



# AI Governance Framework

for India 2025-26

A National Policy Report



# AI Governance Framework for India

Whitepaper for the National Cyber and AI Center (NCAIC)

Published: September 1, 2025 | [www.ncaic.in](http://www.ncaic.in)







# Executive Summary

India stands at a critical inflection point in its artificial intelligence journey. The IndiaAI Mission has successfully unlocked national compute infrastructure, established comprehensive datasets, and created a dedicated safety institute. Simultaneously, the Data Protection and Digital Privacy (DPDP) Act and CERT-In Directions are fundamentally reshaping how organizations approach privacy governance and incident management. However, the rapid pace of AI deployment across sectors is significantly outpacing the development of adequate control mechanisms.

This comprehensive whitepaper presents an actionable, risk-based AI Governance Framework specifically tailored to India's unique legal-regulatory environment and diverse sectoral realities. The framework synthesizes global best practices with practical implementation blueprints designed for ministries, regulatory bodies, public sector undertakings (PSUs), and large enterprises operating within the Indian context.

**Risk Classification**  
Common language to classify AI risk from prohibited to high-risk to low-risk use cases, aligned with India's public interest priorities

**Lifecycle Controls**  
Complete control set covering data, model, application, and operations with security, privacy, and safety engineered by design

**Assurance System**  
Comprehensive audits, evaluations, and attestations mapped to international standards including ISO 42001 and NIST AI RMF

**Implementation Roadmaps**  
Detailed 100-day, 12-month, and 24-month plans with templates and checklists for accelerated adoption

# The Imperative for AI Governance

The scale and stakes of AI deployment in India demand immediate attention. Artificial intelligence systems are now deeply embedded in critical national infrastructure including digital payments, healthcare triage systems, citizen service delivery platforms, agricultural decision support systems, and educational assessment tools. When these systems fail or operate with bias, the consequences can harm individuals at population scale, undermining public trust and democratic institutions.

The Data Protection and Digital Privacy Act has elevated the importance of consent mechanisms, purpose limitation principles, and fiduciary duties in data processing. AI systems must operationalize these foundational principles from design through deployment, rather than treating them as afterthoughts or compliance checkboxes. This requires sophisticated technical controls, clear governance processes, and continuous monitoring capabilities.



## Population-Scale Impact

AI failures in payments, healthcare, and citizen services can harm millions simultaneously, requiring robust safeguards and rapid response capabilities.



## Trust and Security

CERT-In's requirements for 6-hour incident reporting and supply chain oversight intersect directly with AI model operations and security posture.



## Global Competitiveness

ISO 42001 adoption and third-party assurance unlock cross-border trust, international procurement opportunities, and export markets for Indian AI solutions.

Election integrity and the proliferation of deepfakes present immediate challenges requiring coordinated response mechanisms. Content labeling, provenance tracking, and incident response playbooks must be implemented across platforms, public bodies, and enterprise systems to maintain democratic processes and public discourse integrity.

# Foundational Design Principles

The AI Governance Framework is built upon eight foundational principles that reflect India's constitutional values, technological capabilities, and regulatory requirements. These principles provide the philosophical and operational foundation for all governance mechanisms, technical controls, and implementation processes.



## Human-Centric Approach

Human oversight for consequential decisions, meaningful opt-out pathways, and accessible appeal mechanisms ensure that AI systems serve human welfare and preserve individual agency.



## Risk Proportionality

Controls scale appropriately with potential impact through clear prohibited, high-risk, medium-risk, and low-risk categories with corresponding governance requirements.



## Privacy by Design

Consent mechanisms, purpose limitation, data minimization, retention controls, and verifiable deletion capabilities embedded throughout the AI lifecycle.



## Security by Design

Adversarial robustness, prompt injection defenses, secrets isolation, and immutable audit trails protect against sophisticated threats and attacks.

## Transparency and Explainability

Model cards, data sheets, decision rationales, and content provenance metadata enable understanding and accountability for AI system behavior and outputs.

## Inclusivity and Fairness

Comprehensive bias testing across Indian demographics, languages, and cultural contexts ensures equitable outcomes and representation.

## Accountability and Traceability

Named ownership, documented approvals, and immutable logging throughout the AI supply chain enable clear responsibility and forensic analysis.

## Continuous Assurance

Pre-deployment testing and ongoing production evaluation for safety, privacy, and performance ensure sustained compliance and effectiveness.



# Regulatory Landscape Alignment

The AI Governance Framework operates within India's evolving regulatory ecosystem, ensuring seamless integration with existing legal requirements and institutional mandates. This alignment is crucial for practical implementation and regulatory compliance across multiple jurisdictions and sectors.

The Data Protection and Digital Privacy Act 2023, combined with the Draft DPDP Rules 2025, establishes fundamental requirements for lawful processing bases, enhanced duties for Significant Data Fiduciaries, mandatory breach notifications, and Data Protection Impact Assessment-like procedures. AI systems must demonstrate compliance with these requirements through technical controls and governance processes.



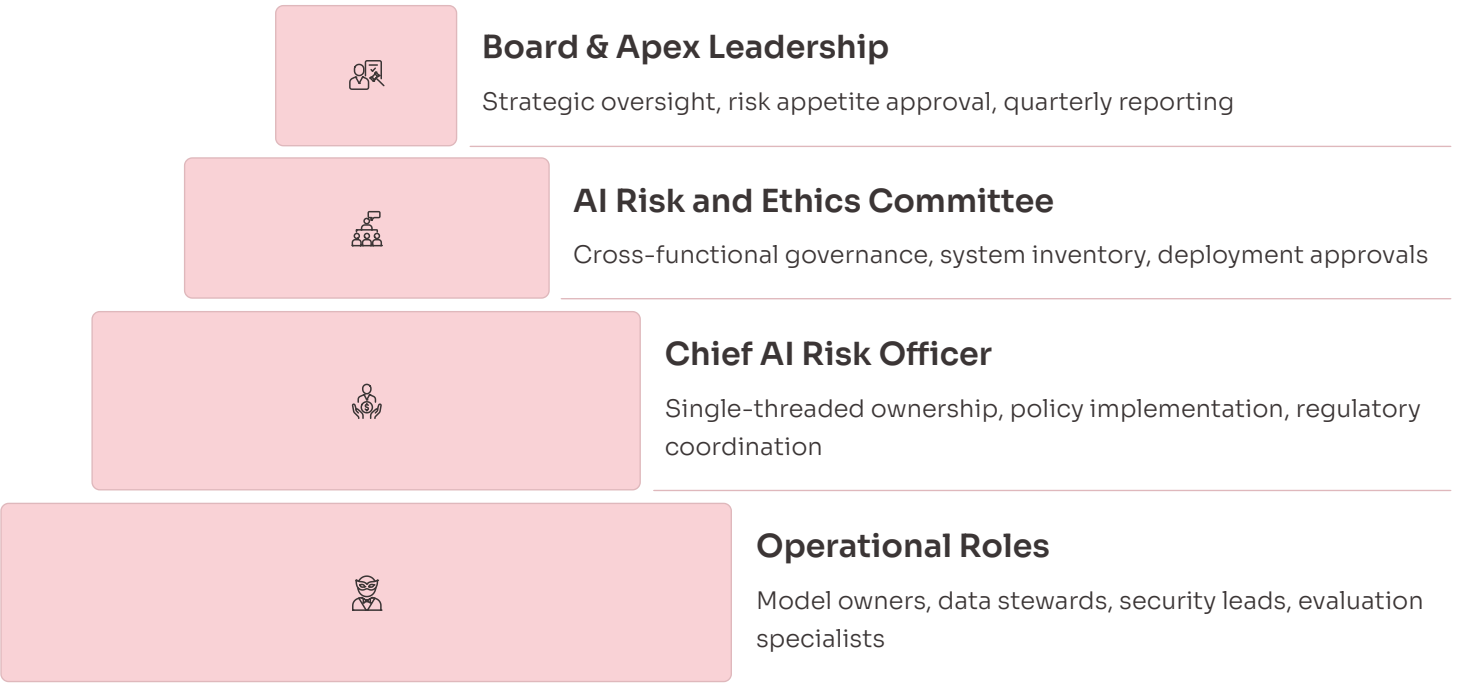
International standards including ISO/IEC 42001 for AI Management Systems, ISO/IEC 23894 for AI risk management, NIST AI Risk Management Framework 1.0, and taxonomical approaches from the EU AI Act inform the framework while adapting to India's specific constitutional, legal, and operational context.



# Governance Model and Organizational Structure

Effective AI governance requires clear organizational structures, defined roles, and explicit accountability mechanisms. The governance model establishes a multi-tiered approach that spans from board-level oversight to operational implementation, ensuring appropriate governance at each organizational level.

Board and apex-level accountability includes approval of AI risk appetite statements, prohibited use case definitions, and high-risk approval criteria. The board receives quarterly AI risk and incident reports, maintaining strategic oversight while delegating operational decision-making to specialized committees and roles.



The AI Risk and Ethics Committee (AIREC) serves as the central coordination body, comprising cross-functional representation from security, privacy, legal, risk, product, operations, compliance, public policy, and domain expertise. AIREC maintains the authoritative AI System Inventory, oversees risk classification processes, manages exceptions and waivers, and provides final approval for high-risk system deployments to production environments.

Activity	Responsible	Accountable	Consulted
System Inventory	Model Owner	CARO	Data Steward
Risk Classification	Eval Lead	AIREC	Legal/Privacy
Production Approval	AIREC	Board Delegate	Security, Legal
Incident Reporting	Security Lead	CARO	Legal/Privacy/PR



# Risk Classification Taxonomy

A clear risk classification taxonomy enables organizations to apply appropriate controls, governance processes, and assurance mechanisms based on the potential impact and societal consequences of AI system deployment. This taxonomy provides specific examples relevant to the Indian context while establishing clear boundaries for different risk categories.

Prohibited AI systems represent use cases that are fundamentally incompatible with constitutional rights, democratic principles, or established legal protections. These systems are banned regardless of technical safeguards or governance mechanisms, reflecting societal values and legal requirements that cannot be overridden by risk mitigation measures.

## Prohibited AI Systems

- Social scoring of citizens for public benefit access
- Biometric categorization by sensitive personal attributes
- Emotion inference for employment, education, or credit decisions
- Subliminal manipulation techniques targeting vulnerable populations

High-risk AI systems operate in domains where failures can cause significant individual or societal harm, requiring comprehensive governance, continuous monitoring, and regulatory oversight. These systems demand the highest levels of technical controls, human oversight, and assurance mechanisms.

### Financial Services

Credit scoring and underwriting systems in banking, financial services, and insurance that determine access to financial products and services.

### Employment and Education

Hiring screening algorithms, educational admissions systems, and performance evaluation tools that impact career and educational opportunities.

### Critical Infrastructure

Power grid operations, telecommunications core systems, transportation safety, and medical device control systems affecting public safety.

### Law Enforcement

Biometric identification systems, risk scoring algorithms, and predictive policing tools used in criminal justice contexts.

## Medium-Risk AI

- Fraud detection for payments and e-commerce platforms
- Content moderation and spam filtering systems
- Customer support copilots with sensitive data access
- Remote proctoring with biometric verification

## Low-Risk AI

- Code development assistants and productivity tools
- Content generation without personal data processing
- General-purpose chatbots with public information
- Internal automation tools without external impact





# AI System Inventory and Registration

The AI System Inventory serves as the authoritative source of truth for all artificial intelligence applications within an organization. This comprehensive registry enables risk management, compliance monitoring, vendor oversight, and incident response by maintaining detailed information about each system's purpose, technical architecture, data flows, and operational status.

Every AI system must be registered in the inventory regardless of risk classification, development stage, or deployment status. The inventory captures essential metadata including use case description, business owner, technical owner, data sources and types, model lineage and provenance, third-party vendors and dependencies, risk classification, deployment environment, and current operational status.

01	02	03
<b>System Identification</b>	<b>Metadata Collection</b>	<b>Risk Classification</b>
Comprehensive discovery of all AI systems including shadow IT, third-party integrations, and embedded AI capabilities within larger applications.	Detailed documentation of technical specifications, data flows, vendor relationships, and operational characteristics for each identified system.	Assignment of appropriate risk category based on use case, data sensitivity, potential impact, and regulatory requirements.
04	05	
<b>Governance Assignment</b>	<b>Continuous Maintenance</b>	
Designation of system owners, data stewards, and responsible parties for ongoing management and compliance.	Regular updates to reflect system changes, new deployments, decommissioning, and evolving risk profiles.	

High-risk systems require additional registration with the AI Risk and Ethics Committee (AIREC) for formal approval processes. This enhanced registration includes detailed technical documentation, risk assessment results, mitigation measures, monitoring plans, and incident response procedures. Regulatory notification may be required for specific high-risk use cases in regulated sectors.

The inventory integrates with existing IT asset management, security information systems, and compliance monitoring tools to provide real-time visibility into AI system deployments and changes. Automated discovery tools help identify new or modified AI capabilities, while integration APIs enable systematic data collection and reporting to regulatory authorities when required.

# Data Governance for AI Systems

Data governance forms the foundation of responsible AI development and deployment, ensuring that personal information, sensitive datasets, and proprietary content are processed lawfully, ethically, and securely throughout the AI lifecycle. The Data Protection and Digital Privacy Act requirements must be operationalized through technical controls and governance processes that span data collection, processing, storage, and deletion.

Lawful basis for data processing must be clearly documented for both training and inference phases of AI system operations. This includes explicit consent mechanisms where required, legitimate interest assessments for business processing, and public interest justifications for government applications. Purpose limitation principles require that data collected for specific purposes cannot be used for incompatible AI training or inference without additional legal basis.



Sensitive personal data and children's data require enhanced protections including additional consent requirements, purpose restrictions, and technical safeguards. Data quality and representativeness for Indian populations, languages, and demographic groups must be systematically evaluated to prevent algorithmic bias and ensure inclusive AI system performance.

Privacy-enhancing technologies including pseudonymization, anonymization, differential privacy, and synthetic data generation should be employed where appropriate to minimize privacy risks while maintaining AI system utility. However, these techniques must be properly implemented and validated to ensure they provide meaningful privacy protection rather than privacy theater.

**Implementation Note:** Organizations should establish clear data lineage tracking from source to model deployment, enabling efficient response to data subject rights requests including access, rectification, and deletion under the DPDP Act.

# Secure Model Development

Secure model development practices protect AI systems from threats throughout the development lifecycle, including data poisoning, model theft, supply chain attacks, and adversarial manipulation. These practices integrate security considerations from initial design through production deployment, establishing defense-in-depth strategies that protect both the development environment and the resulting AI systems.

Development environments must implement comprehensive isolation and access controls, including secure computing environments for sensitive model training, isolated secrets management using key management systems or hardware security modules, network segmentation between development and production systems, and role-based access controls with principle of least privilege. All development activities must maintain detailed audit logs with cryptographic integrity protection.



## Reproducible Development

Version control for code, data, and model artifacts with cryptographic signing and provenance tracking throughout the development pipeline.



## Documentation Standards

Model cards and data sheets initiated at design time, capturing assumptions, limitations, and intended use cases for transparency and accountability.



## Threat Modeling

Systematic analysis of attack vectors including prompt injection, data poisoning, model extraction, and inference-time manipulation vulnerabilities.



## Supply Chain Security

AI Bill of Materials (AIBOM) capturing datasets, model weights, software libraries, and tools with verification of integrity and provenance.

The AI Bill of Materials (AIBOM) represents a critical innovation for supply chain transparency, documenting all components that contribute to AI system functionality. This includes training datasets with their sources and licenses, pre-trained models and fine-tuning data, software libraries and framework versions, development tools and platforms, third-party APIs and services, and cryptographic attestations for critical components.

Model cards and data sheets must be initiated during the design phase and continuously updated throughout development. These documents serve multiple purposes: enabling informed deployment decisions, supporting regulatory compliance, facilitating third-party audits, and providing transparency to end users and stakeholders about system capabilities and limitations.



## Development Security Controls

- Multi-factor authentication for all development access
- Code review requirements for AI-specific components
- Dependency scanning and vulnerability management
- Secure storage for training data and model artifacts
- Network monitoring and anomaly detection
- Regular security assessments and penetration testing



# Pre-Deployment Evaluation Gates

Pre-deployment evaluation gates establish mandatory checkpoints that AI systems must pass before production deployment, ensuring safety, security, privacy, fairness, and performance standards are met. These evaluations provide objective evidence of system readiness and regulatory compliance while identifying potential risks and mitigation requirements.

Safety evaluations assess the AI system's ability to avoid harmful outputs, follow instructions appropriately, resist attempts to bypass safety controls, and maintain appropriate boundaries in tool usage and information access. These evaluations include adversarial testing, jailbreak resistance assessment, harmful content generation testing, and evaluation of the system's response to edge cases and unexpected inputs.

## 1 Safety Assessment

Harmful content detection, instruction following evaluation, jailbreak resistance testing, and tool abuse prevention validation across diverse attack vectors and scenarios.

## 2 Security Evaluation

Adversarial robustness testing, model extraction resistance, backdoor detection, and retrieval-augmented generation (RAG) security hardening assessment.

## 3 Privacy Protection

Membership inference risk analysis, personally identifiable information leakage detection, output filtering effectiveness, and data redaction capability validation.

## 4 Fairness Analysis

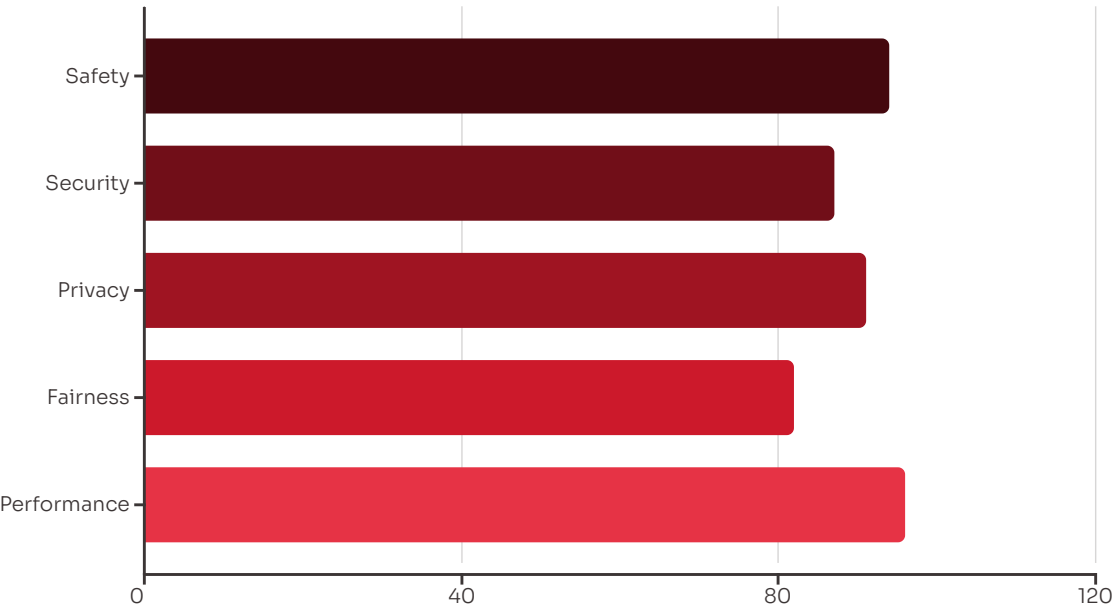
Group fairness evaluation, error parity assessment across Indian demographic segments, language performance consistency, and counterfactual bias testing.

## 5 Performance Validation

Accuracy benchmarking, system stability testing, latency and throughput measurement, cost efficiency analysis, and regression detection across system updates.

Fairness evaluations require particular attention to India's diverse demographic landscape, including assessment across linguistic groups, regional variations, socioeconomic segments, gender identities, age groups, and other protected characteristics. These evaluations must use representative datasets and culturally appropriate metrics that reflect local contexts and values rather than simply adopting international benchmarks.

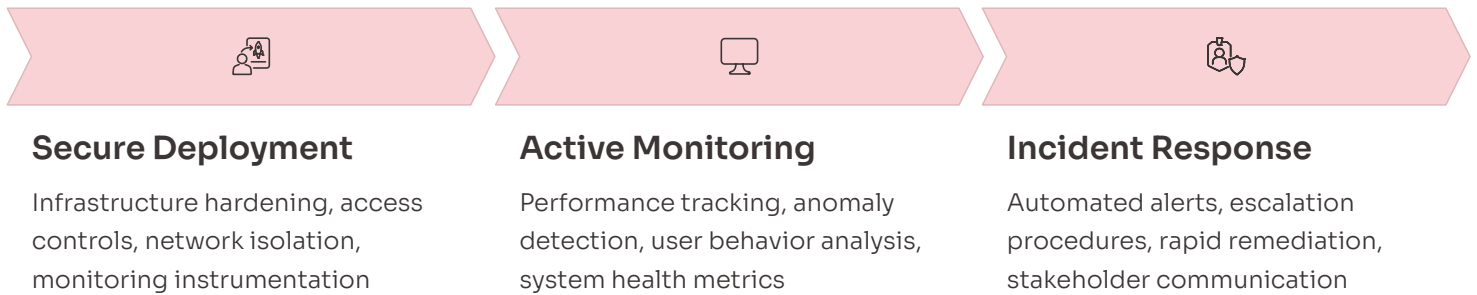
Performance evaluations encompass both technical performance metrics and operational readiness indicators. Technical metrics include accuracy, precision, recall, and F1 scores across different user segments and use cases. Operational metrics cover system latency, throughput capacity, resource utilization, cost per inference, and degradation patterns under load conditions.



# Production Deployment and Operations

Production deployment marks the transition from development to live operations, requiring robust infrastructure, monitoring systems, and operational procedures that maintain security, performance, and compliance throughout the AI system lifecycle. Deployment practices must balance system availability with risk management, ensuring that AI systems can be rapidly updated or disabled when issues are identified.

Environment hardening establishes the foundational security posture for AI system operations, including network isolation and segmentation, rate limiting and abuse detection mechanisms, authentication and authorization controls, encrypted communications and data storage, and automated kill-switch capabilities for rapid system shutdown. These controls protect against both external attacks and internal misuse while enabling legitimate system usage.



Content provenance capabilities become increasingly important as AI-generated content proliferates across digital platforms and communication channels. Where applicable, AI systems should embed provenance metadata using standards such as C2PA (Coalition for Content Provenance and Authenticity) manifests or equivalent technologies. This metadata enables verification of content origins, detection of AI-generated materials, and maintenance of information integrity across distribution chains.

Audit logging requirements align with CERT-In directions and DPDP Act obligations, capturing comprehensive records of system interactions, decisions, and data processing activities. These logs must include user prompts and system responses, tool invocations and results, access attempts and authorization decisions, data processing operations, system configuration changes, and security events and anomalies. All log entries must include synchronized timestamps and cryptographic integrity protection.

Human-in-the-loop workflows ensure that high-risk decisions maintain appropriate human oversight and provide meaningful appeal mechanisms for affected individuals. These workflows must define clear criteria for human intervention, specify qualified human reviewers, establish reasonable response timeframes, and provide accessible appeal procedures that comply with natural justice principles.

❌ **Operational Requirement:** All production AI systems must implement automated rollback capabilities that can restore previous system versions within 15 minutes of detecting safety, security, or performance issues.



# Continuous Monitoring and Change Control

Continuous monitoring ensures that AI systems maintain their intended performance, safety, and compliance characteristics throughout their operational lifecycle. Unlike traditional software systems, AI applications can exhibit performance degradation due to data drift, model staleness, adversarial attacks, and changing operational contexts that require sophisticated detection and response mechanisms.

Data drift detection identifies when the statistical properties of input data diverge from training distributions, potentially indicating changing user behavior, environmental conditions, or adversarial manipulation. Statistical tests, distribution comparisons, and anomaly detection algorithms provide early warning systems that trigger evaluation and potential model retraining processes.

## Real-Time Monitoring

Continuous performance tracking, anomaly detection, and automated alerting for immediate response to system degradation or security incidents.

## Triggered Assessments

Event-driven evaluations following significant incidents, data changes, or regulatory updates that may impact system performance or compliance.

## Scheduled Evaluations

Weekly fairness assessments, monthly security scans, and quarterly comprehensive evaluations to identify gradual changes in system behavior.

## Annual Reviews

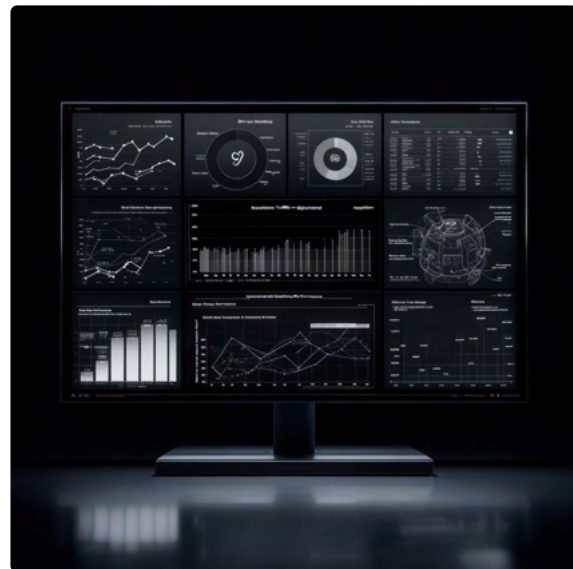
Comprehensive system audits including security assessments, fairness evaluations, and compliance verification by independent third parties.

Model output monitoring tracks the quality, consistency, and appropriateness of AI system responses across different user segments, use cases, and time periods. This monitoring includes accuracy metrics, response quality assessments, bias detection across demographic groups, safety violation detection, and user satisfaction indicators. Automated monitoring systems must be complemented by human review processes that can identify subtle quality degradation or emerging safety issues.

Change control processes ensure that updates to AI systems, including model retraining, parameter adjustments, prompt engineering changes, and infrastructure modifications, are properly tested, approved, and documented before deployment. The AI Bill of Materials must be updated to reflect all changes, and impact assessments should evaluate how changes affect safety, security, privacy, and fairness characteristics.

## Monitoring Metrics

- Model accuracy and performance indicators
- Response latency and system availability
- Error rates and failure patterns
- Security event frequency and severity
- User satisfaction and feedback scores
- Regulatory compliance indicators



Patch management for AI systems presents unique challenges due to the complex dependencies on training data, model weights, software libraries, and external services. Updates to any component may impact system behavior in unexpected ways, requiring comprehensive testing and gradual rollout procedures. The AIBOM enables systematic tracking of all dependencies and their versions, facilitating impact assessment and rollback procedures when updates cause problems.



# System Decommissioning and Data Deletion


System decommissioning represents a critical but often overlooked aspect of AI governance, requiring systematic procedures to safely retire AI systems while protecting sensitive data, maintaining audit trails, and complying with regulatory obligations. Improper decommissioning can create security vulnerabilities, privacy violations, and compliance failures that persist long after system retirement.

Verified model retirement encompasses the secure deletion of model weights, training data, intermediate artifacts, and cached outputs from all storage locations including primary systems, backup repositories, development environments, and third-party services. This process requires cryptographic verification of deletion operations and comprehensive audit trails documenting the decommissioning process.

01	02	03
<b>Decommissioning Planning</b>	<b>Data Inventory</b>	<b>Access Revocation</b>
Assessment of data retention requirements, regulatory obligations, audit trail preservation needs, and stakeholder notification procedures.	Comprehensive identification of all data locations including primary storage, backups, logs, caches, and third-party systems containing AI system data.	Systematic removal of user access, service account permissions, API keys, and authentication credentials associated with the AI system.
04	05	
<b>Secure Deletion</b>	<b>Documentation Update</b>	
Cryptographically verified deletion of sensitive data while preserving required audit trails and compliance documentation.	Updates to AI system inventory, risk registers, and compliance documentation to reflect system retirement status.	

Data retention obligations under the DPDP Act and sector-specific regulations may require preservation of certain records even after system decommissioning. Organizations must carefully balance these retention requirements with privacy principles and data minimization obligations, maintaining only the minimum necessary data for the required retention periods.

The AI System Inventory must be updated to reflect the decommissioned status, including decommissioning date, responsible parties, verification of data deletion, preserved audit records, and any ongoing obligations related to the retired system. This documentation supports compliance audits and provides historical context for future AI governance activities.

 **Regulatory Requirement:** Under DPDP Act provisions, individuals maintain the right to request deletion of their personal data even from decommissioned AI systems, requiring organizations to maintain deletion capabilities and audit trails.

Third-party relationships and vendor agreements must be systematically terminated as part of the decommissioning process. This includes data processing agreements, API access permissions, shared model access, and any ongoing support or maintenance contracts. Vendors must provide deletion certificates and compliance attestations confirming their compliance with decommissioning requirements.



# Technical Control Catalog


The technical control catalog provides specific implementation guidance for security, privacy, fairness, and operational controls that organizations should implement based on their AI system risk classifications. These controls represent proven practices that address common vulnerabilities and regulatory requirements while enabling beneficial AI system functionality.

Security controls protect AI systems from malicious attacks, unauthorized access, and data breaches through multiple layers of technical safeguards. Retrieval-Augmented Generation (RAG) systems require specialized security patterns including isolated vector storage with encrypted databases, allowlisted tool access with permission boundaries, contextual compression to minimize data exposure, and retrieval filtering to prevent unauthorized information access.


<b>Input Security</b> Prompt injection detection, content scanning, input validation, and malicious payload filtering	<b>Processing Security</b> Model isolation, secrets management, access controls, and execution sandboxing
<b>Output Security</b> Data loss prevention, PII redaction, policy enforcement, and response filtering	<b>Infrastructure Security</b> Network isolation, encryption, monitoring, and incident response capabilities

Prompt injection mitigation represents a critical security control for language model applications, requiring multiple defensive layers including content scanning systems that detect malicious inputs, instruction firewalls that enforce system boundaries, tool segregation that limits privilege escalation, and canary prompts that detect attempted manipulation. These controls must be regularly updated as attack techniques evolve.


Privacy controls operationalize DPDP Act requirements through technical mechanisms that protect personal data throughout the AI lifecycle. Consent capture systems must provide clear, specific, and verifiable consent mechanisms with proof of consent storage and withdrawal capabilities. Purpose flags must propagate through data processing pipelines to ensure data is only used for authorized purposes.



**Privacy-Enhancing Technologies**  
Differential privacy, federated learning, homomorphic encryption, and secure multi-party computation for privacy-preserving AI training and inference.



**Fairness and Bias Controls**  
Balanced dataset curation, demographic parity monitoring, equalized odds evaluation, and bias mitigation techniques tailored to Indian contexts.



**Content Provenance**  
Digital watermarking, C2PA manifests, blockchain attestation, and verifiable credential systems for AI-generated content authentication.

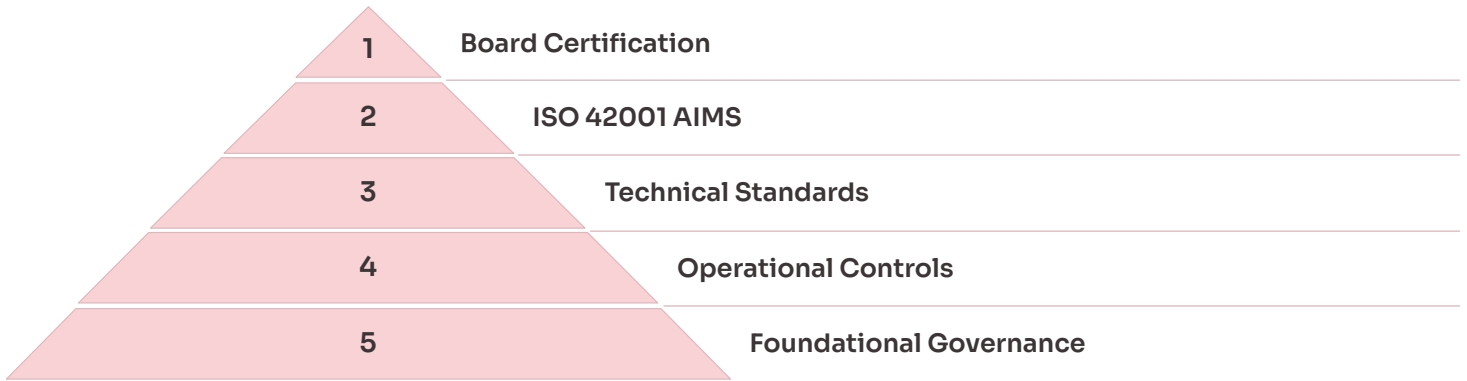
Fairness and inclusivity controls ensure that AI systems provide equitable outcomes across India's diverse population segments. Dataset balancing requires representative sampling across demographic groups, linguistic communities, regional variations, and socioeconomic segments. Fairness metrics must be carefully selected based on the specific application context and stakeholder requirements, with thresholds established through multi-stakeholder consultation processes.



# Assurance and Certification Framework

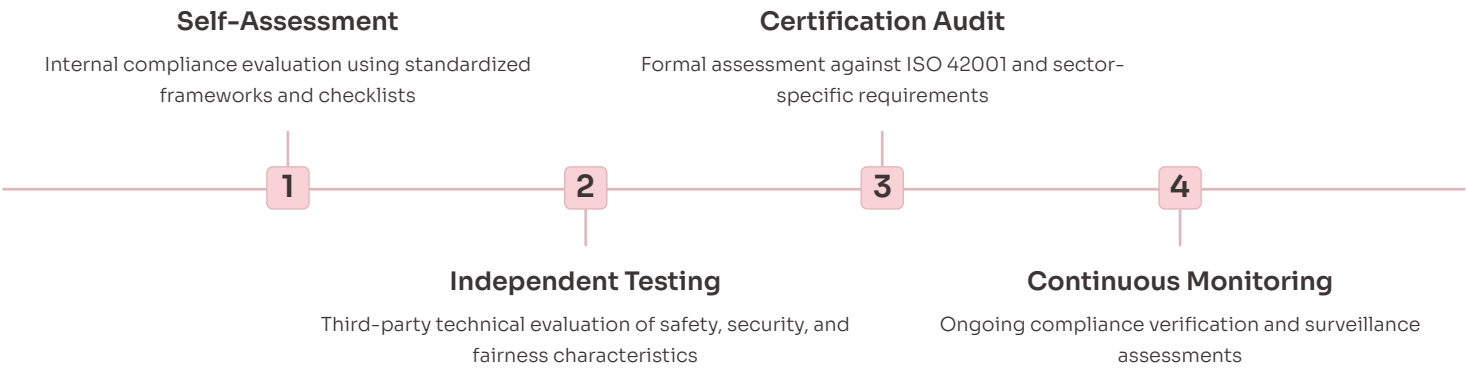
The assurance and certification framework establishes systematic approaches for validating AI system compliance with governance requirements, regulatory standards, and international best practices. This framework provides organizations with clear pathways to demonstrate responsible AI practices while enabling regulators and stakeholders to verify compliance through standardized assessment methods.

ISO/IEC 42001 adoption establishes comprehensive AI Management Systems (AIMS) that integrate with existing ISO 27001 information security management and privacy management programs. This integration creates cohesive governance frameworks that address AI-specific risks while leveraging established management system maturity and operational processes.



Independent testing leverages the IndiaAI Safety Institute laboratories for red-teaming exercises, benchmark evaluations, and conformance testing at scale. These testing capabilities provide standardized evaluation environments, validated test datasets, reproducible evaluation procedures, and credible third-party assessments that support regulatory compliance and public trust.

Third-party audit requirements apply to high-risk AI systems and organizations processing significant volumes of personal data through AI applications. Annual audits assess governance effectiveness, technical control implementation, incident response capabilities, and compliance with applicable regulations. Event-driven audits may be triggered by significant incidents, regulatory changes, or stakeholder concerns.



Transparency reporting enables organizations to demonstrate responsible AI practices through publication of model cards, data sheets, safety reports, and regular transparency reports. These disclosures provide stakeholders with sufficient information to make informed decisions while protecting proprietary information and security-sensitive details.

Technical file maintenance represents a critical component of the assurance framework, requiring organizations to maintain comprehensive documentation including system architecture diagrams, risk assessment results, evaluation reports, incident logs, change control records, and compliance attestations. These files support regulatory inspections, audit activities, and internal governance processes.

Certification Level	Requirements	Audit Frequency
Basic Compliance	Self-assessment, basic controls	Self-certification
Enhanced Assurance	Independent testing, technical files	Biennial audit
Premium Certification	ISO 42001, continuous monitoring	Annual audit

# Public Sector Implementation Blueprint

Public sector AI implementation requires specialized approaches that address unique accountability mechanisms, transparency obligations, citizen rights protections, and democratic governance principles. Government entities face distinct challenges including public trust requirements, constitutional obligations, electoral considerations, and complex stakeholder ecosystems that demand tailored governance frameworks.

Central government playbooks establish standardized procedures for high-risk system approvals, incident response coordination, and crisis communications management. These playbooks enable coordinated responses across ministries and departments while maintaining appropriate escalation pathways to political leadership and senior civil service officers. Specialized procedures for deepfake incidents and election-related AI threats ensure rapid response capabilities during critical periods.

## Ministry-Level Governance

Dedicated AI governance committees, risk assessment procedures, citizen impact evaluations, and parliamentary reporting mechanisms

## Regulatory Coordination

Inter-agency coordination mechanisms, shared evaluation resources, common technical standards, and unified incident response

## PSU Implementation

Board-level oversight, commercial viability assessments, public interest evaluations, and performance monitoring systems

Procurement controls represent a critical lever for driving responsible AI adoption across the public sector ecosystem. Government procurement requirements should mandate ISO 42001 certification or equivalent governance frameworks, require comprehensive AI Bill of Materials documentation, insist on model cards and transparency reports, demand fairness testing on Indian demographic datasets, and specify localization requirements for language and cultural contexts.

Regulatory and civic sandboxes provide controlled environments for piloting innovative AI applications while maintaining appropriate oversight and risk management. These sandboxes enable rapid iteration and learning while establishing clear boundaries, monitoring requirements, and graduation criteria for broader deployment. Sandbox programs should include clear entry criteria, defined testing parameters, monitoring and evaluation procedures, and paths to production deployment or termination.

Open ecosystem promotion supports India's broader AI strategy through advancement of open-source models and datasets, transparent licensing frameworks, collaborative research initiatives, and shared evaluation resources. Government leadership in open ecosystem development demonstrates commitment to democratic AI governance while building national capabilities and reducing dependence on proprietary foreign technologies.



## Citizen-Centric Design

- Accessible appeal mechanisms for AI-driven decisions
- Plain-language explanations of AI system usage
- Opt-out capabilities where legally permissible
- Regular public consultation on AI policy
- Transparency reporting on government AI usage
- Digital literacy programs for AI awareness

**Public Accountability:** All government AI systems that significantly impact citizens must publish annual transparency reports detailing usage statistics, accuracy metrics, bias assessments, and citizen feedback analysis.





# Enterprise Implementation Blueprint

Enterprise AI implementation varies significantly across sectors based on regulatory requirements, risk tolerance, technical capabilities, and business models. The framework provides sector-specific guidance while maintaining consistent core principles and governance structures that can be adapted to diverse organizational contexts and operational environments.

Banking, Financial Services, and Insurance (BFSI) sector implementations must address stringent regulatory oversight, systemic risk considerations, and consumer protection obligations. Credit scoring applications require comprehensive fairness thresholds with regular testing across demographic segments, reasonable explanation capabilities for adverse decisions, complete audit trails for regulatory examination, and stress testing under adverse economic scenarios to ensure system stability.

### Banking & Financial Services

Credit scoring fairness, explainable decisions, regulatory reporting, systemic risk assessment, consumer protection mechanisms

### Healthcare

Clinical safety oversight, patient data protection, post-market surveillance, medical device regulations, professional liability

### Telecommunications

Network security, lawful intercept compliance, supply chain assurance, critical infrastructure protection, service reliability

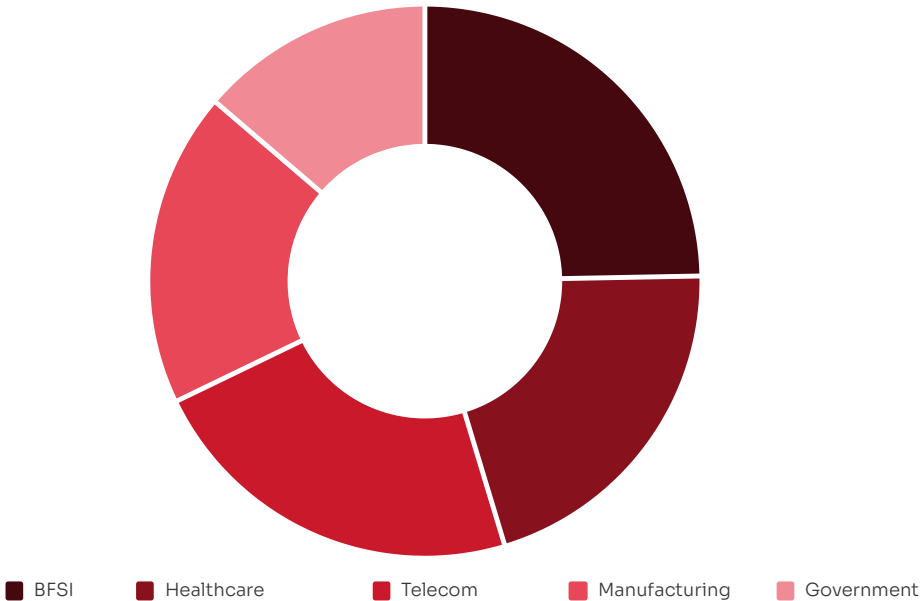
### Manufacturing

Operational technology security, safety interlocks, quality control, supply chain optimization, predictive maintenance

Healthcare AI implementations require specialized governance addressing clinical safety, patient privacy, and medical professional oversight. Clinical safety officers must maintain decision-making authority for AI-assisted medical applications, comprehensive dataset governance must protect patient health information while enabling beneficial research, and post-market surveillance systems must monitor AI diagnostic and treatment recommendation systems for safety and efficacy in real-world clinical settings.

Telecommunications sector implementations must balance AI innovation with critical infrastructure protection, national security considerations, and universal service obligations. AI applications in 5G network cores require specialized security architectures, lawful intercept capabilities must be maintained while protecting privacy rights, and supply chain assurance becomes critical given the national security implications of telecommunications infrastructure.

Manufacturing and Operational Technology (OT) environments present unique challenges due to safety-critical operations, legacy system integration, and industrial control system architectures. AI implementations must respect established Purdue model network segmentation, maintain safety interlocks and emergency shutdown capabilities, and ensure firmware integrity for machine learning applications deployed at the network edge.



Cross-sector considerations include data localization requirements, cross-border data transfer compliance, third-party vendor risk management, and integration with existing enterprise risk management frameworks. Organizations must develop capabilities for managing AI-specific risks while leveraging established governance processes and regulatory relationships.



# 100-Day Quick Start Implementation

The 100-day implementation plan provides organizations with a structured approach to establish foundational AI governance capabilities rapidly while building momentum for longer-term program development. This accelerated timeline focuses on critical path activities that provide immediate risk reduction and regulatory compliance while establishing the organizational foundation for comprehensive AI governance.

Leadership and governance establishment represents the highest priority during the initial implementation period. Organizations must appoint a Chief AI Risk Officer (CARO) with appropriate authority and resources, constitute the AI Risk and Ethics Committee (AIREC) with cross-functional representation, and secure board-level approval for the AI governance program charter, risk appetite statement, and prohibited use cases list.

### Days 1-30: Foundation

Appoint CARO, constitute AIREC, secure board approval, establish governance charter, define prohibited uses

### Days 31-60: Discovery

Complete AI system inventory, classify risks, halt shadow IT deployments, assess current capabilities

### Days 61-100: Controls

Implement evaluation gates, draft AIPIA procedures, establish monitoring capabilities, train staff

System discovery and inventory activities must be completed comprehensively to establish accurate baseline understanding of existing AI system deployments. This includes systematic identification of all AI applications across the organization, documentation of high-risk systems requiring immediate attention, classification of systems according to the risk taxonomy, and immediate freeze on new deployments pending governance implementation.

Policy and procedure development should focus on essential templates and processes that enable immediate compliance and risk management. Priority documents include AI Privacy and Impact Assessment (AIPIA) templates, evaluation gate requirements and minimum thresholds, incident response procedures for AI-specific events, and vendor assessment criteria for AI service providers.

30

Days

Governance foundation establishment

100%

Coverage

AI system inventory completion

5

Templates

Essential policy documents

24/7

Monitoring

Incident response capability

Staff training and capability development must begin immediately to ensure organizational readiness for ongoing governance activities. Training programs should address AI risk concepts and terminology, organizational roles and responsibilities, policy and procedure requirements, incident reporting obligations, and regulatory compliance expectations. These programs should be tailored to different organizational roles and technical skill levels.

### Week 1-4 Priorities

- Leadership appointments and committee formation
- Stakeholder communication and change management
- Resource allocation and budget approval
- External advisor and vendor engagement

### Week 5-8 Priorities

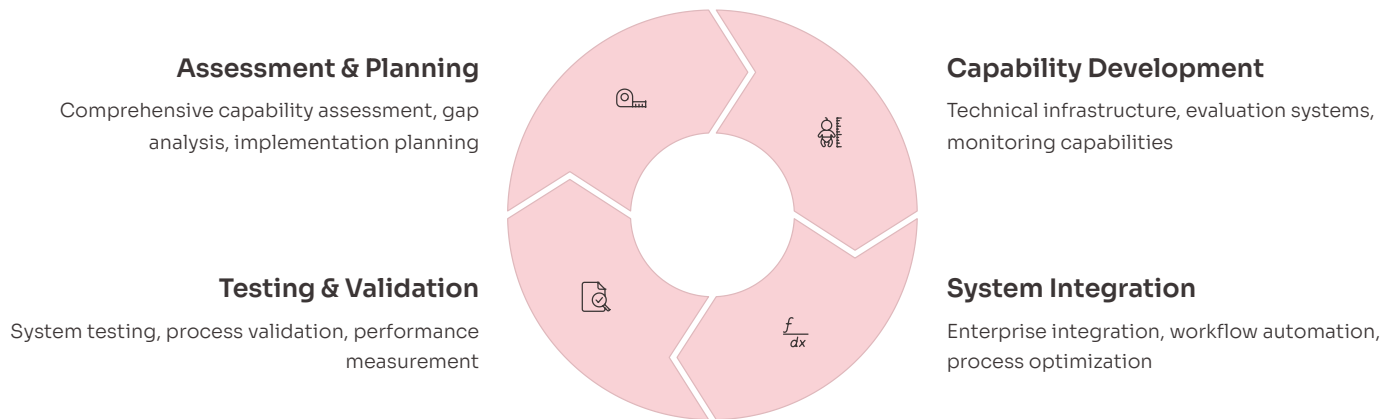
- Comprehensive system discovery and documentation
- Risk assessment and classification activities
- Shadow IT identification and control measures
- Vendor and third-party risk assessment

✔ **Success Metric:** Organizations completing the 100-day implementation should achieve 100% AI system inventory coverage, established governance committees, and operational incident response capabilities.

# 12-Month Maturity Development

The 12-month implementation plan builds comprehensive AI governance capabilities on the foundation established during the initial 100-day period. This phase focuses on operational maturity, systematic process implementation, technical capability development, and integration with existing organizational systems and regulatory requirements.

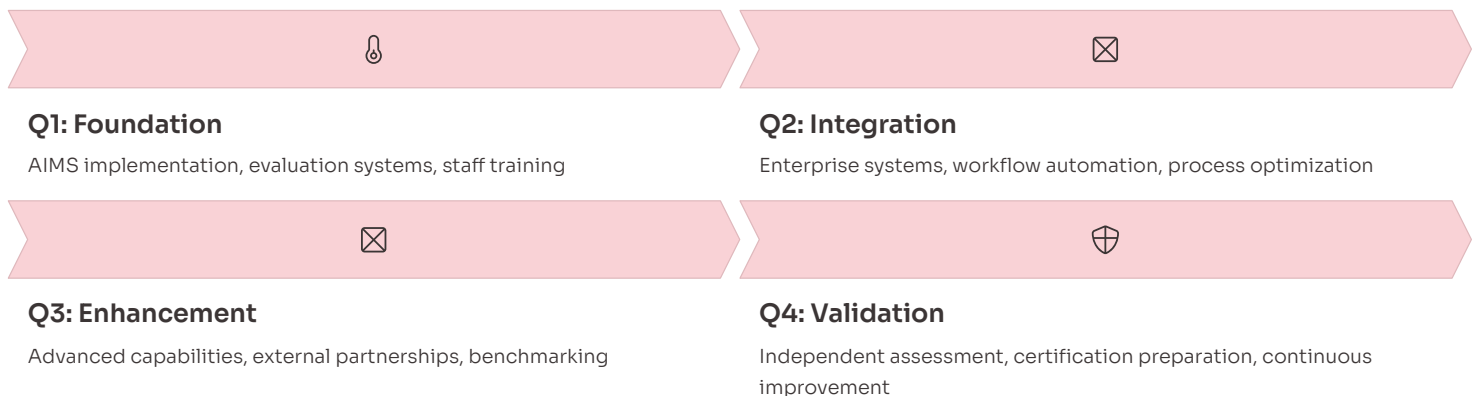
ISO 42001-aligned AI Management System implementation represents the central organizing framework for this maturity phase. Organizations must establish systematic risk management processes, documented procedures for all AI governance activities, performance monitoring and measurement systems, continuous improvement mechanisms, and integration with existing ISO 27001 information security and DPDP Act privacy compliance programs.



Evaluation harness development enables systematic testing of AI systems across safety, security, privacy, fairness, and performance dimensions. These evaluation systems must incorporate Indian demographic datasets and linguistic requirements, provide reproducible testing environments, support continuous evaluation workflows, and generate compliance reports for regulatory requirements. Integration with IndiaAI Safety Institute resources accelerates capability development and ensures consistency with national standards.

Content provenance and watermarking implementation addresses the growing challenge of AI-generated content authentication and misinformation prevention. Organizations should implement digital watermarking systems for AI-generated media, content provenance metadata using C2PA or equivalent standards, verification credential systems for attestations, and public transparency mechanisms that enable content authentication without compromising privacy or security.

Advanced monitoring and analytics capabilities provide organizations with sophisticated visibility into AI system performance, risk indicators, and compliance status. These systems should include real-time performance dashboards, automated anomaly detection and alerting, trend analysis and predictive indicators, regulatory reporting automation, and executive summary reports for board and senior management oversight.

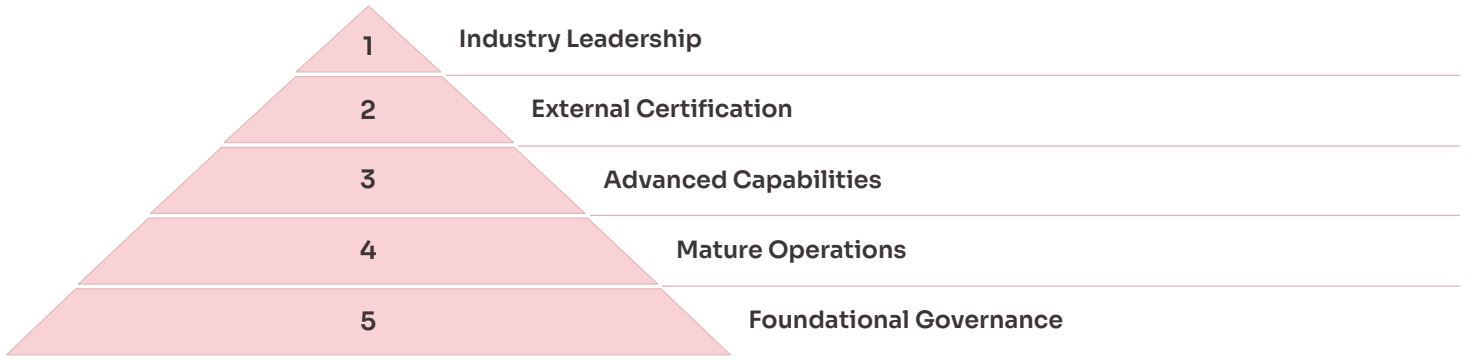


Supply chain assurance capabilities become increasingly important as organizations deploy more complex AI systems with multiple vendor dependencies. Organizations must develop vendor assessment frameworks, contract terms and conditions for AI services, ongoing monitoring of third-party AI capabilities, incident coordination with vendor partners, and termination procedures that protect organizational interests and data.

# 24-Month Strategic Excellence

The 24-month implementation milestone represents achievement of strategic AI governance excellence, positioning organizations as leaders in responsible AI deployment and regulatory compliance. This advanced maturity phase focuses on external validation, industry leadership, ecosystem participation, and continuous innovation in AI governance practices.



External certification achievement demonstrates organizational commitment to responsible AI practices through independent validation of governance systems, technical capabilities, and operational processes. Organizations should pursue ISO 42001 certification with specialized modules for their industry sector, participate in regulatory certification programs where available, and engage with international standard-setting bodies to influence future AI governance standards.



Transparency reporting establishes organizations as trusted leaders in responsible AI development and deployment. Annual transparency reports should detail AI system usage statistics, performance and fairness metrics across demographic segments, incident response and resolution statistics, regulatory compliance achievements, and contributions to AI safety research and open-source communities. These reports demonstrate accountability while building public trust and stakeholder confidence.

Ecosystem participation enables organizations to contribute to and benefit from collaborative AI governance development. Organizations should join sector-specific threat intelligence sharing programs, participate in regulatory sandboxes and pilot programs, contribute to open-source AI safety and evaluation tools, engage in academic research partnerships, and support policy development through expert consultation and public commentary.

Advanced technical capabilities distinguish leading organizations through sophisticated AI risk management and assurance systems. These capabilities include automated red-teaming and adversarial testing, continuous fairness monitoring across demographic segments, supply chain attestation and verification systems, cross-organizational incident response coordination, and predictive risk analytics that anticipate potential issues before they manifest.

<b>Certification</b> ISO 42001, sector-specific standards, international recognition	 <b>Transparency</b> Public reporting, stakeholder engagement, trust building
<b>Ecosystem</b> Industry collaboration, research partnerships, policy contribution	 <b>Innovation</b> Advanced capabilities, thought leadership, standard setting

Continuous improvement and innovation ensure that AI governance capabilities evolve with technological advancement, regulatory development, and emerging risk scenarios. Organizations should establish research and development programs focused on AI governance innovation, participate in international standard development activities, and contribute to the broader AI governance knowledge base through publication and knowledge sharing.



### Strategic Benefits

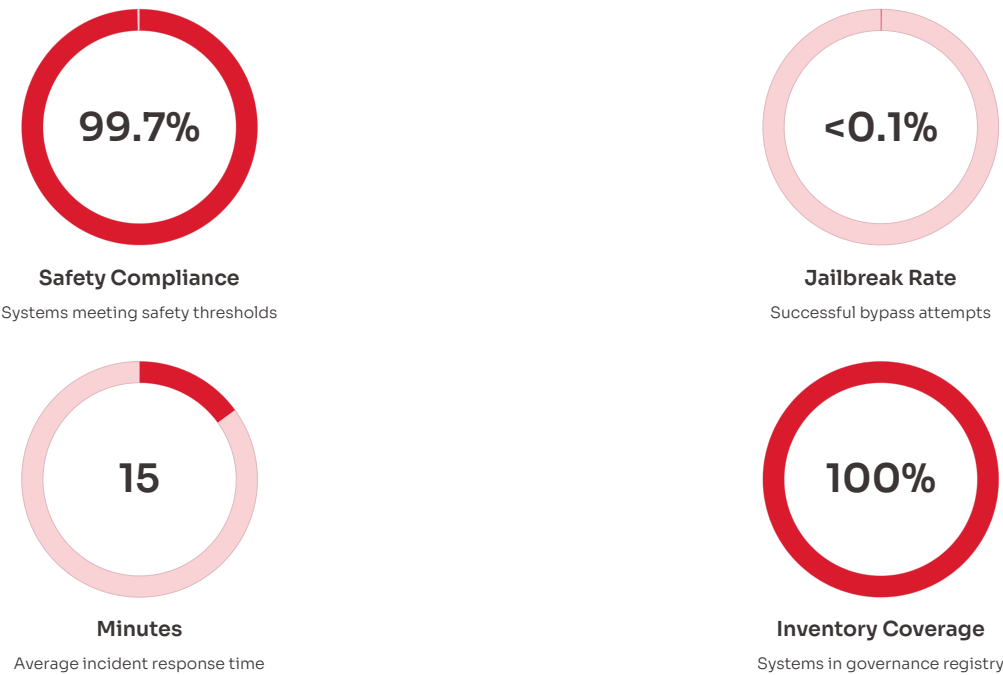
- Enhanced regulatory relationships and reduced compliance costs
- Competitive differentiation through demonstrated responsibility
- Access to restricted markets and high-assurance procurement
- Reduced insurance premiums and risk-adjusted financing
- Attraction and retention of top AI talent
- Board and investor confidence in AI risk management



# Performance Metrics and Reporting

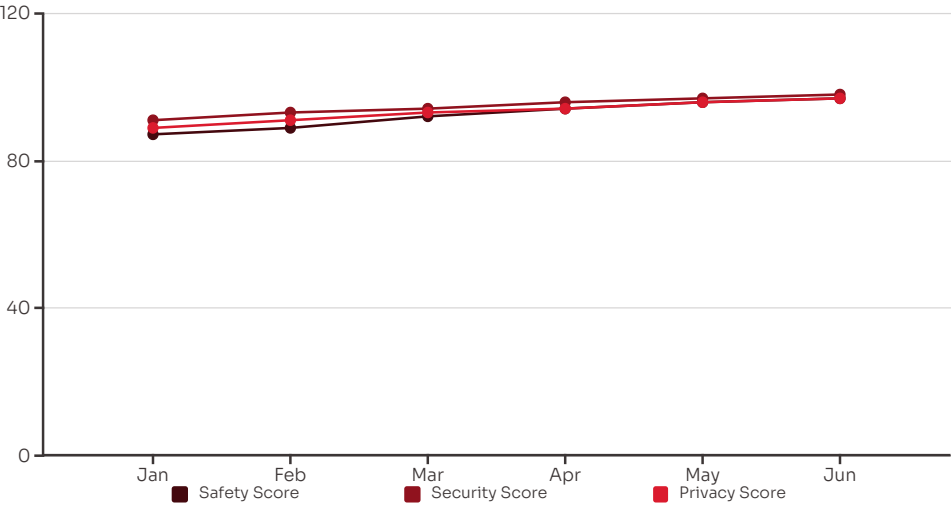
Comprehensive performance metrics enable organizations to measure AI governance effectiveness, demonstrate regulatory compliance, identify improvement opportunities, and provide transparency to stakeholders. The metrics framework spans technical performance, governance effectiveness, regulatory compliance, and stakeholder satisfaction across multiple organizational levels and reporting timeframes.

Safety metrics assess the AI system's ability to avoid harmful outputs and maintain appropriate operational boundaries across diverse usage scenarios. Jailbreak success rates measure the system's resistance to attempts to bypass safety controls, while harmful content generation rates track the frequency of inappropriate, offensive, or dangerous outputs. Tool abuse detection rates evaluate the system's ability to identify and prevent misuse of integrated capabilities and external service connections.



Security metrics evaluate the AI system's resilience against malicious attacks, unauthorized access attempts, and data compromise scenarios. Model extraction resistance measures the system's protection against attempts to steal or reverse-engineer proprietary models, while poisoning detection capabilities assess the ability to identify malicious training data or adversarial inputs designed to corrupt system behavior.

Privacy metrics align with DPDP Act requirements and international privacy standards, measuring the system's effectiveness in protecting personal information throughout the AI lifecycle. Personally Identifiable Information (PII) leakage rates track inadvertent disclosure of sensitive data in system outputs, while membership inference risk assessments evaluate whether attackers can determine if specific individuals' data was used in model training.



Fairness metrics ensure that AI systems provide equitable treatment across India's diverse demographic segments, linguistic communities, and cultural contexts. Error parity measurements compare system performance across different population groups to identify potential bias, while demographic parity assessments evaluate whether system outcomes are distributed fairly across protected characteristics.

Governance effectiveness metrics assess the operational success of AI governance processes, organizational capabilities, and compliance activities. Inventory coverage rates measure the comprehensiveness of AI system registration and oversight, while evaluation completion rates track adherence to pre-deployment testing requirements. Audit finding closure rates indicate the organization's responsiveness to identified governance gaps and improvement opportunities.

Metric Category	Key Indicators	Reporting Frequency
Safety	Harmful content rate, jailbreak resistance	Daily





# Enforcement and Accountability Mechanisms

Effective AI governance requires robust enforcement mechanisms that provide clear consequences for non-compliance while incentivizing responsible AI practices through safe harbors and regulatory recognition. The enforcement framework balances deterrence with encouragement, creating structured pathways for organizations to demonstrate compliance while establishing clear consequences for violations.

High-risk system registration and attestation requirements establish mandatory disclosure and transparency mechanisms for AI applications that can significantly impact individuals or society. Organizations deploying high-risk AI systems must maintain comprehensive technical files documenting system architecture, risk assessments, evaluation results, and ongoing monitoring data. These files must be available for regulatory inspection and support independent auditing activities.



Safe harbor provisions provide regulatory protection for organizations that demonstrate proactive risk management, transparent reporting, and collaborative engagement with regulatory authorities. These protections include reduced penalties for organizations that voluntarily report incidents, expedited approval processes for pre-cleared AI applications, regulatory recognition of certified governance systems, and participation benefits for regulatory sandbox programs.

Penalty frameworks align with existing regulatory structures while addressing AI-specific violations and harms. Privacy-related violations should align with DPDP Act penalty structures, while sector-specific regulators maintain authority over safety and integrity breaches within their jurisdictions. Criminal referrals may be appropriate for intentional violations involving significant harm, manipulation of democratic processes, or willful circumvention of safety controls.

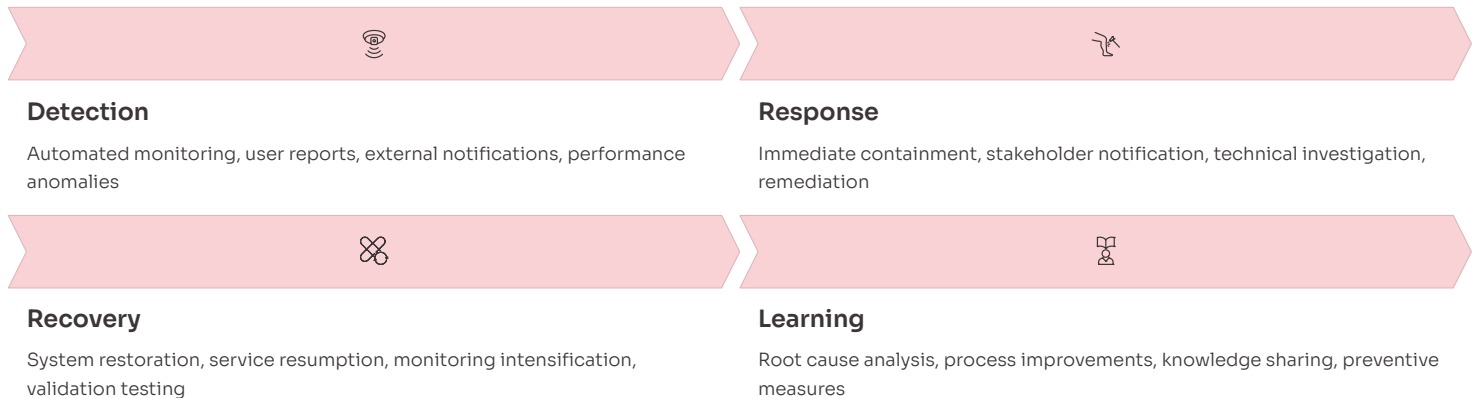
Grievance redressal mechanisms ensure that individuals affected by AI system decisions have meaningful recourse and appeal rights. These mechanisms must provide accessible complaint procedures, reasonable response timeframes, qualified human reviewers for complex technical issues, and independent oversight through ombudsperson or equivalent institutions. Alternative dispute resolution procedures can provide efficient resolution for many AI-related grievances while preserving judicial review for fundamental rights issues.

01	02
<b>Initial Complaint</b>  Accessible submission channels, acknowledgment within 48 hours, initial triage and classification	<b>Investigation</b>  Technical review by qualified personnel, stakeholder consultation, evidence gathering
03	04
<b>Resolution</b>  Corrective action implementation, individual remedy provision, system-wide improvements	<b>Appeal</b>  Independent review, escalation procedures, final determination with reasoning
<b>Enforcement Tools</b> <ul style="list-style-type: none"><li>Administrative sanctions and penalties</li><li>License suspension or revocation</li><li>Mandatory audits and remediation</li><li>Public disclosure of violations</li><li>Criminal referral for serious violations</li><li>Civil liability for damages</li></ul>	<b>Compliance Incentives</b> <ul style="list-style-type: none"><li>Regulatory recognition of certified systems</li><li>Expedited approval processes</li><li>Reduced audit frequency for compliant organizations</li><li>Safe harbor protections for voluntary reporting</li><li>Procurement preferences for certified systems</li><li>Industry leadership recognition programs</li></ul>

# Incident Response and Crisis Management

AI incidents can escalate rapidly and cause widespread harm, requiring specialized response capabilities that address technical failures, security breaches, privacy violations, bias-related harms, and societal disruption. Incident response frameworks must integrate with existing CERT-In reporting requirements while addressing AI-specific challenges such as model behavior changes, data poisoning attacks, and deepfake proliferation.

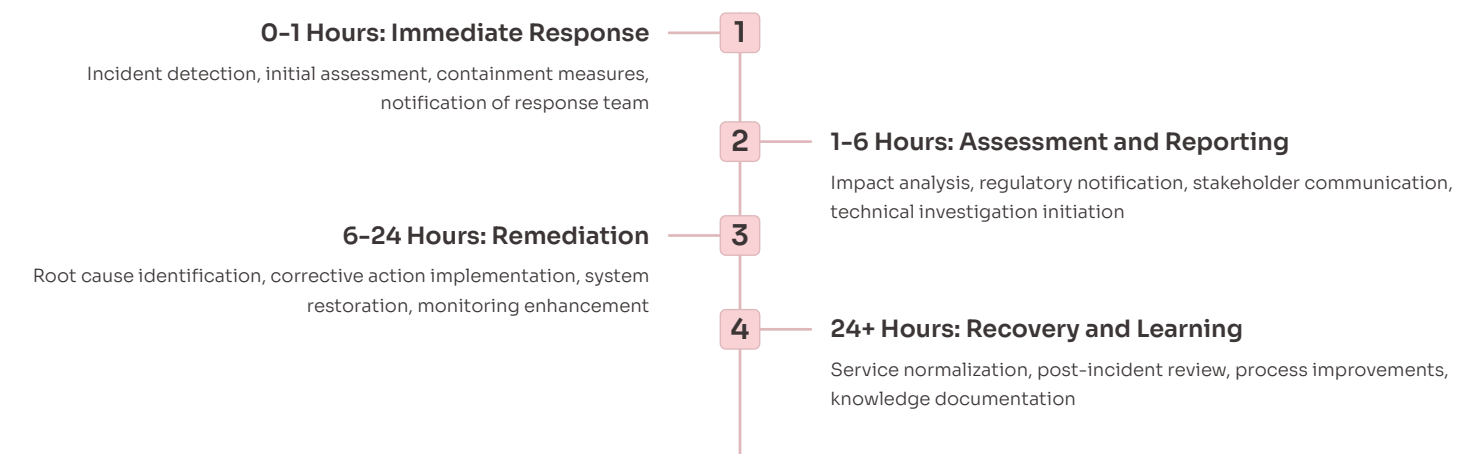
Incident classification systems enable appropriate response prioritization and resource allocation based on potential impact, affected populations, technical severity, and regulatory implications. Critical incidents include safety failures causing physical harm, large-scale privacy breaches, election-related misinformation campaigns, and systemic failures affecting critical infrastructure or essential services.



Deepfake crisis management requires coordinated response across platforms, media organizations, and government entities to limit the spread of malicious synthetic content while preserving legitimate discourse. Response procedures should include content identification and labeling systems, rapid takedown coordination mechanisms, public communication strategies that address misinformation without amplifying harmful content, and collaboration with fact-checking organizations and media literacy initiatives.

Technical response capabilities must address AI-specific incident characteristics including model rollback and version control procedures, data pipeline isolation and contamination assessment, adversarial attack detection and mitigation, bias identification and correction mechanisms, and supply chain incident coordination with vendors and partners.

Regulatory reporting obligations under CERT-In directions require incident notification within six hours for specified categories, while DPDP Act breach notification requirements apply to privacy-related incidents. Organizations must develop integrated reporting procedures that satisfy multiple regulatory requirements without duplication or inconsistency, maintaining detailed incident logs that support forensic analysis and regulatory inspection activities.



Cross-organizational coordination becomes critical for incidents affecting multiple entities or requiring specialized expertise. Incident response networks should include relationships with IndiaAI Safety Institute for technical analysis, sector regulators for compliance guidance, law enforcement for criminal investigations, and peer organizations for threat intelligence sharing and coordinated response activities.

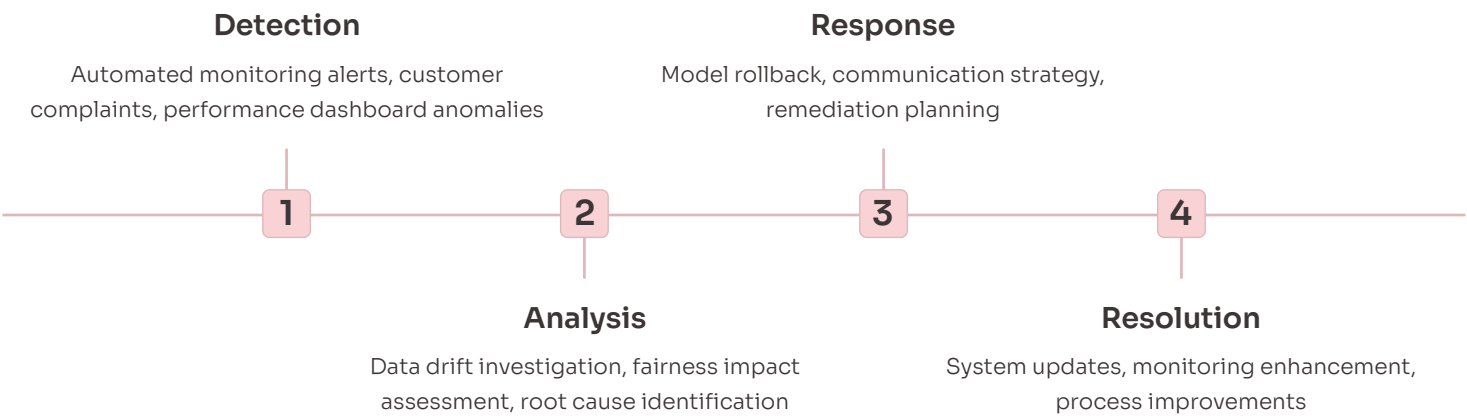
**⚠ Critical Requirement:** All AI incidents with potential safety, security, or rights implications must be reported to CERT-In within 6 hours and documented with sufficient technical detail to support investigation and prevention of similar incidents.



# Case Study: Payment Fraud Detection System

This case study examines a representative scenario involving drift in device signal processing within a payment fraud detection system, illustrating how the AI governance framework addresses real-world operational challenges while maintaining system effectiveness and regulatory compliance. The scenario demonstrates the integration of technical controls, governance processes, and stakeholder management in resolving AI system issues.

The payment fraud detection system experienced sudden increases in false positive rates affecting legitimate transactions, with disproportionate impact on users from specific geographic regions and device types. Initial investigation revealed that changes in mobile device software updates had altered behavioral signals used by the machine learning model, causing drift in input data distributions that degraded system performance.



Governance framework activation began with automated drift detection systems triggering alerts within hours of performance degradation. The AI Risk and Ethics Committee (AIREC) was immediately notified and convened emergency sessions to assess impact scope, approve immediate containment measures, coordinate stakeholder communications, and authorize resource allocation for investigation and remediation activities.

Fairness and performance gate evaluation revealed that the system was failing demographic parity requirements, with error rates significantly higher for users from certain states and economic segments. This triggered mandatory bias assessment procedures and human oversight requirements until system performance could be restored to acceptable thresholds across all population segments.

Technical remediation involved rolling back to the previous model version while engineers developed updated training datasets that accounted for new device signal patterns. The rollback process followed established change control procedures, with comprehensive testing in staging environments and gradual deployment to production systems with continuous monitoring and rollback capabilities.

**Immediate Actions**  
Automated alerts triggered, AIREC convened, model rollback initiated, customer support briefed

**Investigation Findings**  
Data drift from device updates, geographic bias in error rates, inadequate monitoring thresholds

**Resolution Measures**  
Enhanced monitoring systems, bias detection improvements, communication protocols

Communication strategy included proactive customer notification explaining service disruptions, regulatory reporting to RBI and CERT-In as appropriate, internal stakeholder briefings for executive leadership, and public transparency reporting documenting lessons learned and preventive measures implemented to prevent similar incidents.

Post-incident review identified improvements including enhanced data drift detection capabilities, more granular fairness monitoring across demographic segments, improved incident escalation procedures, and additional training for technical teams on bias detection and remediation techniques. These improvements were incorporated into the organization's AI governance framework and shared with industry peers through sector collaboration mechanisms.

# Case Study: Healthcare Diagnostic AI System

This healthcare case study illustrates governance framework application when an AI diagnostic system demonstrates performance degradation for specific population segments, highlighting the intersection of clinical safety, regulatory compliance, and ethical AI deployment in healthcare settings. The scenario emphasizes the critical importance of continuous monitoring and rapid response in safety-critical applications.

A radiology AI system used for preliminary screening of chest X-rays began showing decreased sensitivity for detecting pneumonia in patients from specific regional backgrounds. The performance degradation was initially subtle but became statistically significant over several weeks, raising concerns about health equity and patient safety implications across diverse population segments served by the healthcare system.



## Performance Alert

Statistical monitoring detected decreased sensitivity for specific demographic segments



## Clinical Review

Radiologist investigation confirmed systematic performance degradation patterns



## Root Cause Analysis

Training data bias and population representation gaps identified



## Clinical Response

Enhanced human oversight, system recalibration, equity improvements

Clinical safety officer involvement ensured that patient safety remained the paramount concern throughout the incident response process. Immediate measures included enhanced radiologist review for affected patient populations, modification of AI confidence thresholds to increase sensitivity, communication with clinical teams about potential AI limitations, and retrospective review of recent cases to identify any missed diagnoses that could require patient follow-up.

Root cause analysis revealed training data limitations that inadequately represented the genetic and phenotypic diversity of the patient population served by the healthcare system. The AI model had been trained primarily on datasets from different geographic regions with different disease presentation patterns, population genetics, and imaging equipment characteristics.

Regulatory obligations required notification to medical device regulators about the performance issues, documentation of corrective actions taken to ensure patient safety, submission of updated risk assessments and mitigation measures, and ongoing post-market surveillance reporting to track system performance improvements and any additional safety concerns.



## Clinical Governance Integration

The healthcare AI governance framework integrated seamlessly with existing clinical governance structures, including medical staff committees, quality assurance programs, and patient safety initiatives. This integration ensured that AI-related incidents were handled with the same rigor and transparency as other clinical quality issues.

Multidisciplinary teams including clinicians, data scientists, bioethicists, and patient advocates collaborated to develop comprehensive remediation strategies that addressed both technical performance and broader health equity concerns.

Data augmentation and retraining efforts focused on acquiring more representative training datasets through partnerships with regional healthcare institutions, implementing transfer learning techniques to adapt the model to local population characteristics, and establishing ongoing data collection procedures to maintain model performance across diverse patient populations.

Long-term improvements included establishment of health equity monitoring as a standard component of AI system evaluation, enhanced collaboration with regional healthcare providers to ensure representative datasets, development of clinical decision support tools that explicitly account for population diversity, and creation of patient engagement mechanisms to ensure community input on AI system development and deployment decisions.

- ① **Clinical Integration:** Healthcare AI systems require specialized governance that integrates with clinical quality assurance, medical staff oversight, and patient safety programs while maintaining the highest standards of clinical care and professional accountability.



# Templates and Implementation Tools

Comprehensive implementation tools accelerate AI governance adoption by providing organizations with pre-developed templates, checklists, and frameworks that can be customized to specific organizational contexts and regulatory requirements. These tools reduce implementation time, ensure consistency with best practices, and provide clear guidance for complex governance activities.

The AI Privacy and Impact Assessment (AIPIA) template provides structured methodology for evaluating privacy risks, regulatory compliance, and societal impacts of AI system deployments. The template includes stakeholder identification and consultation procedures, lawful basis analysis and documentation, risk assessment matrices with quantitative and qualitative factors, mitigation strategy development and implementation planning, and residual risk evaluation with sign-off procedures.



## AIPIA Template

Comprehensive privacy and impact assessment framework with stakeholder consultation, risk analysis, and mitigation planning components



## Model Card

Standardized documentation for AI model capabilities, limitations, intended use cases, and performance characteristics



## AIBOM Framework

AI Bill of Materials capturing datasets, models, libraries, and dependencies with provenance tracking



## Evaluation Harness

Systematic testing procedures for safety, security, privacy, fairness, and performance validation

Model Card templates enable consistent documentation of AI system characteristics, capabilities, and limitations in formats accessible to both technical and non-technical stakeholders. These templates include system purpose and intended use cases, training data characteristics and limitations, performance metrics across different population segments, known biases and fairness considerations, safety limitations and prohibited use cases, and recommended human oversight requirements.

The AI Bill of Materials (AIBOM) framework provides systematic methodology for documenting all components that contribute to AI system functionality, enabling supply chain transparency and risk management. AIBOM components include training datasets with source documentation and licensing terms, pre-trained models and fine-tuning procedures, software libraries and framework dependencies, development tools and platform services, third-party APIs and external integrations, and cryptographic attestations for critical components.

01

## Template Customization

Adapt standard templates to organizational context, regulatory requirements, and sector-specific needs

02

## Workflow Integration

Embed templates into existing development and governance processes with automation where possible

03

## Training and Adoption

Provide staff training on template usage, quality standards, and completion requirements

04

## Continuous Improvement

Regular template updates based on user feedback, regulatory changes, and emerging best practices

Evaluation Harness Checklists provide systematic procedures for pre-deployment testing across safety, security, privacy, fairness, and performance dimensions. These checklists include specific test procedures and acceptance criteria, minimum passing thresholds for different risk categories, documentation requirements for evaluation results, escalation procedures for systems that fail evaluation gates, and remediation guidance for addressing identified issues.

Incident Report Templates ensure consistent documentation of AI-related incidents with sufficient detail to support investigation, regulatory reporting, and organizational learning. Templates include incident timeline and chronology, affected systems and user populations, impact assessment and severity classification, immediate response actions and containment measures, root cause analysis and contributing factors, corrective actions and preventive measures, and lessons learned and process improvements.

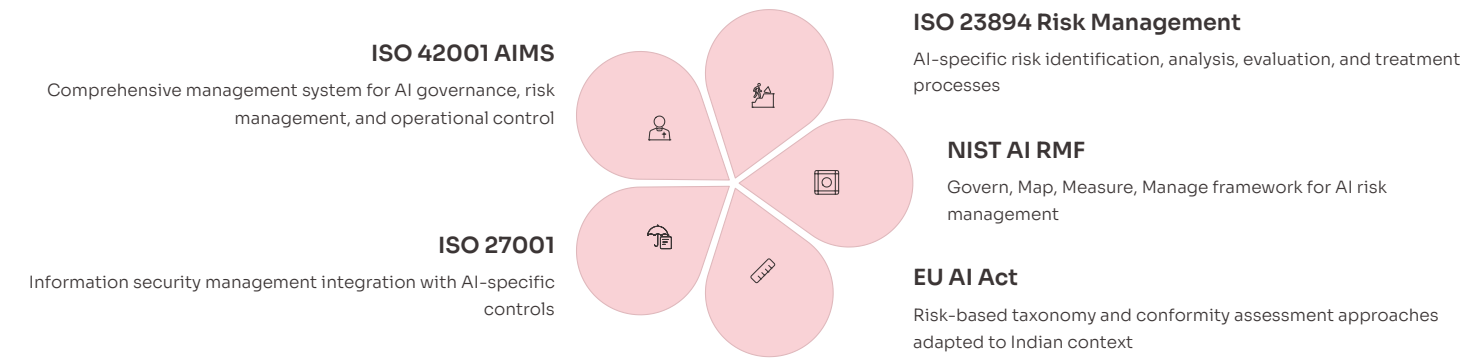
Template Type	Primary Use Cases	Update Frequency
AIPIA	Privacy impact assessment, regulatory compliance	Per system deployment
Model Card	System documentation, transparency reporting	Per model version
AIBOM	Supply chain management, audit preparation	Per system change
Evaluation Harness	Pre-deployment testing, continuous monitoring	Quarterly review
Incident Report	Incident response, regulatory reporting	Per incident



# International Standards Alignment

Alignment with international standards enables organizations to demonstrate global best practices while facilitating cross-border operations, international partnerships, and export opportunities. The framework provides specific mapping to key international standards while adapting requirements to India's unique legal, cultural, and operational context.

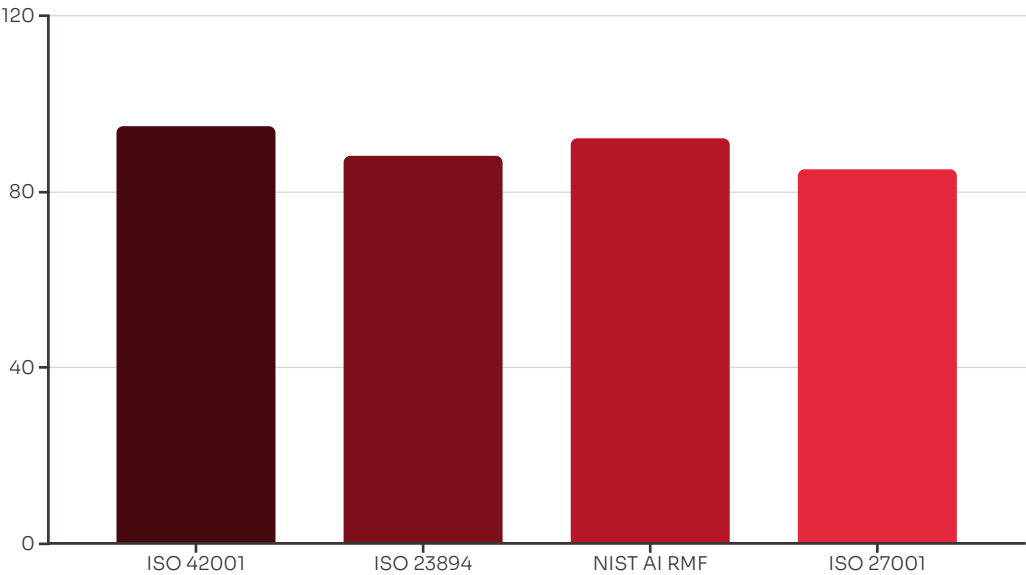
ISO/IEC 42001 AI Management Systems standard provides the foundational framework for systematic AI governance, establishing organizational context and leadership commitment, risk management processes and procedures, operational planning and control mechanisms, performance evaluation and monitoring systems, and continuous improvement methodologies. Organizations implementing this framework can achieve ISO 42001 certification with minimal additional requirements.



ISO/IEC 23894 AI Risk Management standard provides detailed guidance for AI-specific risk identification, analysis, evaluation, and treatment throughout the system lifecycle. The framework's risk classification taxonomy aligns directly with ISO 23894 risk categories, while technical controls and governance processes operationalize the standard's risk treatment recommendations.

NIST AI Risk Management Framework 1.0 structure of Govern, Map, Measure, and Manage phases corresponds directly to the framework's governance model, risk classification approach, evaluation requirements, and operational controls. Organizations can use NIST AI RMF profiles to document their specific risk management approaches for high-risk use cases in Indian contexts.

EU AI Act taxonomical approaches inform the framework's risk classification system while adapting prohibited and high-risk categories to Indian constitutional principles, legal requirements, and societal values. The framework's conformity assessment procedures draw inspiration from EU AI Act approaches while integrating with India's regulatory structures and institutional capabilities.



Integration with existing management systems leverages organizational investments in ISO 27001 information security management, ISO 9001 quality management, and other established frameworks. AI governance processes integrate with existing risk management, change control, incident response, and audit procedures while adding AI-specific requirements and capabilities.

Cross-border recognition enables organizations to leverage Indian AI governance certifications for international operations, partnerships, and market access. Alignment with international standards facilitates mutual recognition agreements, reduces duplicative compliance requirements, and demonstrates commitment to global best practices that build trust with international stakeholders.

🕒 **Certification Pathway:** Organizations implementing this framework can achieve ISO 42001 certification with 90%+ alignment to standard requirements, significantly reducing certification timeline and costs while demonstrating international best practices.



# Operational Handbook: Daily Rhythms

Sustainable AI governance requires embedding governance activities into daily operational rhythms, creating systematic procedures that maintain compliance and risk management without disrupting business operations. The operational handbook establishes regular cadences for monitoring, decision-making, and improvement activities that scale with organizational size and complexity.



Daily operational activities focus on immediate risk identification and response, ensuring that AI systems maintain appropriate safety, security, and performance characteristics. Anomaly triage procedures evaluate automated alerts from monitoring systems, distinguishing between routine fluctuations and genuine issues requiring intervention. Prompt injection and content policy violations require immediate assessment and potential system adjustments to maintain safety boundaries.



Weekly operational reviews provide systematic assessment of AI system performance trends, emerging risks, and governance effectiveness. Data drift detection requires regular evaluation of input data characteristics compared to training distributions, identifying gradual changes that may impact system performance over time. Fairness dashboards enable tracking of equity metrics across demographic segments, ensuring that bias issues are identified and addressed promptly.

Monthly governance activities include formal AI Risk and Ethics Committee meetings, comprehensive audit trail sampling, privacy request processing metrics, and risk register updates reflecting current threat landscape and organizational changes. These activities provide structured oversight while enabling rapid response to emerging issues or regulatory requirements.

Quarterly strategic reviews enable board-level oversight, external red-teaming exercises, comprehensive policy reviews, stakeholder engagement activities, and strategic planning for AI governance evolution. These reviews ensure that governance activities remain aligned with organizational objectives and regulatory expectations while adapting to technological and regulatory developments.

 <b>Monitoring Dashboard Reviews</b> Daily assessment of system performance metrics, user feedback patterns, security event logs, and compliance indicators across all deployed AI systems	 <b>Stakeholder Communication</b> Regular updates to business stakeholders, technical teams, compliance officers, and executive leadership on governance activities and risk status	 <b>Continuous Improvement</b> Systematic identification and implementation of governance process improvements based on operational experience and emerging best practices
---	---	--

Integration with existing operational procedures ensures that AI governance activities complement rather than duplicate established risk management, compliance, and operational processes. Organizations should leverage existing morning briefings, incident response procedures, change management processes, and performance review cycles while adding AI-specific components and considerations.

Escalation procedures define clear criteria and pathways for elevating issues from operational teams to governance committees and executive leadership. These procedures ensure that significant risks receive appropriate attention while preventing unnecessary escalation of routine operational issues that can be resolved through established procedures.

## Daily Checklist

- Review overnight system alerts and anomalies
- Assess performance metric dashboards
- Evaluate user feedback and complaints
- Monitor security event logs
- Check compliance status indicators
- Update incident response activities

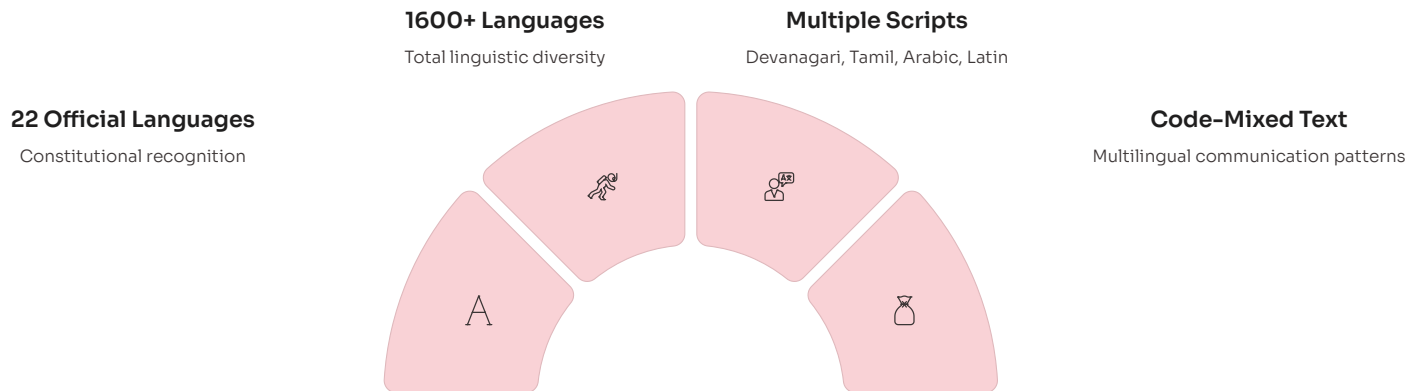
## Weekly Activities

- Comprehensive drift analysis reports
- Fairness metric trending evaluation
- System evaluation scheduling and completion
- Pipeline maintenance and updates
- Trend analysis and forecasting
- Team coordination and planning

# Building for India: Inclusion and Innovation

AI governance for India must address the nation's extraordinary linguistic, cultural, and socioeconomic diversity while supporting indigenous innovation and technological sovereignty. The framework prioritizes inclusive AI development that serves all segments of Indian society while building national capabilities for responsible AI leadership on the global stage.

Multilingual AI capabilities require comprehensive benchmarks and evaluation procedures for Indian languages, scripts, and code-mixed text patterns that reflect authentic communication practices across diverse communities. Evaluation frameworks must account for dialectical variations, regional linguistic patterns, cultural context sensitivity, and cross-lingual performance consistency. These capabilities ensure that AI systems serve all Indian language communities equitably rather than privileging English or major regional languages.



Energy efficiency and sustainability considerations reflect India's commitment to environmental responsibility and resource optimization. AI training and inference operations should prioritize energy-efficient algorithms, optimize compute resource utilization, leverage renewable energy sources where available, and implement carbon footprint monitoring and reduction strategies. These priorities align with India's climate commitments while reducing operational costs and improving system accessibility.

Open ecosystem development supports democratic AI governance principles through transparent, collaborative development approaches that enable broad participation and innovation. Government leadership in open model development, public dataset creation with clear licensing frameworks, collaborative research initiatives with academic institutions, and shared evaluation resources reduces dependence on proprietary foreign technologies while building national AI capabilities.

Cultural sensitivity and contextual appropriateness ensure that AI systems respect Indian values, traditions, and social norms while supporting beneficial innovation and progress. This includes understanding of family structures and social relationships, respect for religious and cultural practices, sensitivity to regional customs and preferences, and awareness of historical and contemporary social dynamics that affect AI system acceptance and effectiveness.



Digital inclusion initiatives ensure that AI governance frameworks support rather than hinder access to AI benefits across all segments of Indian society. This includes accommodating varying levels of digital literacy, providing accessible interfaces for users with disabilities, supporting low-bandwidth and offline operational modes, and designing systems that work effectively with basic mobile devices and limited connectivity infrastructure.

Indigenous innovation approaches recognize and incorporate traditional knowledge systems, local problem-solving approaches, and grassroots technological development that can enhance AI system effectiveness while respecting intellectual property and cultural rights. Collaboration with traditional knowledge holders, integration of local expertise in system design, and support for community-driven innovation initiatives strengthen AI systems while building inclusive development pathways.

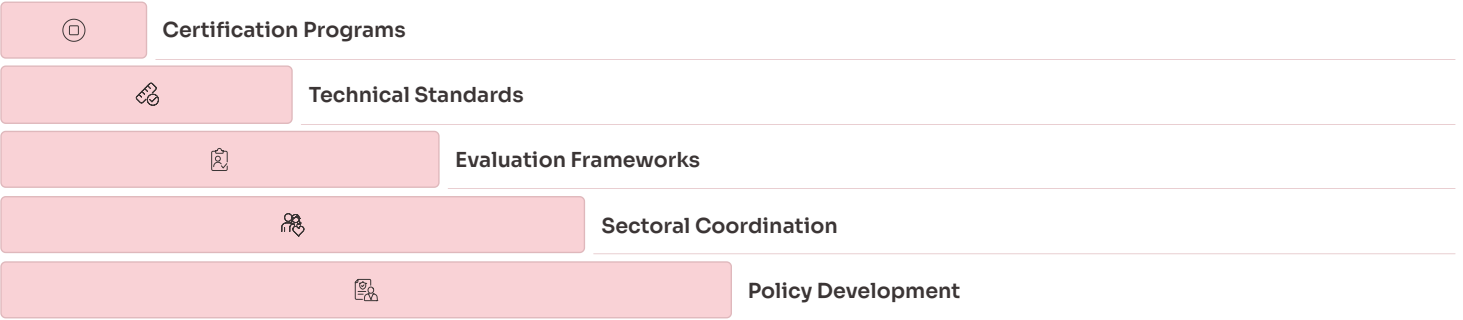
**Innovation Imperative:** India's AI governance framework should enable rather than constrain beneficial innovation, providing clear pathways for responsible development while maintaining appropriate safeguards for safety, privacy, and social welfare.



# NCAIC: National Leadership Role

The National Cyber and AI Center (NCAIC) plays a pivotal role in coordinating AI governance across India's complex institutional landscape, providing leadership in policy development, technical standards, capacity building, and international cooperation. NCAIC's unique position enables it to bridge regulatory silos, facilitate cross-sector collaboration, and represent India's interests in global AI governance discussions.

National conformance program development represents NCAIC's core contribution to systematic AI governance, establishing reference implementation profiles for high-risk AI applications, developing standardized test suites and evaluation procedures, creating certification frameworks for AI governance systems, and maintaining registries of certified AI systems and service providers. This program provides consistency across sectors while enabling innovation and competition.



Sectoral coordination mechanisms enable NCAIC to facilitate collaboration among regulatory bodies, public sector organizations, and industry stakeholders while respecting jurisdictional boundaries and institutional autonomy. Regular convening of regulatory forums, development of common technical standards, coordination of incident response activities, and facilitation of information sharing create synergies while avoiding duplication of effort.

Capacity building initiatives support widespread adoption of responsible AI practices through training programs for government officials, technical assistance for smaller organizations, development of educational resources and best practices, and support for academic research and development activities. These initiatives build national capabilities while ensuring that AI governance benefits extend beyond large organizations to the broader ecosystem.

International cooperation and representation position India as a leader in responsible AI governance through participation in global standard-setting bodies, bilateral and multilateral cooperation agreements, knowledge sharing with other nations, and advocacy for developing country interests in international AI governance discussions. This leadership role enhances India's influence while building beneficial partnerships and knowledge exchange opportunities.

01	02
<b>Policy Development</b>	<b>Standard Setting</b>
Coordinated policy framework development with stakeholder consultation and regulatory alignment	Technical standards development and maintenance with industry and academic collaboration
03	04
<b>Capacity Building</b>	<b>International Engagement</b>
Training, education, and technical assistance programs for government and industry	Global cooperation, knowledge sharing, and advocacy for developing country interests
Annual AI Governance State of Practice reporting provides comprehensive assessment of India's progress in responsible AI deployment, identifying trends in adoption rates, compliance achievements, incident patterns, and emerging challenges. This reporting enables evidence-based policy development while providing transparency and accountability to stakeholders and the general public.	
Research and development coordination leverages India's academic and research capabilities through funding for AI safety research, collaboration with international research institutions, support for open-source AI development, and facilitation of public-private research partnerships. This coordination ensures that governance frameworks remain current with technological developments while supporting indigenous innovation and capability development.	



## Strategic Priorities

- Cross-sector coordination and harmonization
- International leadership and cooperation
- Capacity building and education
- Research and development support
- Public-private partnership facilitation
- Transparency and accountability promotion



# Future Roadmap and Evolution

The AI governance landscape will continue evolving rapidly as technology advances, regulatory frameworks mature, and societal understanding of AI impacts deepens. The framework must remain adaptive and forward-looking, anticipating emerging challenges while maintaining stability and predictability for organizations investing in responsible AI development and deployment.

Technological evolution drivers include advancement of large language models and multimodal AI systems, emergence of artificial general intelligence capabilities, development of quantum computing applications for AI, integration of AI with Internet of Things and edge computing, and evolution of human-AI collaboration paradigms. The governance framework must anticipate these developments while maintaining relevance and applicability across diverse technological implementations.



Regulatory evolution patterns suggest increasing coordination among international regulatory bodies, development of mutual recognition agreements for AI governance certifications, harmonization of technical standards and evaluation procedures, and emergence of specialized regulatory bodies focused exclusively on AI governance. India's early adoption of comprehensive AI governance positions the nation to influence these global developments while protecting national interests.

Sectoral specialization will likely emerge as different industries develop domain-specific governance requirements, technical standards, and evaluation procedures. Healthcare AI governance may diverge significantly from financial services requirements, while critical infrastructure applications may require specialized security and safety frameworks. The foundational governance principles remain constant while implementation approaches adapt to sectoral needs.

Emerging challenges require proactive governance development including AI-generated misinformation and deepfake proliferation, algorithmic manipulation of democratic processes, AI-enabled cyber attacks and defense systems, cross-border data governance and AI supply chains, and human rights implications of AI deployment in authoritarian contexts. The framework's adaptability mechanisms enable rapid response to these emerging challenges while maintaining core protective principles.

**Technological Adaptation**  
Framework evolution to address emerging AI capabilities and applications while maintaining core governance principles

**International Harmonization**  
Increased cooperation and mutual recognition with global governance frameworks and standards

**Sectoral Specialization**  
Development of domain-specific governance requirements and technical standards

**Continuous Innovation**  
Ongoing development of governance tools, methodologies, and best practices

Innovation opportunities enable India to contribute leadership in AI governance development through novel approaches to multilingual AI evaluation, energy-efficient AI governance, privacy-preserving AI techniques, and culturally-sensitive AI design. These innovations can be exported globally while addressing India's specific challenges and priorities, creating economic opportunities while advancing responsible AI development worldwide.

Long-term strategic objectives include establishing India as a global leader in responsible AI development, creating sustainable economic opportunities through AI innovation, protecting fundamental rights and democratic institutions, building technological sovereignty and reduced dependence on foreign AI systems, and contributing to global AI safety and beneficial AI development for all humanity.

**Future Consideration:** How can India's AI governance framework evolve to address challenges we cannot yet anticipate while maintaining the flexibility and adaptability needed for emerging technological and social developments?



# Conclusion: A Framework for Responsible AI Leadership

The AI Governance Framework for India represents more than regulatory compliance or risk management—it embodies a comprehensive approach to realizing artificial intelligence's transformative potential while safeguarding individual rights, democratic institutions, and social welfare. This framework positions India as a global leader in responsible AI development, demonstrating that technological advancement and ethical governance can advance together rather than in tension.

The framework's success depends on widespread adoption across government, industry, and civil society, supported by strong leadership commitment, adequate resource allocation, continuous capability development, and sustained stakeholder engagement. Early implementers will gain competitive advantages while contributing to a responsible AI ecosystem that benefits all participants through reduced risks, enhanced trust, and improved outcomes.

1.4B	35	22	100%
Citizens	States/UTs	Languages	Coverage
Protected by responsible AI governance	Coordinated implementation across jurisdictions	Inclusive AI evaluation and deployment	Comprehensive governance across sectors

Implementation success requires sustained commitment from organizational leadership, adequate investment in governance capabilities, comprehensive staff training and development, integration with existing business and regulatory processes, and continuous improvement based on operational experience and emerging best practices. Organizations that treat AI governance as a strategic capability rather than compliance overhead will achieve better outcomes while contributing to broader ecosystem development.

The global implications of India's AI governance leadership extend far beyond national boundaries, influencing international standards development, demonstrating effective approaches for diverse democracies, contributing to global AI safety research and development, and advocating for developing country interests in international AI governance discussions. India's scale, diversity, and democratic values provide unique insights that benefit the global community while advancing national interests.

Future developments will test the framework's adaptability and resilience as technological capabilities advance, regulatory requirements evolve, and societal expectations change. The framework's foundation in fundamental principles, commitment to continuous improvement, and integration with democratic governance processes provide the flexibility needed to address emerging challenges while maintaining core protective functions.

<b>National Leadership</b> Position India as a global leader in responsible AI governance and development	<b>Rights Protection</b> Safeguard fundamental rights and democratic institutions in the AI age
<b>Innovation Enablement</b> Support beneficial AI innovation while managing risks and ensuring accountability	<b>Global Contribution</b> Contribute to worldwide AI safety and beneficial development for all humanity

The call to action is clear: India stands at a pivotal moment where proactive governance can shape AI's role in society rather than merely responding to its consequences. Government leaders, industry executives, civil society organizations, and individual citizens all have roles to play in implementing responsible AI governance that serves India's democratic values and development aspirations while contributing to global AI safety and beneficial development.

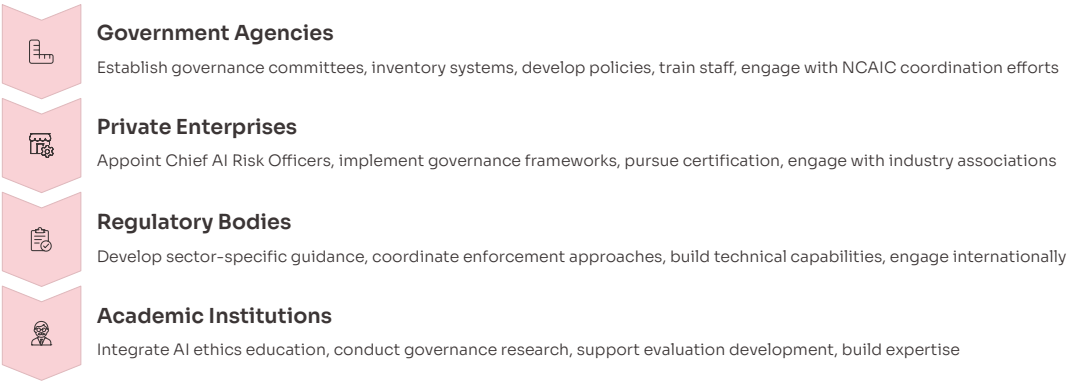




# Call to Action: Implementation Steps

Successful implementation of the AI Governance Framework requires coordinated action across multiple stakeholders, clear accountability mechanisms, and systematic approaches to capability building and change management. This call to action provides specific next steps for different stakeholder groups while emphasizing the urgency and importance of immediate action to establish responsible AI governance before risks escalate.

Government agencies and public sector organizations should immediately establish AI governance committees with appropriate authority and resources, conduct comprehensive inventories of existing AI deployments and planned systems, develop internal policies aligned with the framework's principles and requirements, and initiate staff training programs to build necessary capabilities. These foundational steps enable more sophisticated governance activities while ensuring immediate risk management.



Private sector enterprises should appoint Chief AI Risk Officers with appropriate authority and budget, implement governance frameworks appropriate to their risk profile and sector requirements, pursue third-party certification where beneficial for competitive positioning, engage with industry associations to share best practices and coordinate approaches, and contribute to policy development through consultation and expert commentary.

Regulatory bodies across sectors should develop specific guidance for AI applications within their jurisdictions, coordinate enforcement approaches with other regulators to avoid conflicts or gaps, build internal technical capabilities to evaluate AI systems effectively, engage with international regulatory bodies to share experiences and develop common approaches, and participate in NCAIC coordination mechanisms to ensure systematic coverage.

Academic and research institutions should integrate AI governance and ethics education into technology curricula, conduct research on governance effectiveness and emerging challenges, support development of evaluation methodologies and tools, build expertise in AI safety and responsible development practices, and engage with policy development through research, consultation, and expert testimony.



Civil society organizations and advocacy groups should monitor AI deployment impacts on communities and vulnerable populations, advocate for inclusive governance that protects rights and promotes equity, participate in policy development and consultation processes, build public awareness of AI governance importance, and hold organizations accountable for responsible AI practices through transparency and engagement.

International cooperation opportunities include participation in global standard-setting bodies, bilateral cooperation agreements with like-minded nations, knowledge sharing through multilateral forums, joint research and development initiatives, and coordination on cross-border challenges such as AI-enabled misinformation and cyber threats.

The path forward requires sustained commitment, adequate resources, and recognition that AI governance represents an investment in India's technological future rather than a compliance burden. Organizations that embrace comprehensive AI governance will gain competitive advantages while contributing to a responsible AI ecosystem that benefits all participants through reduced risks, enhanced trust, and improved outcomes for society as a whole.

## Success Indicators

- 100% AI system inventory coverage
- Established governance committees
- Staff training program completion
- Policy framework implementation
- Incident response capability
- Stakeholder engagement activities

## Resources Required

- Dedicated leadership and staff time
- Technology infrastructure investment
- Training and capability development
- External expertise and consultation
- Legal and compliance support
- International coordination activities



# Acknowledgments and Contact Information

This whitepaper represents collaborative effort among government agencies, industry leaders, academic institutions, and civil society organizations committed to responsible AI development and deployment in India.

“

## Contributing Organizations

National Cyber and AI Center (NCAIC), Ministry of Electronics and Information Technology, CERT-In, IndiaAI Mission, various sectoral regulators, leading technology companies, premier academic institutions, and civil society advocacy groups

”

The framework benefits from international collaboration and knowledge sharing with global partners committed to responsible AI governance and development. Special recognition goes to the technical experts, policy analysts, legal scholars, and community representatives who contributed expertise and insights throughout the development process.

## Contact Information

### National Cyber and AI Center

Website: [www.ncaic.in](http://www.ncaic.in)

Email: [governance@ncaic.in](mailto:governance@ncaic.in)

Policy Consultation: [policy@ncaic.in](mailto:policy@ncaic.in)

Technical Support: [technical@ncaic.in](mailto:technical@ncaic.in)

## Framework Updates

This framework will be updated regularly to address emerging technologies, regulatory developments, and implementation feedback. Current version, updates, and implementation resources available through the NCAIC website and associated consultation mechanisms.

© 2025 National Cyber and AI Center. This framework is made available under open license terms to support widespread adoption and implementation of responsible AI governance practices across India and internationally.

Together, we build the future of responsible artificial intelligence.

