Database Security

RAG-based systems shall implement controls covering knowledge-source provenance, secure ingestion, document and chunk-level access control, tenant isolation, vector-store security, indirect prompt injection, poisoned knowledge sources, retrieval relevance, stale content, deletion propagation and retrieval logging.

Auditors shall verify that retrieved content is treated as untrusted input and cannot override higher-priority system instructions or bypass user-level access restrictions.

9.7 Agentic AI Security

Agentic AI systems shall have documented identities, authorised objectives, approved tools, defined permission boundaries and action limits. Agents shall operate under least privilege and shall not automatically inherit permissions from users, tools or other agents.

Consequential actions shall use appropriate human approval gates. Agent execution metadata, state transitions, tool calls, delegated tasks, approvals, external actions and outputs required to reconstruct the execution chain shall be logged and protected against unauthorised modification.

High-risk agentic systems shall support emergency interruption, credential revocation, rollback and prevention of uncontrolled automatic restart.

Agentic AI systems that interact with external tool servers, plugins, Model Context Protocol (MCP) services, agent-to-agent communication mechanisms or dynamically discovered services shall maintain an approved inventory of permitted connections. Tool/service identities shall be authenticated, permissions shall follow least privilege, connection metadata and tool descriptions shall be validated, and security-relevant tool calls shall be logged. Dynamic or newly introduced tool connections shall undergo appropriate provenance, integrity and security assessment before production use.

9.8 Multimodal AI Security

Multimodal systems shall be assessed for attacks and failures arising from text, image, audio, video, documents, metadata and other supported modalities.

Testing shall include malicious or hidden instructions embedded in media, cross-modal prompt injection, manipulated files, spoofed inputs, modality-specific adversarial examples and inconsistent behaviour across equivalent inputs presented in different modalities.

9.9 Multi-Agent AI Security

Multi-agent AI systems introduce additional security risks arising from communication, delegation, distributed identity, shared context and cascading actions across multiple autonomous or semi-autonomous agents.

Organisations deploying multi-agent systems shall implement the following controls:

Table 18. Multi-Agent AI Security Controls

Control Area	Audit Requirement
Agent Identity	Each agent shall have a uniquely identifiable service identity with authenticated and attributable actions. Shared or anonymous agent credentials shall be avoided for high-risk systems.
Inter-Agent Authentication	Communication between agents shall be authenticated and integrity-protected using approved mechanisms.
Delegation Boundaries	Agents shall delegate tasks only to approved agents, tools or services and shall not automatically propagate their full privileges to delegated agents.
Privilege Containment	Each agent shall operate under least privilege appropriate to its assigned function. Privilege inheritance across agent chains shall be explicitly controlled.
Message Trust	Agent-to-agent messages shall be treated as potentially untrusted input unless their source and integrity are independently verified.
Context and Memory Isolation	Shared context, memory and state shall be protected against poisoning, cross-agent leakage and unauthorised modification.
Circular / Recursive Delegation	Controls shall detect and limit recursive delegation, uncontrolled task spawning and agent loops that may cause resource exhaustion or uncontrolled behaviour.
Cascading Failure Controls	Failure or compromise of one agent shall not automatically cause uncontrolled downstream actions across the multi-agent workflow.
Action Traceability	Logs shall enable reconstruction of agent-to-agent requests, delegated tasks, permissions, decisions, tool invocations and resulting external actions.

Emergency Containment	High-risk systems shall support suspension of individual agents or the complete multi-agent workflow and revocation of associated credentials.

Auditor checks:

- ✓ Unique agent identities and permissions are documented.
- ✓ Inter-agent communication is authenticated and logged.
- ✓ Delegation does not result in uncontrolled privilege inheritance.
- ✓ Agent-to-agent inputs are assessed for prompt injection and message spoofing.
- ✓ Circular delegation, uncontrolled agent spawning and cascading-failure scenarios have been tested.
- ✓ Shared memory/context is tested for poisoning and cross-agent leakage.
- ✓ A complete multi-agent execution chain can be reconstructed from logs.
- ✓ Emergency isolation or shutdown mechanisms are periodically tested.

9.10 Autonomous and Cyber-Physical AI Security

AI systems capable of controlling or influencing physical processes, equipment or safety-critical operations shall be assessed for both cybersecurity and operational safety implications.

Table 19. Autonomous and Cyber-Physical AI Security Controls

Control Area	Audit Requirement
Operational Boundaries	Authorised operating conditions, physical boundaries, environmental constraints and conditions under which autonomous operation is permitted shall be defined.
Sensor and Input Integrity	Controls shall detect spoofed, manipulated, unavailable or contradictory sensor and environmental inputs.
Independent Safety Controls	Where practicable, safety interlocks and emergency controls shall operate independently of the AI model being controlled.
Safe State / Degraded Mode	The system shall transition to a predefined safe state or restricted operating mode when confidence falls below threshold, critical inputs fail or unsafe behaviour is detected.
Manual Takeover	Authorised personnel shall be able to assume control or suspend autonomous operation within defined operational requirements.
Physical Action Boundaries	Geographical, speed, transaction, actuator, movement or other operational constraints shall be technically enforced where applicable.

Simulation and Fault Testing	Applicable Tier 3 and Tier 4 systems shall undergo simulation, fault injection, edge-case and hardware-in-the-loop testing before material production deployment or change.
Secure Updates	Model, firmware and software updates shall be authenticated, tested, approved and capable of rollback.
Physical Event Traceability	Sensor inputs, material AI outputs, physical commands, human interventions and safety events shall be time-synchronised and logged.
Emergency Shutdown	Safety-critical systems shall provide an independently controlled mechanism to interrupt unsafe autonomous operation.

10. Cybersecurity & Resilience

This section consolidates cybersecurity and resilience requirements specific to AI systems, expanding upon the controls in Section 8.6 with detailed infrastructure hardening, vulnerability management, SOC integration, incident response and business-continuity provisions. The requirements are informed by applicable CERT-In Directions, advisories and technical guidance, recognised cybersecurity standards, and AI-specific threat frameworks including MITRE ATLAS and OWASP.

While Section 8.6 addresses security from an audit domain perspective (what to audit), this section addresses security from an operational and infrastructure perspective (how to secure), providing the technical depth that organisations need for implementation and auditors need for verification.

10.1 AI-Specific Attack Surface

AI systems introduce new and amplified attack surfaces that extend beyond conventional IT security boundaries. Understanding this expanded attack surface is essential for effective security design and audit:

Table 20. AI Attack Surfaces, Threats and Security Controls

Attack Surface Component	Threat Description	Key Security Controls
Training Data Pipeline	Data ingestion endpoints, ETL processes, data storage, feature engineering pipelines. Vulnerable to data poisoning, data exfiltration, and supply chain contamination.	Hash verification of datasets, access control on data pipelines, anomaly detection on incoming data, encrypted storage, lineage tracking.
Model Development Environment	Notebooks, experiment tracking, hyperparameter stores, model registries, code repositories. Vulnerable to insider threats, IP theft, and backdoor insertion.	RBAC for development environments, code review processes, model artifact signing, secure experiment tracking, access logging.
Model Artifacts	Serialised model files (.pkl, .onnx, .pt, .h5), configuration files, weights, and checkpoints. Vulnerable to model theft, tampering, and malicious code injection.	Digital signatures, integrity hashing (SHA-256), encrypted storage, access-controlled model registry, version control with audit trail.
Inference API Endpoints	REST/gRPC endpoints serving model predictions. The primary external-facing attack surface. Vulnerable to adversarial inputs, model extraction, DDoS, prompt injection (LLMs), and data exfiltration.	TLS enforcement, authentication/authorisation, rate limiting, input validation and sanitisation, output filtering, API gateway with WAF.

GPU/TPU Infrastructure	Specialised compute resources for training and inference. Vulnerable to resource hijacking (cryptomining),side-channel attacks, and shared-tenancy risks in cloud environments.	GPU resource isolation, namespace separation in Kubernetes, dedicated compute for sensitive workloads, monitoring of GPU utilisation anomalies.
MLOps/CI-CD Pipeline	Automated pipelines for model training, validation, testing, and deployment. Vulnerable to supplychain compromise, pipeline poisoning, and unauthorised deployment.	Pipeline integrity verification, automated security scanning in CI/CD, approval gates before production deployment, immutable build artifacts.
Monitoring and Logging Infrastructure	Performance dashboards, drift detection systems, SIEM integration, log aggregation. Vulnerable to log tampering, monitoring evasion, and alert suppression.	Immutable logging, centralised log aggregation with integrity verification,SIEM alert rules for AI-specific events, 180-day retention (CERT-In).

10.2 CERT-In Cybersecurity Requirements and Technical Guidance for AI Systems

Organisations shall identify CERT-In Directions, advisories, guidelines, vulnerability notes, blueprints and other technical references applicable to the AI system and its supporting digital infrastructure.

Auditors shall determine applicability based on the system architecture, exposure, technology stack, threat environment, deployment model and supporting infrastructure.

Applicable guidance shall be considered while assessing governance, asset management, access control, infrastructure hardening, vulnerability management, monitoring, incident response, third-party dependencies and technical assurance.

10.3 AI Infrastructure Hardening Requirements

Organisations shall implement the following infrastructure security controls for all AI system deployments:

Table 21. AI Infrastructure Hardening Controls

Control Area	Requirements
Network Segmentation	AI workloads shall be appropriately segmented based on system risk, trust boundaries, data sensitivity, exposure and potential attack paths. Development/training and production inference environments shall be segregated where compromise of one environment could materially affect another. Higher-risk AI workloads shall use enhanced logical or physical isolation where warranted.

Container and Orchestration Security	AI workloads deployed in containers (Docker, Kubernetes) shall use: hardened base images scanned for vulnerabilities (Trivy, Grype), RBACfor Kubernetes namespaces and pods, Pod Security Admission enforcing applicable Pod Security Standards or equivalent admission-control mechanisms preventing privilege escalation and insecure workload configurations, network policies restricting inter-pod communication, and image signing with verification before deployment.
GPU Resource Isolation	GPU resources shall be allocated to specific AI workloads with isolation from other tenants. In cloud environments, dedicated GPU instances are recommended for Tier 3–4 workloads. GPU utilisation shall be monitored for anomalies (cryptomining, unauthorised training jobs).
Secrets Management	API keys, model-signing keys, database credentials, service credentials and other secrets used by AI systems shall be generated, stored, distributed, rotated and revoked using approved secrets-management mechanisms. Secrets shall not be hard-coded in source code or stored in unsecured configuration. Runtime secret injection shall use controlled and auditable mechanisms.
Encryption Standards	Approved industry-standard cryptography shall be used for data at rest and in transit. AES-256 and TLS 1.2+ (preferably TLS 1.3 where supported), or cryptographically equivalent mechanisms, may be used based on applicable organisational and external requirements.
Patch Management	Vulnerabilities affecting AI frameworks, libraries, operating systems, container platforms, model-serving infrastructure and other dependencies shall be remediated according to risk-based vulnerability-management SLAs. Prioritisation shall consider exploitability, known exploitation, asset exposure, system criticality, dependency reach and availability of compensating controls. Where immediate patching is not feasible, documented interim mitigation and risk acceptance shall be implemented.

10.3.1 Secure MLOps / LLMOps Pipeline

AI development, training, evaluation and deployment pipelines form part of the security boundary of the AI system and shall be protected against unauthorised modification, supply-chain compromise and uncontrolled model promotion.

Table 69. Secure MLOps / LLMOps Pipeline Controls

Control Area	Requirement
Environment Segregation	Development, testing, staging and production AI environments shall be appropriately segregated.
Version Control	Model code, training code, datasets, model artefacts, prompts, guardrails, configurations and deployment manifests shall be version controlled where applicable.
Pipeline Identity	Training, evaluation and deployment jobs shall execute using

	authenticated identities with least-privilege permissions.
Source and Dependency Security	Code, libraries, containers and dependencies shall undergo appropriate code, dependency, secret and vulnerability scanning.
Dataset Validation Gates	Training and fine-tuning data shall undergo defined provenance, quality, integrity and security checks before use in production model pipelines.
Model Artefact Integrity	Production model artefacts shall be cryptographically hashed or signed and integrity shall be verified before deployment/loading.
AIBOM/SBOM Generation	Material model, software, dataset and service dependencies shall be captured through applicable AIBOM/SBOM processes.
Evaluation Gates	Model promotion shall require completion of applicable performance, security, fairness, robustness and regression evaluations.
Separation of Duties	Production model promotion shall require an approval mechanism independent from the individual initiating the model build or training job where risk warrants.
Automated Retraining	Automated retraining shall not automatically promote a new model into production without defined validation and approval criteria.
Controlled Deployment	Tier 3 and Tier 4 model changes should use staged, canary, shadow or equivalent controlled deployment where technically feasible.
Rollback	The pipeline shall maintain the ability to restore a known-good model, configuration and associated dependencies.
Pipeline Monitoring	Unauthorised pipeline modification, anomalous training jobs, unexpected model artefacts and abnormal compute usage shall generate security alerts where applicable.

Auditor Checklist:

- ✓ Development, test and production AI environments are appropriately segregated.
- ✓ Model/code/data/prompt changes are traceable to authorised versions and approvals.
- ✓ Training and deployment jobs use authenticated service identities and least privilege.
- ✓ Applicable source-code, dependency, secret and container scanning is integrated into AI pipelines.
- ✓ Production model artefacts are integrity-verified before deployment.
- ✓ Model promotion requires completion of defined security, performance and regression criteria.
- ✓ Automated retraining cannot bypass model-validation and release-approval gates.
- ✓ Rollback to a known-good model/configuration state has been tested for applicable high-risk systems.

10.4 AI-Specific Vulnerability Assessment and Penetration Testing

The tools referenced in this section are illustrative. Equivalent commercial, open-source or internally developed tools and techniques may be used where they provide appropriate test coverage, evidence quality and reproducibility.

Standard VAPT methodologies must be supplemented with AI-specific testing to address the unique vulnerability classes of AI systems:

Table 22. AI-Specific Vulnerability Assessment and Penetration Tests

Test Category	Description and Tools
Traditional Infrastructure VAPT	Standard vulnerability scanning and penetration testing of AI servers, network infrastructure, cloud configuration, and web application interfaces. Tools: Nessus, Qualys, Burp Suite, OWASP ZAP.
API Security Testing	Authentication bypass testing, authorisation testing, input validation/fuzzing, rate limit verification, TLS configuration assessment, anderror handling review for all model-serving API endpoints.
Adversarial Robustness Testing	Testing model resilience against adversarial examples using appropriate adversarial techniques, such as FGSM, PGD, C&W or architecture-specific alternatives where technically applicable.
Model Extraction Testing	Simulating high-volume query attacks to assess whether the model's decision boundaries can be reconstructed. Testing rate limiting effectiveness,query pattern detection, and output information leakage.
RAG Security Testing	Test indirect prompt injection, poisoned documents, unauthorised retrieval, document/chunk-level access-control bypass, vector-store tenant isolation and stale/deleted content retrieval.
Agentic AI Security Testing	Test excessive permissions, unauthorised tool invocation, approval-gate bypass, action-budget bypass, credential abuse, unsafe delegation and emergency-stop effectiveness.
Agent Memory Testing	Test memory poisoning, cross-user leakage, retention of sensitive information and unauthorised reuse of persistent context.
Multimodal Security Testing	Test hidden or malicious instructions in images, audio, video and documents, cross-modal prompt injection, malformed media and modality-specific adversarial inputs.

Data Poisoning Detection	Anomaly detection on training datasets using isolation forests or statistical divergence measures. Verification of data source integrity. Assessment of label corruption risk.
Prompt Injection Testing (LLMs)	Red-team testing of prompt boundaries: direct injection (user input manipulation), indirect injection (context poisoning via retrieved documents), system prompt extraction attempts,and jailbreak testing.
Supply Chain Security Scanning	Scanning of third-party model files for embedded malicious code (ModelScan). Dependency vulnerability scanning for ML libraries. Container image scanning for all AI deployment artifacts.
AI Privacy Attack Testing	Assess resistance to model inversion, membership inference, attribute inference and training-data extraction attacks where relevant. Testing shall evaluate whether model outputs, confidence information, embeddings or repeated queries can expose sensitive information about training data or users.

AI-specific security testing shall be repeated following material changes to the model, system prompt, RAG corpus, embedding model, tool/plugin set, agent permissions, deployment architecture, major dependencies or provider model version.

10.5 Security Operations Centre (SOC) Integration

The organisation's SOC shall be equipped to monitor, detect, investigate, and respond to AI-specific security events. AI monitoring shall not be treated as a separate function but integrated into the existing SOC workflow:

Table 71. Security Operations Centre Integration Requirements

SOC Capability	Requirements
AI Log Ingestion	SOC SIEM (Splunk, ELK, QRadar, Sentinel) shall ingest logs from: model inference API calls, data pipeline processing events, model training and retraining runs, model loading and version change events, user authentication to AI systems, and GPU/compute resource utilisation.
AI-Specific Detection Rules	Custom SIEM correlation rules shall address, as applicable: unusual spikes in API calls (potential model scraping or extraction); high-volume or high-similarity repeated queries (potential automated probing or extraction); repeated edge-case inputs (adversarial probing); sudden prediction-confidence degradation; unexpected model-version changes; anomalous data-ingestion volumes; unusual agent/tool invocation patterns; and other AI-specific security events identified through threat modelling.
Alert Triage Procedures	AI-specific alert triage playbooks integrated into SOC standard operating procedures. Each AI alert type shall have: severity classification, initial triage steps, escalation criteria, and response procedures.

AI Incident Playbooks	Documented playbooks for: adversarial attack containment and model rollback, data poisoning investigation and dataset quarantine, model extraction incident response and API lockdown, prompt injection breach containment (LLMs), deepfake incident response and evidence preservation, AI-driven fraud investigation.
SOC Analyst Training	SOC analysts trained on AI-specific threat indicators and attack techniques. Training shall cover MITRE ATLAS taxonomy, OWASP ML/LLM Top 10, AI log interpretation, and adversarial attack recognition. Annual refresher training mandatory.
Red Team Exercises	For Tier 4 systems: annual red team exercises simulating AI-specific attacks (adversarial inputs, model extraction, prompt injection, supply chain compromise). Results fed into SOC detection rule improvement and playbook updates.

10.6 AI Incident Response and Business Continuity

AI-specific incident response extends the organisation's existing incident response plan with procedures tailored to AI failure modes:

Table 23. AI Incident Response and Business Continuity Requirements

Capability	Requirements
AI Incident Classification	AI operational incidents shall be classified independently from audit findings. Incident severity shall consider: • active exploitation status; • confidentiality, integrity or availability impact; • model or dataset compromise; • scope of affected users/systems; • safety and operational impact; • persistence of the compromise; • recoverability and rollback capability; and • potential cascading impact. The incident-classification methodology shall be documented and aligned with the organisation's existing cybersecurity incident-management process.
CERT-In Reporting Integration	AI incidents meeting CERT-In reportable criteria shall be reported within 6 hours.

Model Rollback Capability	Production AI systems shall maintain a documented rollback or fallback capability appropriate to system criticality and risk. Tier 3 and Tier 4 systems shall define and test recovery objectives for restoring a known-good model or alternate processing mode. Recovery objectives shall consider operational criticality, safety impact and applicable RTO/RPO requirements.
Data Quarantine Procedures	Where data poisoning or dataset compromise is suspected, affected data shall be isolated or quarantined within a risk-appropriate containment objective. Procedures shall address isolation of the affected dataset, suspension of impacted training/retraining activities, assessment of contamination scope, preservation of evidence and validation prior to reuse.
Business Continuity for AI	AI systems classified as Tier 3–4 shall have documented business continuity plans including: fallback to previous model version, manual processing procedures for critical decisions, degraded-mode operation (rule-based fallback), recovery time objectives (RTO) and recovery point objectives (RPO), and communication plans for affected stakeholders.

10.7 Tier-Specific Cybersecurity Requirements

Table 24. Cybersecurity Requirements by Risk Tier

Risk Tier	Cybersecurity Requirements
Tier 1 (Low)	Standard cybersecurity controls; baseline AI attack-surface assessment; access control, encryption and applicable vulnerability testing.
Tier 2 (Moderate)	AI-specific threat modelling; API security testing; model integrity verification; AI-related logging and monitoring; periodic AI security assessment.
Tier 3 (High)	Comprehensive AI VAPT covering all applicable test categories in Section 10.4; adversarial testing; AI SOC integration; supply-chain/dependency scanning; tested rollback/fallback capability; enhanced incident readiness.
Tier 4 (Critical)	All Tier 3 requirements plus continuous monitoring, advanced adversarial testing, periodic red-team exercises, dedicated AI incident playbooks, enhanced forensic readiness, tested business continuity and safe-state/recovery mechanisms.

10.8 Cybersecurity & Assurance Completion Check

Before concluding the cybersecurity and resilience assessment, the auditor shall confirm that:

- ✓ AI-specific attack surface and trust boundaries have been identified and assessed;
- ✓ Applicable CERT-In Directions, advisories, guidelines and technical references have been identified and considered;
- ✓ All applicable technical test categories under Section 10.4 have been completed or

documented as scope limitations;

- ✓ AI-specific security logs, monitoring and SOC detection coverage have been assessed;
- ✓ Incident-response, rollback/fallback and recovery mechanisms have been practically tested where required by the risk tier; and
- ✓ Vulnerabilities, control failures and residual risks identified through technical testing have been incorporated into the audit findings and remediation process under Section 12.

11. Contestability and Human Decision Review

11.1 Purpose and Applicability

For the purpose of these Guidelines, contestability refers to the ability to subject an AI-generated decision, recommendation or action to a defined process of review, validation, intervention, override, correction or escalation where its accuracy, safety, reliability or appropriateness is questioned.

Contestability provides an additional assurance mechanism for AI systems whose outputs materially influence operational, financial, safety-related or other consequential decisions. It enables incorrect, anomalous, biased, low-confidence or unsafe AI outputs to be identified and reviewed by authorised personnel before or after execution, depending on the system architecture and risk profile.

Section 8.8 defines the human-oversight controls that organisations are expected to implement. This section supplements those requirements by defining the decision-review mechanisms and audit procedures used to verify that such oversight operates effectively in practice.

Contestability and decision-review requirements shall be proportionate to the AI system's risk tier, autonomy, potential impact, reversibility of outcomes and level of human involvement.

11.2 Technical Control Requirements

Table 25. Contestability and Human Decision-Review Controls

Control Area	Audit Requirement
Decision Review Criteria	The organisation shall define the circumstances under which an AI-generated decision, recommendation or action requires human review. Triggers may include low confidence, anomalous output, high-impact action, policy exception, security alert, material model disagreement or other predefined conditions.
Human Review Mechanism	Authorised reviewers shall be provided with sufficient information to assess the AI output, including relevant inputs, model output, confidence or uncertainty information where available, applicable explanation and supporting context.
Escalation and Exception Handling	Defined escalation mechanisms shall exist for cases where the reviewer cannot validate the AI output, where repeated exceptions occur or where the potential impact exceeds the reviewer's authority.
Explanation Availability	Sufficient technical information shall be available to support meaningful review of material AI outputs. Explanation methods shall be appropriate to the model architecture, use case and risk level.
Review Independence	For high-risk decisions, the organisation shall evaluate whether the reviewer has sufficient independence from the model development or automated decision process to provide effective challenge.
Review Effectiveness Monitoring	The organisation shall periodically assess whether human review is effective, including analysis of overrides, reviewer agreement/disagreement, repeated exceptions, review latency and evidence of inappropriate automatic

	acceptance of AI outputs.
Feedback to Model Governance	Repeated overrides, corrections or review failures shall be analysed to identify potential model, data, workflow or control deficiencies and shall feed into model monitoring and re-validation processes.

11.3 Audit Testing of Contestability and Decision Review

Auditors shall assess both the design and operating effectiveness of contestability and human decision-review mechanisms.

Audit procedures should include, as applicable:

- ✓ inspect the documented criteria and triggers requiring human review;
- ✓ sample AI-generated decisions or actions and verify that required reviews were performed;
- ✓ verify that reviewers received sufficient information to make an informed assessment;
- ✓ test whether mandatory approval gates can be bypassed through alternate interfaces, APIs or workflows;
- ✓ verify that AI outputs can be rejected, corrected or overridden by authorised personnel;
- ✓ test escalation mechanisms for low-confidence, anomalous, disputed or high-impact AI outputs;
- ✓ inspect audit logs to verify traceability of AI outputs, human approvals, interventions and overrides;
- ✓ assess whether reviewer access and permissions are appropriate and segregated from unauthorised users;
- ✓ review training and competence records of personnel performing human review;
- ✓ analyse override frequency, reviewer agreement rates and recurring exceptions for indicators of ineffective oversight or model weakness;
- ✓ verify that repeated override patterns trigger investigation, model review or re-validation where appropriate; and
- ✓ for high-risk or critical systems, perform or observe practical testing of review, approval, override, interruption and recovery mechanisms.

Auditors should not conclude that human oversight is effective solely because a manual-review procedure exists. The audit shall determine whether reviewers are technically capable, sufficiently informed and operationally empowered to challenge or override AI outputs when necessary.

11.4 Tier-Specific Applicability

Table 75. Decision-Review Requirements by Risk Tier

Risk Tier	Contestability and Human Decision Review Requirements
Tier 1 – Low	Defined system owner and basic review mechanism recommended for abnormal or disputed outputs. Basic output and intervention logging where applicable.
Tier 2 – Moderate	Defined human-review mechanism for material exceptions, anomalous outputs or predefined review triggers. Override capability and escalation procedure shall be available.
Tier 3 – High	Human review or approval shall be required for consequential decisions as defined by the risk assessment. Review criteria, explanation support, override, escalation and decision traceability shall be implemented and periodically tested.
Tier 4 – Critical	Enhanced supervisory review shall be implemented for critical decisions or actions. Approval, intervention, emergency override, safe-state/fallback and recovery mechanisms shall be independently tested. Review effectiveness shall be continuously monitored or assessed at defined intervals.

12. Audit Reporting & Remediation

This section defines the standards for classifying audit findings, structuring audit reports, developing corrective action plans, tracking remediation to closure, and integrating audit outcomes with the continual improvement mechanisms. The reporting framework ensures that audit results are actionable, traceable and drive measurable improvement in AI governance.

12.1 Finding Classification Framework

All audit findings shall be classified using a four-level severity scale. The severity determines remediation timelines, escalation requirements, and reporting obligations:

Table 26. Audit Finding Severity and Remediation Requirements

Severity	Definition	Remediation Requirements	Examples
Critical	Non-conformity or control failure creating an immediate or severe security, safety, operational or systemic risk, including active exploitation, significant model/data compromise, uncontrolled consequential actions or failure of essential protective controls.	Immediate escalation to designated senior management and prompt containment or restriction of the affected AI system/function based on risk. A detailed Corrective Action Plan shall be established on an expedited basis. External reporting shall be initiated only where the identified condition independently triggers an applicable external incident or reporting requirement.	Active discriminatory AI in production; unpatched critical vulnerability in Tier 4 system; AI decisions causing demonstrable harm; complete absence of governance for high-risk AI.
High	Significant non-conformity that materially weakens AI governance, creates substantial risk exposure, or violates a mandatory control requirement, but does not require immediate system shutdown.	High findings shall be monitored through the organisation's governance and risk-escalation process until effective closure. External reporting or competent-authority escalation shall occur only where the underlying condition independently triggers an applicable external requirement.	Missing independent validation for Tier 3 model; bias metrics exceeding thresholds without mitigation; absent CERT-In incident reporting process; non-functional human override mechanism.

Moderate	Partial non-conformity or control weakness that reduces the effectiveness of AI governance but does not create immediate high-risk exposure. The control exists but is incomplete or inconsistently applied.	Remediation plan within 30 business days. Implementation within 60 days. Tracked through standard audit follow-up. AI Governance Committee informed in aggregate reporting.	Incomplete model cards; data lineage gaps for non-sensitive datasets; fairness monitoring configured but not regularly reviewed; training records for human reviewers outdated.
Low / Observation	Minor non-conformity or improvement opportunity. The control is largely effective but could be enhanced. May also include best-practice recommendations beyondmandatory requirements.	Addressed within 3 months or incorporated into continuous improvement programme. No mandatory escalation. Tracked for trend analysis.	Documentation formatting inconsistencies; recommended (not mandatory) controls not implemented for Tier 1 systems; minor process efficiency improvements.

12.2 Finding Documentation Structure

Each audit finding shall be documented using the following standardised structure to ensure consistency, completeness, and actionability:

Table 27. Audit Finding Documentation Structure

Element	Description
Finding ID	Unique identifier (e.g., AISAG-2026-001) for tracking and cross-referencing.
Finding Title	Concise descriptive title summarising the non-conformity or observation.
Severity Classification	Critical, High, Moderate, or Low/Observation, using the finding classification framework.
Audit Domain Reference	The applicable core audit domain or supplementary audit requirement, identified by its section and specific control objective.
Standards Mapping	Corresponding ISO 42001 clause/Annex A control, NIST AI RMFmapping, and/or MITRE ATLAS mapping as applicable.
Condition (What Was Found)	Factual description of the current state observed during the audit, supported by specific evidence references.
Criteria (What Was Expected)	The applicable framework requirement, external requirement, standard or control objective against which the condition was assessed. Where an external requirement is cited, the finding should identify the issuing authority/source, applicable document or reference, relevant clause/control where available, and the basis for its applicability to the auditee or AI system.

Cause (Root Cause Analysis)	Analysis of why the non-conformity exists. Root cause identification enables systemic remediation rather than superficial fixes.
Effect (Risk / Impact)	Assessment of the actual or potential effect of the finding on confidentiality, integrity, availability, model/data integrity, safety, reliability, operational continuity, service delivery, financial exposure or other applicable risk objectives. Applicable external requirements may additionally be considered where relevant.
Recommendation	Specific, actionable recommendation for remediation, includingsuggested approach and priority.
Management Response	The auditee's response including agreement/disagreement with the finding, planned corrective actions, responsible owner, and target completion date.
Evidence References	Specific working paper references, document identifiers, system screenshots, or test result references supporting the finding.

12.3 Audit Report Structure

The final audit report shall follow a standardised structure to ensure completeness and comparability across audit engagements:

Table 28. Audit Report Structure

Report Section	Content Requirements
1. Executive Summary	High-level summary for senior leadership and governance bodies. Includes: audit objectives, scope, overall assurance opinion, summary of findings by severity, and key recommendations. Maximum 2 pages.
2. Audit Scope and Methodology	Description of: AI systems audited, risk tiers assessed, audit domains covered, methodology applied, standards referenced, audit team composition, and engagement timeline.
3. AI System Landscape Summary	Overview of the organisation's AI system portfolio as assessed during the audit: number of systems, risk tier distribution, model types, data categories, and third-party dependencies.
4. Detailed Findings	Complete finding documentation for all findings using the structure defined in Section 12.2, organised by audit domain. Each finding includes severity classification, root cause, impact assessment, and recommendation.
5. Finding Summary Matrix	Tabular summary of all findings: Finding ID, Domain, Severity, Title, Status, Owner, and Target Date. Enables at-a-glance tracking and prioritisation.
6. Management Response and Corrective Action Plans	Consolidated management responses for each finding, including agreed corrective actions, responsible owners, target completion dates, and any disagreements with findings (with auditor's rejoinder).

7. Positive Observations	Areas of strong practice identified during the audit. Recognition of effective controls, mature governance, and exemplary AI management practices. Positive observations build trust and incentivise compliance.
8. Overall Assurance Opinion	The auditor's overall opinion on the state of AI governance and compliance: Satisfactory, Satisfactory with Reservations, Requires Significant Improvement, or Unsatisfactory. Criteria for each opinion level defined below.

12.4 Assurance Opinion Criteria

The overall assurance opinion shall not be determined solely by the numerical count of findings. The auditor shall consider the materiality, systemic nature and severity of findings; effectiveness of compensating controls; affected AI-system risk tier; scope limitations; residual risk; recurrence of previous findings; and potential impact on security, safety, reliability and operations.

Table 79. Assurance Opinion Criteria

Opinion	Criteria
Satisfactory	Controls are appropriately designed and operating effectively across the material in-scope areas. No findings create material residual risk that would undermine the audit objective.
Satisfactory with Reservations	The control environment is generally effective, but one or more material weaknesses require remediation. Findings do not indicate fundamental or systemic failure of the overall control environment.
Requires Significant Improvement	Significant or systemic control weaknesses materially reduce assurance. Remediation is required across one or more important domains before a satisfactory level of assurance can be achieved.
Unsatisfactory	Critical or systemic control failures result in unacceptable residual risk or prevent the auditor from obtaining reasonable assurance over the in-scope AI systems. Immediate management action and targeted re-assessment are required.

12.5 Corrective Action Plans (CAPs)

For every finding classified as Critical, High, or Moderate, the auditee shall develop a Corrective Action Plan. The Corrective Action Plan is an approved remediation plan defining the actions, responsibilities, milestones, evidence and verification activities required to address the identified non-conformity.

Table 29. Corrective Action Plan Requirements

CAP Element	Requirement
Finding Reference	Link to the specific finding ID and severity classification.

Root Cause Acknowledgement	Management's acknowledgement of the root cause (or documented disagreement with rationale).
Corrective Action Description	Specific actions to be taken, including both immediate containment (for Critical/High) and long-term systemic correction.
Responsible Owner	Named individual with authority and accountability for implementing the corrective action. Not a team or department—a specific person.
Target Completion Date	Committed date for remediation completion, aligned with severity-driven timelines (Section 12.1).
Interim Milestones	For complex remediation: intermediate milestones with dates, enabling progress tracking.
Evidence of Completion	Definition of what evidence will demonstrate successful remediation (e.g., updated policy, test results, system configuration, retrained model).
Verification Method	How the auditor or governance function will verify that the corrective action has been effectively implemented (re-test, documentation review, demonstration).

12.6 Remediation Tracking and Closure

The audit lifecycle does not end with report issuance. Effective remediation tracking ensures that findings drive actual improvement:

Table 81. Remediation Tracking and Closure Mechanisms

Mechanism	Requirements
CAP Tracking Register	Centralised register maintained by the AI Governance Committee or internal audit function, tracking all open findings: Finding ID, severity, owner, target date, current status, and evidence of closure.
Progress Reporting	Monthly progress updates for Critical and High findings reported to the AI Governance Committee. Quarterly updates for Moderate findings. Low/Observations tracked at annual review.
Verification of Closure	Before a finding is closed, the auditor (or internal audit function for internal audits) shall verify that: the corrective action has been implemented as described, evidence of completion is obtained and reviewed, and the root cause has been addressed (not just the symptom).
Overdue Escalation	Findings not remediated within agreed timelines shall be escalated in accordance with the organisation's governance and risk-escalation process. Critical and High overdue findings shall be brought to appropriate senior-management or Board-level attention based on impact and risk. Escalation to an external competent authority shall occur only where specifically required by an applicable external requirement.
Re-Audit Triggers	An Unsatisfactory assurance opinion, or two or more Critical findings, triggers a mandatory re-audit within 3 months (or sooner if directed by the regulator). The re-audit scope covers all Critical and High findings plus their related domains.

12.7 External Reporting and Escalation Requirements

Applicable external requirements shall be identified during audit scoping and incorporated into the audit criteria and Statement of Applicability. This subsection addresses reporting or escalation obligations triggered by audit findings, cybersecurity incidents or other reportable events.

During audit planning and reporting, the auditor and auditee shall identify applicable external reporting requirements relevant to the system, organisation and identified finding or incident.

Where an audit identifies a condition that also constitutes a reportable cybersecurity incident or other externally reportable event, the applicable reporting mechanism and prescribed timeline shall be followed.

The audit report shall distinguish between:

- internal audit escalation;
- corrective-action reporting;
- security-incident reporting; and
- other applicable external reporting requirements.

The classification of an audit finding alone shall not automatically determine whether external reporting is required.

12.8 Reporting and Remediation Auditor Checklist

When conducting a follow-up or subsequent audit, the auditor shall verify:

> ✓ Previous audit findings have been formally tracked in a CAP register with assigned owners and target dates.
>
> ✓ Critical and High findings from the previous audit have been remediated and independently verified.
>
> ✓ Evidence of corrective action completion is documented and sufficient to demonstrate effective remediation.
>
> ✓ Root causes have been addressed (not just symptoms) for recurring or systemic findings.
>
> ✓ Management review (ISO 42001 Clause 9.3) has considered audit results and CAP status.
>
> ✓ Continual improvement actions arising from audit findings are documented and in progress.
>
> ✓ Overdue findings have been appropriately escalated per the framework's escalation requirements.
>
> ✓ Applicable external reporting and escalation obligations triggered by previous findings have been fulfilled.
>
> ✓ Audit report distribution has been controlled per confidentiality requirements.
>
> ✓ Trend analysis of findings across audit cycles has been performed and presented to the AI Governance Committee.

13. Maturity Level Model

This section defines a five-level organisational maturity model for AI governance capability. The model enables organisations to benchmark their current state, identify gaps and chart a structured progression path towards mature, resilient AI governance. It complements, but does not replace, the audit planning, control assessment, technical testing, decision review, reporting and remediation requirements in Sections 7–12.

The maturity model is informed by established capability maturity frameworks (CMMI, COBIT), adapted for the specific requirements of AI governance, and aligned with the nine core audit domains of this framework and the requirements of ISO/IEC 42001:2023.

13.1 Purpose and Use

The maturity model serves four distinct purposes within the framework:

Table 30. Purpose and Use of the Maturity Model

Purpose	Description
Self-Assessment	Organisations can use the model to conduct internal assessments of their AI governance capability, identifying strengths and areas for development prior to formal audit engagements.
Audit Context	Auditors may use maturity information to provide context for capability development and prioritisation of improvement activities. The maturity level shall not alter the severity classification of an audit finding, which shall be determined solely in accordance with Section 12.1.
Progression Planning	The model provides a clear roadmap for organisations to advance from ad hoc, reactive AI governance to optimised, continuously improving practices, with specific criteria for each progression step.
Regulatory Benchmarking	Sectoral regulators and CERT-In can use the model to establish minimum maturity expectations by sector or risk tier, and to monitor industry-wide governance trends over time.

13.2 Maturity Level Definitions

The model defines five progressive maturity levels. Each level represents a qualitative step-change in governance capability, not merely an incremental improvement:

Level 1: Initial / Ad Hoc

Table 83. Maturity Level 1 — Initial / Ad Hoc

Description	AI governance is absent, informal, or entirely reactive. AI systems are deployed without structured oversight, documentation, or risk assessment. Decisions about AI are made on a case-by-case basis without policy guidance. There is no awareness of AI-specific risks beyond general IT security.
Key Characteristics	No AI governance policy or designated AI ownership. No AI system inventory or risk classification. No structured bias testing, explainability, or monitoring. Incident response does not address AI-specific threats. Compliance with this framework is not yet initiated.

Level 2: Developing / Repeatable

Table 31. Maturity Level 2 — Developing / Repeatable

Description	Basic AI governance structures are established but not yet consistently applied across the organisation. Key policies exist but may not cover all AI systems or audit domains. Risk awareness is growing but processes are still maturing. Governance depends on individual effort rather than institutional capability.
Key Characteristics	AI governance policy exists but may be incomplete or inconsistently enforced. Basic AI system inventory maintained. Risk classification initiated for some AI systems. Some bias testing and documentation for customer-facing AI. CERT-In reporting process established. Human oversight defined for high-impact decisions.

Level 3: Defined / Standardised

Table 32. Maturity Level 3 — Defined / Standardised

Description	AI governance is formally defined, documented, standardised, and consistently applied across the organisation. Processes are institutionalised and do not depend on individual knowledge. All nine audit domains are addressed with documented controls. The organisation can demonstrate compliance with this framework's mandatory requirements for its risk tier.
Key Characteristics	Complete AI governance policy suite covering all domains. Comprehensive AI system inventory with risk tier classification. Formal AI Governance Committee operational. ISO 42001-aligned management system in place. Bias testing, explainability, and monitoring are standard practice. Independent model validation for Tier 3+ systems. Decision-review, human intervention and override mechanisms operational. Third-party governance structured.

Level 4: Managed / Measured

Table 33. Maturity Level 4 — Managed / Measured

Description	AI governance is not only standardised but actively measured, monitored, and managed using quantitative metrics. The organisation uses data-driven insights to optimise its AI governance practices. Controls are proactively strengthened based on trend analysis, threat intelligence, and performance data. Governance extends beyond compliance to strategic value creation.
Key Characteristics	Quantitative governance KPIs tracked and reported (fairness scores, drift rates, incident response times). Predictive risk identification using governance analytics. Cross-domain integration verifiedand optimised. Continuous monitoring with automated alerting. Regular benchmarking against industry peers and international standards. Board-level AI governance reporting with strategic insights.

Level 5: Optimising / Leading

Table 87. Maturity Level 5 — Optimising / Leading

Description	AI governance is a strategic organisational capability that is continuously optimised, innovative, and industry-leading. The organisation actively contributes to the evolution of AI governance practices, participates in standards development, and serves as a benchmark for others. Governance is embedded in organisational culture, not just processes.
Key Characteristics	Continuous improvement cycle demonstrably drives governance evolution. Organisation contributes to national/international AI governance standards. Advanced capabilities: automated compliance verification, AI-driven governance analytics, real-time risk dashboards. Governance innovation: novel approaches to emerging AI risks (agentic AI, quantum ML). External recognition as governance leader. Active participation in CERT-In/industry governance forums.

13.3 Domain-Specific Maturity Indicators

The following matrix provides specific maturity indicators for each of the nine core audit domains across all five levels. This enables granular assessment and targeted improvement planning:

Table 34. Domain-Specific Maturity Indicators

Domain	L1 Initial	L2 Developing	L3 Defined	L4 Managed	L5 Optimising

8.1 Governance & Accountability	No AI policy. No designated ownership.	Basic policy exists. AI owner assigned for some systems.	Comprehensive policy. AI Governance Committee. Full RACI. AIIA for all Tier 2+.	Governance KPIs tracked. Policy effectiveness measured. Board reporting with insights.	Governance innovation. Industry benchmarking. Standards contribution.
8.2 Data Lifecycle & Privacy	No data provenance.	Basic lineage for key datasets.	Full provenance documentation. ISO 25012 quality assessment. Bias screening standard.	Automated data quality monitoring. Drift-aware data governance.	Predictive data quality. Novel privacy-enhancing techniques. Data governance thought leadership.
8.3 Model Risk Management	No model registry. No version control.	Basic registry for production models. Manual version tracking.	Complete registry. Automated version control. Independent validation. Decommissioning protocol.	Quantitative MRM metrics. Aggregate model risk reporting. Automated validation pipelines.	AI-driven model risk prediction. Self-healing model pipelines. MRM framework published.
8.4 Bias & Fairness	No bias testing.	Basic bias metrics for some models.	Comprehensive fairness suite. Proxy screening. Ongoing monitoring dashboards.	Fairness metrics integrated into CI/CD. Counterfactual testing automated. Trend analysis.	Sector-leading fairness research.
8.5 Explainability	No explainabilitycapability.	Basic model documentation. Ad hoc explanations on request.	Appropriate explainability techniques implemented and validated; model documentation and decision traceability established.	Explanation quality measured. Consistency verified. Real-time explainability.	Novel XAI techniques. Explainability research contribution. Advanced and independently validated explanation capability.

8.6 Security &Resilience	Standard IT security only. No AI-specific controls.	AI threat awareness. Basic CERT-In reporting. Annual VAPT.	AI threat model. Adversarial testing. Supply chain verification. AI SOC integration.	Continuous security monitoring. Automated adversarial testing. Red team programme.	AI security research. Zero-day AI threat response. Security framework contribution .
8.7 Continuous Monitoring	No production monitoring.	Basic performance tracking. Manual drift checks.	Automated drift detection. Fairness dashboards. Defined re-validation triggers.	Real-time monitoring with predictive alerts. Automated retraining governance.	Self-optimizing monitoring. Novel drift detection. Monitoring thought leadership.
8.8 Human Oversight	No defined oversight model.	HITL for some high-impact decisions. Override capability exists.	HITL/HOTL defined per tier. Reviewer training. Decision-review, intervention and override mechanisms operational and tested.	Override effectiveness measured. Reviewer competence assessed. Decision accountability analytics.	Optimised human-AI collaboration. Novel oversight models. Oversight research contribution .
8.9 Third-Party and AI Supply-Chain Assurance	No vendor AI due diligence.	Basic vendor assessment for key AI providers. Standard contracts.	Structured due diligence. Audit rights in SLAs. SBOM maintained. Structured supplier assurance, AIBOM/SBOM, security requirements and dependency monitoring.	Continuous vendor monitoring. AI supply chain analytics. Vendor risk scoring.	Supply chain innovation. Vendor ecosystem governance leadership. Cross-industry sharing.

13.4 Maturity Assessment Methodology

Maturity assessment may be conducted as a standalone exercise (self-assessment) or as part of a formal audit engagement. The methodology follows these steps:

Table 835. Maturity Assessment Methodology

Step	Methodology
Step 1: Domain-Level Scoring	Assess each of the nine core audit domains independently against the maturity indicators in Section 13.3. Assign a maturity level (1–5) to each domain based on the organisation's demonstrated capability. Evidence-based assessment: the assigned level must be supportable by documentary, observational, or analytical evidence.
Step 2: Overall Maturity Determination	The organisation's overall maturity level is determined by the lowest-scoring domain (weakest-link principle). Rationale: a chain of governance is only as strong as its weakest link. An organisation with Level 4 capability across eight domains but Level 1 in security cannot be considered Level 4 overall.
Step 3: Gap Analysis	For each domain, identify the specific gaps between the current maturity level and the target level. The target maturity level should consider the organisation's AI risk profile, applicable external requirements and improvement objectives
Step 4: Improvement Roadmap	Develop a prioritised improvement roadmap addressing the gaps identified in Step 3. Prioritisation criteria: risk impact (domains affecting the highest-tier AI systems first), regulatory urgency, implementation feasibility, and resource requirements. The roadmap should include specific actions, responsible owners, milestones, and target dates.
Step 5: Reassessment Cadence	Maturity assessments may be conducted annually as part of the AI governance review, after major organisational changes (mergers, new AI deployments or regulatory shifts), and during formal audits where an independent maturity assessment accompanies the assurance opinion

14. Referenced Frameworks & Standards

This section catalogues the international standards, regulatory frameworks, Indian legislation, and technical security resources that form the normative and informative foundation of the AI Security Auditing Guidelines. These references ensure global credibility, compliance interoperability, and alignment with both Indian regulatory requirements and international best practices.

The standards, frameworks, advisories and technical references identified in this section provide normative or informative inputs to the AI Security Auditing Guidelines. Audit criteria shall primarily comprise the controls defined in these Guidelines together with applicable external requirements identified for the engagement. Informative frameworks and good-practice references shall be used to support risk assessment, control design, threat modelling, technical testing and auditor judgement and shall not, solely by virtue of being referenced in this section, be considered mandatory requirements.

14.1 Purpose of Referenced Frameworks

The inclusion of referenced frameworks serves five strategic functions:

Reference Alignment: Providing recognised standards, frameworks, threat taxonomies and technical guidance against which framework controls may be mapped. A referenced document shall be treated as mandatory only where its applicability is established through these Guidelines or an applicable external requirement.

Interoperability: Enabling Indian organisations to achieve cross-border regulatory recognition and international AI certification readiness.

Audit credibility: Providing auditors with externally validated control objectives, risk categories, and assessment methodologies.

Gap coverage: Ensuring no critical governance dimension is overlooked by synthesising complementary frameworks that address different aspects of AI risk.

Future-proofing: Anchoring the framework to evolving international standards (ISO, NIST, EU) whose update cycles will inform periodic revisions of this document.

14.2 International AI Governance Standards

The following international standards and frameworks provide the primary governance architecture upon which this framework's audit requirements are built:

14.2.1 ISO/IEC 42001:2023 – AI Management Systems

Table 36. ISO/IEC 42001:2023 Reference Profile

Attribute	Detail
Authority	International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC).
Status	Normative – Primary alignment standard for this framework.
Relevance	The first international management system standard dedicated to AI. Establishes requirements for organisations providing or using AI-based products or services to ensure responsible development, provision, and use of AI systems.
Core Structure	Follows the ISO Harmonized Structure (Clauses 4–10): Context of the Organisation, Leadership, Planning, Support, Operation, Performance Evaluation, Improvement. Annex A provides 38 AI-specific controls.
Audit Integration	ISO 42001 clauses and Annex A controls are mapped throughout Sections 7, 8, and 12 of this framework documents.
Key Clauses forAuditors	Clause 4 (Context of the organisation), Clause 5 (Leadership), Clause 6 (Planning, including AI risk and impact assessment and treatment), Clause 7 (Support, including competence and awareness), Clause 8 (Operation, including AI system lifecycle processes), Clause 9 (Performance evaluation, including internal audit and management review), Clause 10 (Improvement, including nonconformity and corrective action).
Annex A ControlFamilies	A.2 Policies related to AI; A.3 Internal organization; A.4 Resources for AI systems; A.5 Assessing impacts of AI systems; A.6 AI system life cycle; A.7 Data for AI systems; A.8 Information for interested parties; A.9 Use of AI systems; A.10 Third-party and customer relationships.

14.2.2 ISO 23894:2023 – AI Risk Management

Table 91. ISO 23894:2023 Reference Profile

Attribute	Detail
Authority	ISO
Status	Informative – Cross-referenced for risk management process guidance
Relevance	Provides guidance on managing risks specific to AI systems, building upon the ISO 31000 risk management framework. Addresses AI-specific risk sources including data quality, model behaviour, societal impact, and operational context.
Integration Approach	Rather than a standalone audit domain, ISO 23894 is integrated as a cross-reference throughout the framework: its risk assessment process informs Section 5 (Risk Tiering), its treatment guidance supports Section 7 (Audit Planning & Scoping Methodology), and its monitoring principles align with Section 8.7 (Continuous Monitoring).

Key Contribution	Establishes that AI risk management is a process standard, not a classification standard—complementing this framework's tier-based classification with a structured approach to identifying, analysing, evaluating, and treating AI-specific risks.

14.2.3 NIST AI Risk Management Framework (AI RMF 1.0)

Table 37. NIST AI RMF 1.0 Reference Profile

Attribute	Detail
Authority	U.S. National Institute of Standards and Technology (NIST)
Status	Informative – Key reference for risk management functions andtrustworthy AI characteristics
Relevance	A voluntary, risk-based framework for trustworthy AI, structured around four core functions: Govern, Map, Measure, and Manage. Widely adopted internationally as a practical operationalisation guide for AI governance.
Core Functions	GOVERN: Establish AI governance structures, policies, and accountability. MAP: Contextualise AI risks based on system characteristics and deployment. MEASURE: Assess and quantify AI risksusing metrics and testing. MANAGE: Prioritise, respond to, and monitor identified AI risks.
Trustworthy AI Characteristics	Valid and Reliable, Safe, Secure and Resilient, Accountable and Transparent, Explainable and Interpretable, Privacy-Enhanced, Fair with Harmful Bias Managed.
Audit Integration	NIST AI RMF function mappings are provided throughout Section 8 (Core Audit Domains) and, for technical controls, Section 10 (Cybersecurity & Resilience).

14.2.4 ISO/IEC 42005:2025 – AI System Impact Assessment

Table 38. ISO/IEC 42005:2025 Reference Profile

Attribute	Detail
Authority	ISO/IEC
Status	Informative – AI impact-assessment reference
Relevance	Provides guidance for conducting AI system impact assessments across the AI lifecycle, including identification, evaluation and documentation of potential impacts associated with intended and foreseeable use of AI systems.
Key Provisions	Supports Section 5 risk classification, Section 8.1 AI Risk and Impact Assessment, material-change reassessment and pre-deployment assurance.

Key Audit Use	Auditors may use the standard to evaluate whether AI impact assessments are appropriately scoped, documented, periodically reviewed and updated following material changes.

14.2.5 EU AI Act (Regulation 2024/1689)

Table 39. EU AI Act Reference Profile

Attribute	Detail
Authority	European Parliament and Council of the European Union
Status	Informative – Referenced for risk classification model and conformity assessment principles
Relevance	The world's first comprehensive AI-specific legislation, establishing a risk-based regulatory framework with four categories: Unacceptable Risk (prohibited), High Risk (mandatory conformity assessment), Limited Risk (transparency obligations), and Minimal Risk (voluntary codes of practice).
Key Provisions	Mandatory conformity assessments for high-risk AI in healthcare, law enforcement, credit scoring, and education. Requirements for technical documentation, logging, human oversight, accuracy, robustness, and cybersecurity. Obligations for providers, deployers, importers, and distributors of AI systems.
Alignment with This Framework	The EU AI Act's risk classification model informed the four-tier structure in Section 5. Its conformity assessment requirements parallel the audit intensity calibration in Section 5.5. Indian organisations seeking EU market access will benefit from dual compliance under this framework.

14.3 AI Security and Adversarial Risk Frameworks

Table 40. AI Security, Adversarial Risk and Audit Reference Frameworks

The following frameworks provide the technical security threat intelligence, attack taxonomy, and vulnerability assessment methodologies used in the audit controls throughout this framework:

Framework	Authority	Status	Relevance	Integration Point

MITRE ATLAS	MITRE Corporation	Informative	Adversarial Threat Landscape for Artificial Systems – a curated knowledge base of adversarial tactics, techniques, and real-world case studies targeting AI/ML systems. Organised by attack stage (reconnaissance through impact), providing auditors with a structured threat model for AI-specific security assessments.	Section 8.6 (Security), Section 10.4 (AI-Specific VAPT) – all MITRE ATLAS mappings.
OWASP Machine Learning Top 10	OWASP Foundation	Informative	Identification of the ten most critical security risks in machine learning systems, including data poisoning, model evasion, model theft, adversarial examples, and ML supply chain vulnerabilities.	Section 8.6 (Security), Section 9 (Advanced AI Security Requirements), Section 10 (Cybersecurity)
OWASP GenAI LLM Top 10 – 2026	OWASP Foundation	Informative	Dedicated security risk taxonomy for large language model applications, covering prompt injection (direct and indirect), Sensitive Information Disclosure, Excessive Agency, Supply Chain, Data and Model Poisoning, Unbounded Consumption, Misinformation, Hidden Context Exposure, Vector and Embedding Weaknesses, Improper Output Handling.	Section 9
NIST SP 800-53 Rev. 5	NIST	Informative	Security and Privacy Controls for Information Systems and Organizations. While not AI-specific, provides foundational information security controls (access control, audit and accountability, system integrity, risk assessment) applicable to AI infrastructure	Section 10 (Cybersecurity)
ISO/IEC 27001:2022	ISO/IEC	Informative	Information security management system standard. Organisations with existing ISO 27001 certification can extend their ISMS to cover AI-specific controls, creating	Section 10 (Cybersecurity), Section 7 (Audit Planning)

			an integrated governance approach with ISO 42001.	
ISACA AI Audit Framework	ISACA	Informative	Professional guidance for auditing AI systems, including risk assessment approaches, control objectives, and audit procedures specific to AI governance, data management, and model lifecycle.	Section 12 (Audit Reporting & Remediation)
IIA Artificial Intelligence Auditing Framework	Institute of Internal Auditors	Informative	Internal audit guidance for AI systems covering governance, strategy alignment, data management, model development, deployment controls, and ongoing monitoring from an internal audit perspective.	Section 7 (Audit Planning), Section13 (Maturity Model)
NIST AI 100-2e2025 – Adversarial Machine Learning	NIST	Informative	Provides taxonomy and terminology for adversarial machine-learning attacks and mitigations, considering ML methods, lifecycle attack stages, attacker objectives, capabilities and knowledge.	Sections 8.6, 9 and 10.4
OWASP Top 10 for Agentic Applications 2026	OWASP GenAI Security Project	Informative	Provides a dedicated security-risk taxonomy for autonomous and agentic AI applications, including goal hijacking, tool misuse, identity/privilege abuse, supply-chain risk, memory/context poisoning, inter-agent communication and cascading failures.	Sections 9.7, 9.9, 10.4 and 10.5

14.4 CERT-In Advisories, Guidelines and Technical References Relevant to AI Security

CERT-In periodically issues cybersecurity advisories, guidelines, technical references, vulnerability notes and implementation guidance relevant to emerging cyber threats and

technologies.

During an AI security audit, the auditor shall identify applicable CERT-In publications based on the AI system's technology stack, deployment architecture, attack surface, threat exposure and supporting infrastructure.

Table 41. CERT-In Publications Relevant to AI Security Audits

Advisories & Guidelines	Purpose and applicability	Audit Check
CERT-In Advisory CIAD-2026-0020: Defending Against Frontier AI Driven Cyber Risks	Provides awareness and defensive measures against AI-enabled and accelerated cyber threats, including automated vulnerability discovery and exploitation. Applicable to organisations operating digital infrastructure and AI-enabled systems	Verify inclusion of frontier-AI threats in risk assessment/threat modelling, vulnerability assessment, monitoring, red-team exercises and incident-response procedures
Blueprint for Reducing Exposure and Defending against AI-Assisted Vulnerabilities Exploitation in Digital Infrastructure	Provides a risk-based security blueprint to reduce attack surface and strengthen resilience against AI-assisted exploitation across enterprise, cloud, applications, APIs, AI systems and OT environments.	Verify asset/exposure management, vulnerability and patch management, access controls, segmentation, SOC/SIEM monitoring, supply-chain security, incident response and periodic security testing against applicable Blueprint controls.
Guidelines regarding AI-Accelerated Vulnerability Protection and Response Requirements for Original Equipment Manufacturers (OEMs), and Technology Providers	Establishes enhanced requirements for OEMs and technology providers to identify, disclose, remediate and respond to vulnerabilities at the speed required in an AI-accelerated threat environment.	Verify continuous vulnerability assessment, secure development practices, vulnerability disclosure, patch/remediation timelines, interim mitigations, SBOM, incident response and supporting security-assurance evidence from the OEM/provider.
Technical Guidelines on SBOM, QBOM & CBOM, AIBOM and HBOM Version 2.0	Provides technical guidance for maintaining structured bills of materials covering software, cryptographic, AI and hardware components to improve asset/component visibility and supply-chain assurance.	Verify that applicable AIBOM/SBOM records identify material models, software components, datasets, dependencies and relevant metadata and that the inventories support vulnerability, dependency and supply-chain assessment.

Comprehensive Cyber Security Audit Policy Guidelines	Provides CERT-In cybersecurity audit-policy guidance relevant to audit planning, responsibilities, conduct, evidence and reporting by cybersecurity auditing organisations.	Consider applicable audit-policy requirements during engagement planning, evidence handling, fieldwork, audit reporting and quality assurance.

14.5 Applicability of External Requirements

External requirements relevant to an AI security audit may include applicable CERT-In Directions; CERT-In advisories, guidelines and technical references; sector-specific cybersecurity requirements; contractual security and assurance obligations; formally adopted AI management or cybersecurity standards; and other requirements identified by the auditee or competent authority.

Mandatory requirements shall be incorporated into the audit criteria where applicable. Informative standards, advisories and good-practice references shall support risk assessment, threat modelling, control assessment and technical testing based on their relevance to the in-scope AI system.

Identification and mapping of applicable external requirements during an audit engagement shall be performed in accordance with Section 7.3.

15. Integration with the India AI Governance Guidelines

In 2025, the Ministry of Electronics and Information Technology (MeitY) published the India AI Governance Guidelines—a national policy framework developed by a drafting committee constituted by MeitY, with a foreword by the Secretary, MeitY, and a preface by the Principal Scientific Adviser to the Government of India. The Guidelines represent India's overarching policy position on AI governance: a balanced, agile, flexible, and pro-innovation approach structured around seven guiding principles ("sutras"), six policy pillars, a phased action plan, and practical guidance for industry and regulators.

This section maps the India AI Governance Guidelines to the National AI Security Auditing Guidelines, demonstrating how the two instruments complement each other and ensuring that organisations adopting this audit framework are simultaneously aligned with the national governance vision.

15.1 Relationship Between the Two Instruments

The India AI Governance Guidelines and this audit framework operate at different but complementary levels of governance:

Table 42. Roles of the National Governance and Audit Guidelines

Instrument	Role in AI Governance
India AI Governance Guidelines	Sets the national AI governance philosophy, policy direction, and institutional architecture. Provides high-level principles, recommends policy and regulatory approaches, and establishes institutional bodies (AI Governance Group, AI Safety Institute). Scope is economy-wide and covers enablement, regulation, risk mitigation, and institutional design.
AI Security Auditing Guidelines (This Framework)	Operationalises the governance vision into auditable controls. Provides the technical audit methodology, risk tiering, domain-specific control requirements, and audit & assurance mechanisms that translate governance principles into verifiable compliance. Scope is audit-specific, targeting organisations deploying AI systems.

In essence, the India AI Governance Guidelines answer the question "what should India's AI governance look like?" while this audit framework answers "how do we verify that organisations are actually doing it?" The two instruments are designed to be mutually reinforcing: compliance with this audit framework demonstrates alignment with the national governance vision, and the national governance vision provides the policy authority for the audit requirements.

15.2 Alignment of Seven Sutras with Framework Principles

The India AI Governance Guidelines are anchored on seven guiding principles or "sutras," adapted from the RBI's FREE-AI Committee report for cross-sectoral applicability. This framework's seven guiding principles are directly aligned with these sutras. The following table maps each sutra to the corresponding framework principle and audit operationalisation:

Table 43. Seven-Sutra Alignment with Audit Principles

India AI Sutra	Sutra Description	Framework Principle	Audit Operationalisation
Sutra 1: Trust is the Foundation	Trust must be embedded across the AI value chain—in the technology, the organisations building it, the institutions supervising it, and the individuals using it.	**All seven framework principles collectively**	The entire audit framework operationalises trust through verifiable controls. Trust is not a standalone principle but the cumulative outcome of security, accountability, transparency, fairness, privacy, auditability and resilience.
Sutra 2: People First	Human-centric AI design, human oversight, human empowerment, and capacity development.	**P2: Accountability & Human Oversight**	Section 8.8 (Human Oversight): HITL/HOTL/HOVL requirements by tier. Section 11 (Contestability and Human Decision Review): mechanisms for review, validation, intervention and override of AI-generated decisions. Section 8.1: AI Governance Committee with cross-functional human accountability.
Sutra 3: Innovation over Restraint	Responsible innovation should be prioritised over cautionary restraint, all other things being equal.	**P7: Resilience and Continuous Assurance**	Section 5 (Risk Tiering): proportionate obligations by risk tier. Section 5.7: MSME proportionality provisions.
Sutra 4: Fairness & Equity	AI systems should be fair, unbiased, and must not discriminate, especially against marginalised communities.	**P4: Fairness and Bias Management**	Section 8.4 (Ethics, Bias & Fairness): quantitative bias metrics (disparate impact, equalised odds), proxy variable screening, ongoing fairness monitoring, counterfactual testing.

Sutra 5: Accountability	Clear allocation of responsibility and enforcement of regulations based on function performed and risk level.	**P2: Accountability & Human Oversight; P6: Auditability & Traceability**	Section 8.1: RACI matrix, AI governance policy, AIIA. Section 12: finding classification, CAP, remediation tracking.
Sutra 6: Understandab leby Design	Disclosures and explanations that can be understood by intended users and regulators, tothe extent technically feasible.	**P3: Transparency & Explainability**	Section 8.5 (Explainability): XAI techniques, model cards, user-facing plain-language explanations, decision logging, appropriate explainability techniques, model documentation, decision traceability and audience-appropriate explanations.
Sutra 7: Safety, Resilience & Sustainability	Safe, secure, and robust systems that can withstand shocks, with environmental sustainability considerations.	**P1: Security-by-design**	Section 8.6 (Security): adversarial testing, CERT-In incident reporting, VAPT. Section 10 (Cybersecurity): infrastructure hardening, SOC integration, business continuity. Section 8.7 (Monitoring): drift detection, re-validation triggers.

15.3 Alignment of Six Policy Pillars with Framework Sections

The India AI Governance Guidelines structure their recommendations across six policy pillars spanning enablement, regulation, and oversight. The following table maps each pillar to the corresponding sections of this audit framework:

Table 99. Policy-Pillar Alignment with the Audit Framework

Policy Pillar	Guidelines Recommendation	Framework Sections	Integration Analysis
Pillar 1: Infrastructure	Expand access to data, compute, and digital public infrastructure (DPI) for AI development and adoption.	**Section 8.2 (Data Lifecycle): data quality, provenance, and governance controls.**	The audit framework does not mandate infrastructure provision (that is a policy function) but ensures that AI systems built on shared infrastructure meet governance standards. These data-quality controls directly support the Guidelines' emphasis on high-quality, representative datasets.

Pillar 2: Capacity Building	Education, skilling, and training to buildtrust and increase AI adoption.	**Section 7.6: auditor competence requirements.** **Section 8.8: human reviewer training.**	The audit framework operationalises capacity building through mandatory competence requirements for auditors and human reviewers.
Pillar 3: Policy& Regulation	Balanced, agile frameworks that support innovation and mitigate risk. Review current laws and address gaps.	**Section 14 (Referenced Frameworks): regulatory landscape mapping.**	The audit framework provides the auditable control layer aligned with applicable policy and regulatory requirements.
Pillar 4: Risk Mitigation	India-specific risk assessment framework based on evidence of harm. Voluntary measures, techno-legal approaches, and incident reporting.	**Section 5 (Risk Tiering): India-specific four-tier risk classification with eight-factor scoring. Section 9 (Advanced AI Security Requirements): threat landscape. Section 8.6: incident reporting.**	The audit framework's risk tiering system directly delivers the "India-specific risk assessment framework" recommended by the Guidelines. The incident reportingarchitecture aligns with the Guidelines' call for a national AI incident database, leveraging CERT-In's existing infrastructure.
Pillar 5: Accountabilit y	Graded liability based on function, risk, and due diligence. Transparency about AI value chain operations.	**Section 8.1 (Governance): accountability structures. Section 11 (Contestability and Human Decision Review): review and accountability mechanisms for consequential AI outputs. Section 12 (Reporting): assurance opinions. Section 8.9 (Third-Party): value chain transparency.**	The audit framework implements graded accountability through risk-tiered audit intensity (Section 5.5) and through the finding severity classification (Section 12.1) which calibrates consequences to the level of non-compliance.

Pillar 6: Institutions	Whole-of-government approach with AI Governance Group (AIGG), Technology & Policy Expert Committee (TPEC),and AI Safety Institute (AISI).	**Sections 7.6 (Auditor Competence) and 7.7 (Auditor Independence)**	The audit framework establishes competence and independence requirements for auditors and is designed to support integration with the institutional mechanisms envisaged under the India AI Governance Guideline.

16. Acknowledgement

This document has been prepared as part of the activities of the Working Group on "**Indian Standards for Cybersecurity Auditing**", constituted with the objective of developing a comprehensive and standardised framework for conducting cybersecurity audits tailored to Indian regulatory requirements and the evolving needs of the national cybersecurity ecosystem.

The Indian Computer Emergency Response Team (CERT-In) acknowledges the mentorship and guidance of Dr Sanjay Bahl, Director General, CERT-In, Sh. S.S. Sarma, CERT-In & Sh. Ashutosh Bahuguna, CERT-In and valuable contributions, expertise and insights provided by the following members of the Working Group on "Indian Standards for Cybersecurity Auditing" during the review, deliberation and development of the document titled "**AI Security Auditing Guidelines**":

i. Dr Naveen Dham, G-Info Technology Solutions Private Limited
ii. Ms. Vaishali Mutalik, Shieldbyte Infosec Private Limited
iii. Dr Swati Vashisth, Dr CBS Cyber Security Services LLP
iv. Sh. Sahil Pahwa, The Intect (Hion Security Private Limited)
v. Sh. Abhishek Solanki; Sh. Shashank Gupta; Sh. Mohit Kataria, CERT-In