Office of the Principal Scientific Adviser
to the Government of India

सत्यमेव जयते

STRENGTHENING AI GOVERNANCE THROUGH TECHNO-LEGAL FRAMEWORK

January 2026

India's AI Policy Priorities White paper Series

Acknowledgement

**Office of the Principal Scientific Adviser
to the Government of India**

सत्यमेव जयते

About: White Paper Series on Emerging Policy Priorities for India's AI Ecosystem

To foster informed deliberation and action among stakeholders engaged in shaping India's artificial intelligence (AI) policy and governance landscape, the Office of the Principal Scientific Adviser to the Government of India is producing this White Paper Series. These papers are conceived as explanatory briefs that examine specific policy issues and their associated nuances, with the aim of enabling broader understanding and constructive societal engagement. The White Papers are developed by drawing on collective insights from the extended AI ecosystem, including inputs from multi-stakeholder consultations, bilateral and multilateral AI policy engagements, and subsequent expert reviews. They are intended solely as explanatory documents that highlight identified policy priorities and stimulate further discussion. The views presented in these White Papers should not be construed as formal policy positions of the PSA Office.

"AI is already reshaping our polity, our economy, our security and even our society. AI is writing the code for humanity in this century."

"During our G20 Presidency, we built a consensus on Harnessing AI Responsibly, for Good, and for All. Today, India leads in AI adoption, and techno-legal solutions on data privacy."

Hon'ble Prime Minister Narendra Modi
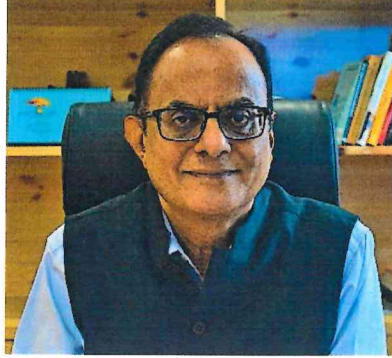during AI Action Summit in Paris, Feb 11, 2025

अजय के. सूद
भारत सरकार के प्रमुख वैज्ञानिक सलाहकार
Ajay K. Sood
Principal Scientific Adviser to the Govt. of India

कर्तव्य भवन 3, जनपथ, नई दिल्ली – 110001
Kartavya Bhavan 3, Janpath, New Delhi-110001
Tel. : +91-11-24011867, 24011868
E-mail : sood.ajay@gov.in, office-psa@nic.in
Website : www.psa.gov.in

Foreword

Artificial Intelligence (AI) is being adopted rapidly across sectors. While it has significant transformative potential, it is essential to ensure that associated risks and harms do not undermine trust or become a barrier to innovation and adoption. Therefore, developing a robust and responsive governance framework is not just a necessity but a prerequisite for sustaining the momentum of technological progress.

Globally, countries and regions are pursuing different approaches for AI governance shaped through their regulatory priorities, institutional capacity, and ecosystem needs. Some are building risk-based frameworks for higher-risk AI uses across sectors. Others rely on principle-based guidance and standards-driven implementation to provide flexibility with strengthening accountability. The recently released India AI Governance Guidelines Report sets out a pro-innovation approach that combines baseline legal safeguards, sectoral regulatory norms, technical measures and institutional mechanisms to safe and trusted AI. It proposes a "techno-legal" model as a viable pathway for AI governance.

This white paper is an attempt to define the "techno-legal" approach in a larger context as an integration of legal instruments, rule-based conditioning, and technical enforcement mechanisms embedded into AI architecture and operation by design. It recognises that effective governance needs to be an ecosystem-wide stakeholder approach. The purpose of this white paper is to serve as an explanatory document designed to highlight the evolving policy position in the country with respect to AI governance and facilitate informed deliberation among all relevant stakeholders. This initiative aims to build further understanding on the nuances of techno legal framework and how India can lead this discussion and provide a pathway for the collective governance framework for AI.

(Ajay K Sood)

Date: 22nd January 2026

Table of
Contents

List of Figures:

List of Tables:

1. Introduction

The rapid advancement of Artificial Intelligence (AI) is creating opportunities for innovation but also poses governance challenges. Owing to AI's rapidly evolving, adaptive, and borderless nature, conventional regulatory approaches are proving inadequate. Traditionally, regulation has relied on "command-and-control" laws, which mean that organizations must comply with the formal rules set by competent authorities, often under penalty of law. In response to the regulatory gaps posed by AI's unique characteristics, policymakers around the world are exploring various methods to govern AI in a manner that harnesses the innovation potential while also safeguarding social and ethical values. India requires an AI governance framework that is pro-innovation yet robust enough to protect society from associated risks and harms.

Currently, India's AI governance is guided by baseline regulations (such as the IT Act 2000, BNS 2023, and DPDP Act 2023 [15][16]), Intellectual Property Rights laws, sectoral guidelines (from regulators such as RBI [8], SEBI, ICMR etc.) and, policy advisory guidelines (e.g., India AI Governance Guidelines 2025) [13]. While this set the broad guidance, there are also sector-agnostic voluntary frameworks that the ecosystem may adopt, which includes ISO/IEC 42001 (AI Management System standard adopted by BIS), fairness standards framed by the Standard for Fairness Assessment and Rating of Artificial Intelligence Systems" (TEC 57050:2023) etc), [18]. This operates alongside MeitY's proposed amendments to the IT Rules, 2021 on synthetically generated information [15], and other government-led initiatives such as Safe & Trusted AI pillar of the IndiaAI mission.

Scan the QR Code to access the IndiaAI Governance Guidelines

While these baseline regulations and frameworks provide safeguards, they are not specifically designed to address the emerging and unknown complexities arising from the AI's development, adoption, deployment, and population scale usage. These complexities may include breach of privacy, security, fairness (violated by being biased), intellectual property and sa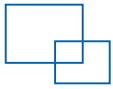fety (misinformation and deepfake), resulting in public harm and loss of trust. Moreover, gaps remain in the application of existing laws to AI-related harms and risks. For instance, certain provisions of the Information Technology (IT) Act, 2000, such as Section 66D (cheating by personation using a computer resource) and Sections 67, 67A, and 67B (relating to obscene or sexually explicit content) apply only to limited categories of deepfakes. Similarly, Section 356 of the Bhartiya Nyaya Sanhita (BNS), 2023, which addresses defamation, may be invoked where a deepfake harms an individual's reputation, but only after the harm has occurred and been reported, highlighting the largely reactive nature of existing laws.

However, as per the AI governance guideline report, these gaps can be addressed through sector-specific guidelines and targeted amendments to existing laws, rather than through the enactment of a separate, standalone AI law. But, at the same time, there is a growing need for a life-cycle based AI governance framework to enable effective implementation of existing laws/guidelines. India's vast socio-economic diversity, evolving digital infrastructure, and innovation-led growth further highlight the need of AI governance framework that reflects its unique requirements rather than mirroring external templates. In view of the contextual realities in the Indian ecosystem, a combinatorial approach with an appropriate balance of technical and legal instruments, collectively referred as "techno-legal" approach is proposed.

Definition: The techno-legal approach to AI governance could be defined as the integration of legal instruments, rule-based conditioning, regulatory oversight and technical enforcement mechanisms embedded with the technical architecture by design. This approach ensures that governance

is not merely a set of external constraints (or post-facto rules) but an intrinsic feature of any AI system, adaptable to evolving risks and contexts.

The techno-legal approach not only supports responsible innovation but also promises that AI technologies, irrespective of whether they are developed domestically or sourced from abroad, are aligned with the technical, legal, and ethical norms of the country. It is identified as a potential governance model that is transparent (for accountability and IPR compliance), explainable (in terms of performance and protection), provable (with reference to technical safeguards) and enabling (towards unlocking data and AI for innovation), making it contextually relevant, and aligned with India's constitutional values and developmental priorities.

The foundation of techno-legal regulation starts with:

Why: To meet the fundamental rights of citizens enshrined in the constitution to live with privacy, security, safety, access to fair information and earn for their work in the digital and AI era.

What: To ensure that the AI systems are trained, developed, deployed, and used in a way that protects citizens' privacy, security, safety, and ensures fair treatment, which are the primary attribute of Safe and Trusted AI.

How: To ensure that primary attributes of safe and trusted AI are respected and followed, there should be technical safeguards and governance mechanisms (which may be non-technical) that together can help ascertain transparency, accountability, explainability, provability, and the enabling nature of the AI system across its lifecycle.

Lifecycle Stages: To ensure that the primary attributes of safe and trusted AI are complied with across all five stages of the AI lifecycle, namely : (1) data collection (which includes conception stage and covers all types of data, extending beyond personal data), (2) data in use protection (by technical and non-technical measures); (3) AI training & model assessment, (4) Safe AI inference cycle (includes responsible AI implementation) , and (5) trusted agents (which includes discriminative AI systems, generative AI systems and agentic AI), a techno-legal framework may be put in place [18].
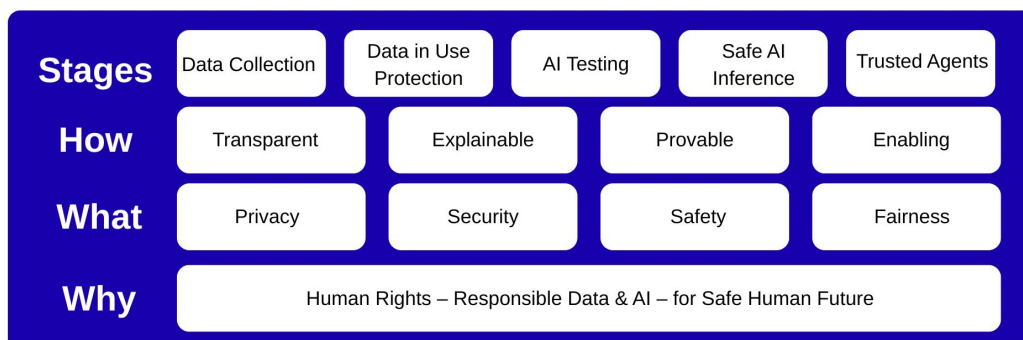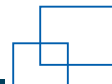
Figure 1: Founding pillars of Techno-legal Regulation (Source: Presented during roundtable discussion on "Techno-Legal Regulation for Responsible, Innovation-Aligned AI Governance" held on 22nd December 2025 at the Office of PSA, New Delhi)

Prioritizing techno-legal governance can strengthen India's AI governance framework in four ways:

01	scale and consistency of enforcement through standardized, automated checks
02	measurable accountability via logs, attestations, and audit trails
03	inclusive, low-cost compliance that leverages Digital Public Infrastructure (DPI) to support smaller firms and public agencies
04	future-readiness, allowing rapid calibration in response to evolving risks or model updates

2. Understanding Techno-legal Approach to AI Governance

Techno-legal approach has a potential to unlock value of technological-innovations for India while ensuring safety and responsible adoption of AI towards the country's population. It is a framework to use technology the right way with right controls to achieve the dual objectives of "innovation and safety" rather than "innovation vs safety." Successful examples of this approach can be seen in India's Digital Personal Data Protection (DPDP) Act 2023, which has matured over a period and in RBI's FREE-AI recommendations, Nov 2025 [36]. Rather than legal and technical requirements operating in silos, the regulatory expectation is to prevent harm to people (and society) with technical safeguards as a legal obligation across the AI life cycle, while not compromising on the innovation potential.

The formulation and implementation of a techno-legal AI governance framework, in addition to ecosystem participants, would primarily engage sectoral regulators and regulated entities required to comply with applicable regulations. The roles and sequencing described below are illustrative and reflect a flow in a sectoral regulatory context, where a regulator issues guidelines or requirements, based on applicable legal instruments, and regulated entities within that sector operationalise them through internal governance, processes, and technical controls. The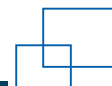 specific steps, actors, and obligations may vary by sector, use case, and applicable legal instrument. Accordingly, a techno-legal approach would broadly constitute the following components/stages with their roles defined as

(a) Law (an applicable legal instrument) to focus on enabling innovation while balancing control with technical safeguards grounded on legality. Prevention of harm to people/society proactively with due process and technical controls to be prioritised and incentivised

(b) Rules, framed under a Law to provide a clearer direction to the industry on the obligations for process and technical safeguards across various scenarios, requirements for exemptions and compliance across the AI lifecycle,

(c) To adapt to the evolution of AI, the techno-legal approach also leaves space for interpretation (as implementation may depend on the size of an organization or to a

particular situation) and for the creation of necessary frameworks and standards.

(d) Regulatory mechanism to be constituted to provide guidance, hear grievances and pronounce decision on complaints, (e.g. the Data Protection Board envisaged under the Digital Personal Data Protection Act.) This role could also be performed by the courts for laws whose enforcement case is before the courts.

(e) Techno-legal approach also to be operated at an organizational level. The Board of Directors of an organization have the ethical responsibility of taking care of their customers, serving them safely and responsibly. For this, they have to construct organizational structure that aligns with the laid-out regulation.

(f) The organization structure may include a Techno-legal governance team, or any other team within an organization responsible for AI initiatives. This team's responsibility is to protect the organization (e.g. data fiduciary)'s data and to manage the risks of AI by putting in place policies, frameworks, practices, processes, and technical controls. Technical safeguards must be taken care of by this team.

(g) Technical and Policy teams to take input from the techno-legal governance team and ensure that the policies, frameworks, practices, processes, and technical controls are implemented properly across the lifecycle.

(h) For clarity, protection of privacy is only one control. The technical controls should be conceptualized by first identifying the Responsible AI attributes to be achieved (such as safety, transparency, and accountability) and then designing relevant controls to implement those features
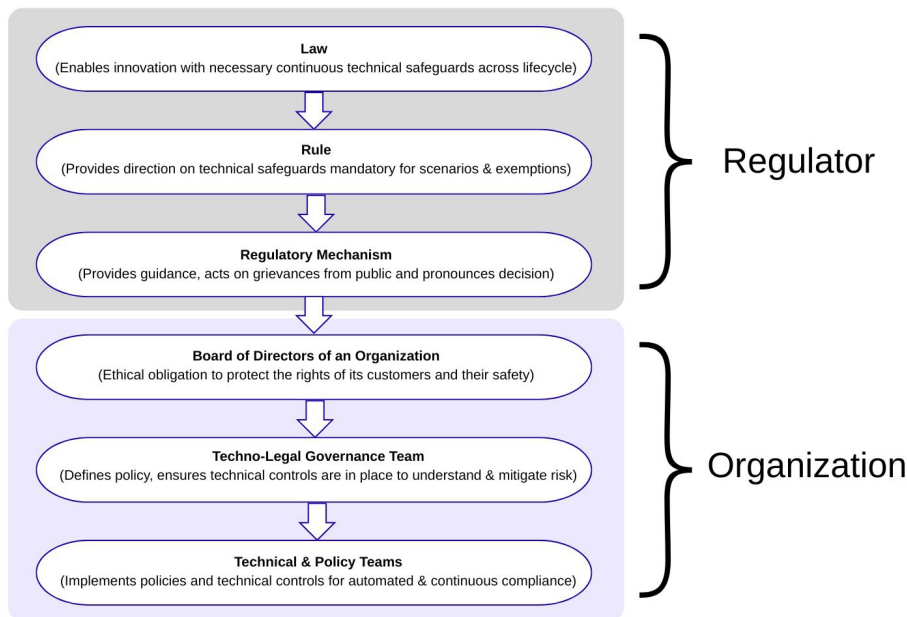
Figure 2: An indicative diagram for techno-legal governance mechanism

At the execution level, techno-legal framework may evolve from a law to rules to guidance to standards to protocols to various frameworks based on organizations and various levels of associated risks. Regulators and/or Courts (if applicable) could provide guidance on various attributes of Responsible AI and its mitigation depending upon the level of maturity of existing laws and obligations for each attribute.

In a nutshell, a techno-legal approach supplements the traditional regulatory mechanisms with modern governance tools and digital processes that embed compliance into technology. Instead of enforcing rules only after deployment, this approach encodes legal obligations directly into technical artefacts and system-level controls (such as consent-based access control, privacy-enhancing technologies, privacy-preserving machine learning, Data poisoning detection, data and AI threat modelling, data and AI impact assessment, LLM and agentic Guardrails [25][27]) directly at the design and development stage, allowing system controls to support compliance on an ongoing basis. This reduces the risk of gaps, misuse, and regulatory breaches during operation. When combined with human oversight, these measures enable compliance to occur at the point of data collection, model training, and deployment. By tracing and
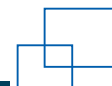
recording actions across the AI value chain, stakeholders can detect early risk, establish accountability, and scale monitoring through regulatory technology ("RegTech")[1] tools, such as real-time bias detection[2]. These technological tools, when integrated with legal aspects may support the evolution of robust techno-legal framework.

India's unique approach to Responsible AI	
Objective	Empowers innovation with governance and technical controls across the AI lifecycle, thus driving "Responsible AI by design" in a techno-legal way. This AI Governance approach is "law plus", including voluntary frameworks, to drive innovation.
Approach	Solution oriented and techno-legal approach to developing, deploying, and using AI across its lifecycle while mitigating risks, providing clarity for AI ecosystem.
Key Difference	Provides a path to unlock data and AI with right governance and technical controls to accelerate innovation. Preventing risks with governance and technical controls across the lifecycle is prioritized and incentivized.
Incentivisation	Larger deployments touching more citizens and /or deployments associated with higher perceived risks should have advanced levels of governance, transparency, and technical controls. (like stricter compliance defined for Significant Data Fiduciaries (SDF) in Digital Personal Data Protect (DPDP) Act). The scale of perceived risk of potential harm and not having "Responsible AI by Design" control can determine penalties (like as provisioned in DPDP Act).
Outcome	Balances innovation with risk mitigation using technical control and governance. Current laws are adhered to and utilized to full extent, and they are complemented with voluntary frameworks. This clarity for AI development can accelerate investment, startups and AI developers, enterprise adoption, and global leadership. "Responsible AI by Design" at population scale can result in global trust and adoption.

The importance of this approach lies in its ability of proactive identification and mitigation of risks before they manifest as harm; efficient oversight at both national and cross-border scales; clear allocation of responsibility among developers, deployers, and operators; and enhanced public trust through demonstrable privacy, security, fairness, and transparency. As emphasized in the India AI Governance Guidelines [13], rather than forcing compliance through regulation, the aim is to encourage voluntary guidelines and create a balanced model, which combines baseline legal requirements with tech-driven enforcement for a stronger and reliable governance framework.

3. Safe and Trusted AI Across the AI Lifecycle

Technology should be focused on solving real world problems while protecting citizens from harm across various dimensions or attributes of Responsible AI. While starting AI journey, an organization must clearly identify with an AI opportunity or an AI use case, which can unlock value for the organization and its users. Proper understanding of business, technical and legal requirements, value generated by the initiative and utility to the customer is essential. Along with it is equally important to understand the risk that may emerge at various stages of the AI life cycle and potential mitigation strategy for the same.

During execution at each stage, it is important to meet the AI development & deployment requirements along with necessary risk governance and control requirements with technical safeguards. This results in building "Responsible AI by Design" solutions which can be the foundation of compliance, trust, and global adoption. Model training

and Safe AI inference may include discriminative AI models and generative AI models. Many Deployment scenarios may stop at Safe AI inference and may not have agentic solutions, which is not mandatory.

Each stage of the lifecycle involves one or more organizations and pass on the result downstream. It is important for the downstream entities and AI deployers to ensure that they choose compliant upstream solutions and build on top of it in a compliant way.

I invite researchers, industry leaders, policymakers and citizens to read, reflect and engage. Together we can build an India where drone technology enhances prosperity, strengthens security and upholds democratic values.The following section outlines various risk types, methods for risk identification, potential impacts arising from negligence, and the corresponding indicative mitigation controls and benefits:
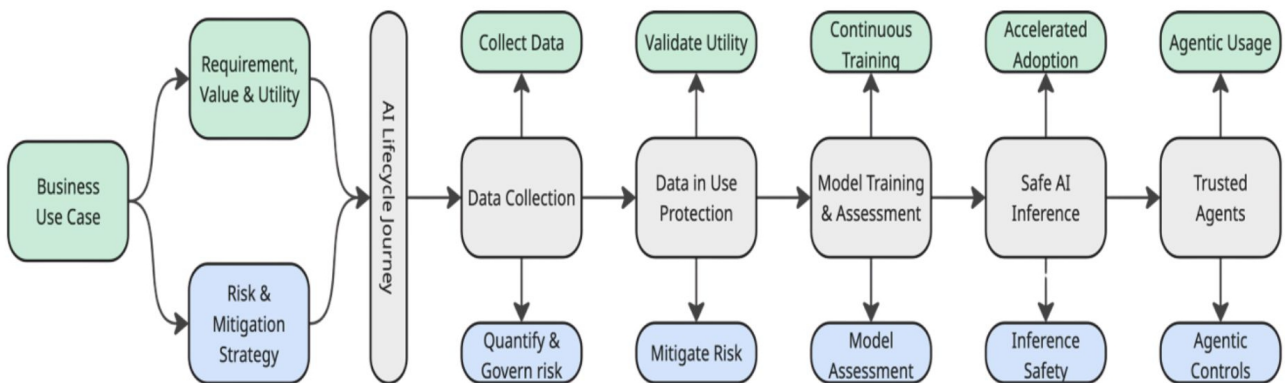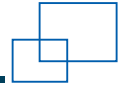
Figure 3: An indicative diagram of lifecycle journey of AI use-case

a. Data Collection Stage:

Risk: The data collection stage is foundational to building compliant, safe, and trustworthy AI systems. This follows the conception stage where an organization has clearly identified an AI opportunity or an AI use case, which can unlock value for the organization and its users. This also covers all types of data. Risks at this stage arise when data is collected without a clear understanding of its privacy, safety, intellectual property, security, and fairness implications.

- Privacy risk: Collection of personal, sensitive, or re-identifiable information without proper consent, lawful basis, or purpose limitation can lead to personal data breaches and regulatory non-compliance.

- Safety risk: Inclusion of malicious, harmful, deepfake, or inappropriate content in datasets can result in unsafe or harmful AI behaviors.

- Intellectual property risk: Inclusion of proprietary data, covered by intellectual property laws, can result in ambiguity regarding its use to build an AI system.

- Security risk: Poorly governed data sources may introduce risks such as data poisoning, credential leakage, or exposure of proprietary data through shared or third-party datasets.

- Fairness risk: Skewed, incomplete, or unrepresentative data collection can embed bias, resulting in discriminatory or unfair AI outcomes.

Potential Impact of Negligence:

- Non-compliance with data protection and AI regulations

- Increased likelihood of biased, unsafe, or untrustworthy AI systems

- Downstream risks that are difficult or costly to remediate post-

• training

Controls & Benefits:

• Data governance frameworks and risk assessment at ingestion

• Data Protection Impact Assessment (DPIA) and AI risk assessments prior to data onboarding

• Data classification, source validation, and consent verification

• Quantification of risk using techniques such as Privacy Threat Modeling [24] and documented recommendations before use

• Enables early risk prevention, transparency, and readiness for global AI deployment

b. Data in Use Protection Stage

Risk: Data in Use protection for model training is one of the most critical phases, which is high on penalty for non-compliance and is also irreversible post-training.

• Privacy risks: unauthorized access, over-retention, secondary use beyond original purpose of the training data, or lack of consent management

• Safety risks: uncontrolled use of inappropriate datasets in training pipelines.

• Security risks: weak access controls, lack of encryption, insider threats, and exposure of data to undesired parties, third-party tools, or external model APIs.

Potential Impact of Negligence:

• Regulatory non-compliance (DPDP, GDPR etc.)

• Breach notifications, penalties, and reputational damage

• Loss of trust in AI outputs due to opaque data handling

• Inadequate representation and biased decision-making

• Breach of intellectual property rights

Controls & Benefits: [20] [21] [23]

• Privacy-enhancing technologies (PETs) for controlling risk at the data in use and model training stage.

• PETs like Expert grade anonymization, Synthetic Data, Differential Privacy with mathematical guarantee of output privacy.

• Fairness, Security, and Safety-enhancing processing of data should be completed and verified before starting model training

• Data lineage, audit logs, and retention controls
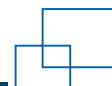
- Data Threat Modeling such as Privacy Threat Modeling, Data Fairness, safety, and security assessment, and DPIA/ AI Impact Assessment is to be approved before using data for model training

- Using digital contracts to standardize data-sharing terms, data privacy preserving techniques and enabling trusted execution environment as confidential computing/ Confidential Clean Rooms, to ensure appropriate security of sensitive datasets (finance, health, commerce) and isolation of models and rest of the systems.

- Enables provable compliance, reduces breach impact, and supports global AI deployment

c. AI Training & Model Assessment Stage

Risk: At the model training and evaluation stage, risks arise from model selection, training data quality, and evaluation rigor.

- Privacy risks: memorization of personal data and model inversion attacks.

- Safety risks: training on unsafe, malicious, or hallucination-prone datasets.

- Security risks: training-time attacks such as data poisoning or backdoor insertion.

- Fairness risks: emerge if bias is not measured and mitigated.

- Explainability risks: occur when opaque or overly complex models are deployed without justification.

Impact of negligence:

- Deployment of vulnerable, privacy-violating, biased, or unsafe models

- Regulatory exposure due to lack of transparency and documentation

- Reduced reliability and stakeholder confidence

Controls & Benefits: [22]

- Model risk assessment and benchmarking

- Scoring models for Privacy, Safety, Security, Fairness and Explainability (e.g., red-teaming, stress testing)

- Validating Transparency, Provability of compliance and Automation for enablement & compliance of model development process

- Enables informed model selection, transparency, and clarity of residual risk

- Do an AI impact assessment for model selection from a host of candidate models based on the red
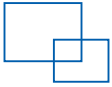
• teaming scoring results

• Clarity and transparency of model selection, resulting in accelerated AI adoption and trust

d. Safe AI Inference Stage

Risk: Responsible AI implementation remains relevant for this stage. During inference, AI systems interact with real users and live data, making this stage highly exposed.

• Privacy risks: unintended disclosure of sensitive information in responses.

• Safety risks: hallucinations, harmful outputs, and prompt-based manipulation.

• Security risks: prompt injection, model extraction, and adversarial attacks.

• Fairness risks: in real-time decision-making impacting individuals or groups.

Impact of negligence:

• User harm or misinformation

• Data leakage during inference

• Exploitation of AI systems by external actors

Controls & Benefits:

• Prompt & RAG (Red (high risk/urgent), Amber (medium risk/caution), and Green (low

risk/on track) level risk identification of usage for privacy, safety, security, and fairness

• Mitigation of risk with inline risk tagging and mitigation using pseudonymous inference or redaction or blocking

• Runtime monitoring, AI adversarial attack detection, and alerting

• Responsible AI firewall that works as a controller for Gen AI usage

• Ensures safe and controlled AI usage, reducing misuse, breach risks and accelerating adoption

e. Trusted Agents Stage

Risk: With the emergence of agentic AI, risks escalate as systems gain autonomy, tool access, and decision-making power. It is clarified that agentic AI systems may include individual AI agents, which may be discriminative AI models or generative AI models.

• Privacy risks: excessive data access and cross-system data leakage.

• Security risks: unauthorized actions, privilege escalation, and misuse of credentials.

• Safety risks: agents acting beyond the intended scope.

• Governance risks: arise if agent behavior is not observable or
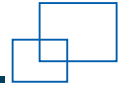
- auditable.

Outcome/Impact:

- Large-scale automated failures or misuse

- Compounded security and compliance breaches

- Loss of organizational control over AI actions

Controls & Benefits:

- Agent identity, authentication, and authorization along with trusted agentic attributes

- Firewalling based on context boundary and description-based usage

- Continuous monitoring, behavior logging, and kill switches

- Policy-based orchestration and governance frameworks

- Establishes a trusted ecosystem of AI agents that unlock value while remaining controlled and compliant.

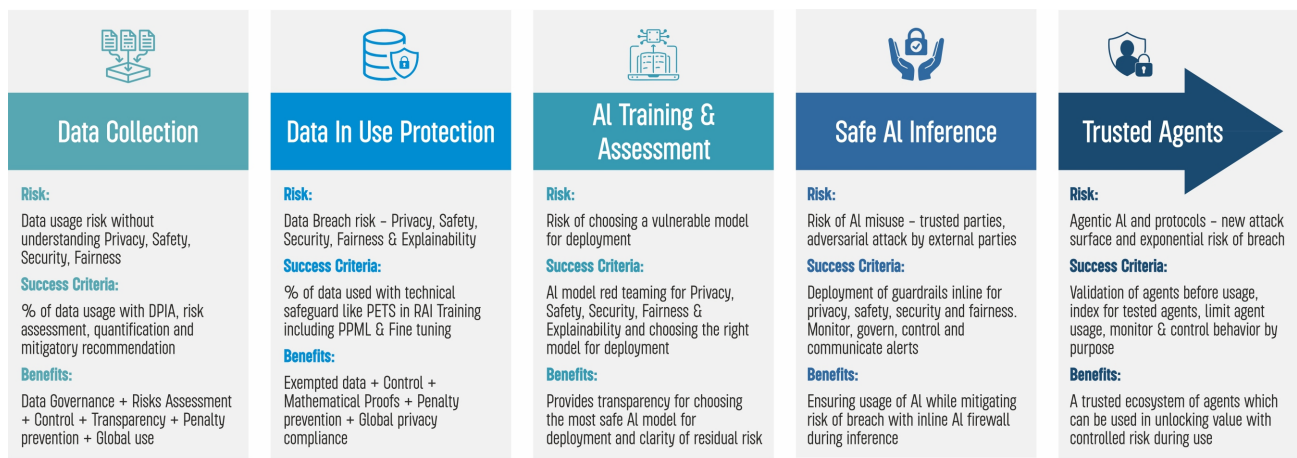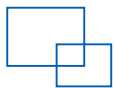| Data Collection	Data In Use Protection	AI Training & Assessment	Safe AI Inference	Trusted Agents
Risk: Data usage risk without understanding Privacy, Safety, Security, Fairness	**Risk:** Data Breach risk – Privacy, Safety, Security, Fairness & Explainability	**Risk:** Risk of choosing a vulnerable model for deployment	**Risk:** Risk of AI misuse – trusted parties, adversarial attack by external parties	**Risk:** Agentic AI and protocols – new attack surface and exponential risk of breach
Success Criteria: % of data usage with DPIA, risk assessment, quantification and mitigatory recommendation	**Success Criteria:** % of data used with technical safeguard like PETS in RAI Training including PPML & Fine tuning	**Success Criteria:** AI model red teaming for Privacy, Safety, Security, Fairness & Explainability and choosing the right model for deployment	**Success Criteria:** Deployment of guardrails inline for privacy, safety, security and fairness. Monitor, govern, control and communicate alerts	**Success Criteria:** Validation of agents before usage, index for tested agents, limit agent usage, monitor & control behavior by purpose
Benefits: Data Governance + Risks Assessment + Control + Transparency + Penalty prevention + Global use	**Benefits:** Exempted data + Control + Mathematical Proofs + Penalty prevention + Global privacy compliance	**Benefits:** Provides transparency for choosing the most safe AI model for deployment and clarity of residual risk	**Benefits:** Ensuring usage of AI while mitigating risk of breach with inline AI firewall during inference	**Benefits:** A trusted ecosystem of agents which can be used in unlocking value with controlled risk during use

*Figure 4: Safe and Trusted AI Across the AI Lifecycle*

4. Technological pathways to Techno-Legal AI Governance

The Government of India, through recent initiatives, has signaled growing consideration of a techno-legal approach through the IndiaAI Missions' "Safe & Trusted AI" pillar, which focuses on developing indigenous tools, frameworks, and mechanisms to embed ethics and safety in AI systems.

In 2024, MeitY launched a national call for proposals on "Responsible AI" to build practical solutions that could be adopted by both government and industry. Eight proposals were selected from Indian universities and partners, covering areas such as Machine Unlearning, Synthetic Data, Bias Mitigation, Explainability, Privacy-Enhancing Strategies, Algorithm Auditing, Governance Testing, and Ethical Certification [3]. The Machine Unlearning (IIT Jodhpur) proposal enables models to "forget" specific data influences, supporting lawful withdrawal, consent revocation, and data erasure requests under emerging data protection regimes [34]. Synthetic Data (IIT Roorkee) facilitates privacy-preserving model testing [35], while Bias Mitigation in Healthcare (NIT Raipur) integrates fairness checks into model

a) Government-Led Initiatives

validation. Explainable and Privacy-Preserving AI (DIAT Pune) and Privacy-Enhancing Strategies (IIT Delhi, IIT Dharwad, IIIT-Delhi, and TEC) advance lawful, secure, and interpretable AI development. Tools such as Nishpaksh and ParakhAI standardize fairness audits and participatory algorithmic evaluation, while Track-LLM introduces a governance testing framework for large language models [11][19].

Building on the success of the first call, MeitY's second Expression of Interest (EoI) [4] focused on enforcement-oriented tools, including mechanisms to distinguish synthetically generated media through clear disclosure and persistent identifiers, stress-testing frameworks, deep-fake detection, and risk assessment protocols. Five projects were selected: Saakshya (IIT Jodhpur & IIT Madras) and AI Vishleshak (IIT Mandi) for Audio-Visual deepfake detection and handwritten signature forgery detection, Bias Mitigation (Digital Futures Lab & karya) for evaluating Gender Bias in Agriculture LLMs-Creating Digital Public Goods (DPG) for Benchmarking and Fair
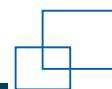
Data Work, and Anvil (Globals ITES Pvt Ltd & IIIT Dharwad) for Penetration Testing & Evaluation Tool for LLM and Generative AI. These projects represent important technology-driven efforts to promote the responsible development of AI [11][19].

b) Enabling Technologies for Techno-Legal Measures

Several technical measures and tools emerging from these initiatives demonstrate how legal and regulatory objectives may be operationalized through technology, particularly in addressing privacy and security breaches, content authentication, cybersecurity risks, and algorithmic bias:

a. AI Auditing Toolkits like Nishpaksh (for fairness auditing) and ParakhAI (for participatory algorithm auditing), highlight the recognized importance of the structured, enforceable measures to ensure fair and responsible development of AI in India, where the existing social diversities can potentially deepen due to the algorithmic bias in AI systems.

b. Bias mitigation techniques, spanning pre-training, in-training, and post-training interventions, reduce discriminatory outcomes. Recent causal-model-based approaches further strengthen bias mitigation by generating fair datasets that preserve sensitive features for auditing without allowing them to influence decisions [10].

c. Privacy-enhancing technologies [21] [23][25][26], such as k-anonymity, t-closeness, and differential privacy enable generalized and anonymized data sharing with mathematical guarantees. Techniques such as synthetic data generation and differentially private querying allow model training and data analysis without exposing sensitive information. Cryptographic measures including zero-knowledge proofs and secure multi-party computation facilitate collaboration among multiple entities or verification of computations without revealing underlying private data.

d. Advanced compute methods such as confidential computing with trusted execution environments, homomorphic encryption, and federated learning support secure

computation on sensitive or distributed data, enabling
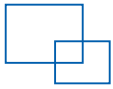
compliance with privacy and security requirements while maintaining analytical utility, thereby fostering responsible, privacy-preserving, and trustworthy AI deployment.[23]

e) Vulnerability scanning for AI algorithms, real-time anomaly detection, AI Red teaming, and secure model update mechanisms ensure cybersecurity and "secure-by-design" principle, also followed by the United States to embed enforceable security compliance, ensuring resilience against adversarial threats [12][25].

f) Prompt-level detection techniques such as tokenization, pseudonymous inference, or Role Based Access Control (RBAC) address privacy risks (personal information leakage), Safety concerns, Security threats (AI adversarial attacks, and confidential information breaches). Context-based firewalling provides an additional layer of protection by enforcing dynamic policies based on the sensitivity and context of the data, ensuring secure, safe, and privacy-preserving AI interactions [28][29][30].

g) Tools like content provenance [7] show potential of a techno-

legal strategy where regulatory obligations (e.g., mandatory labeling of AI-generated content) may be enforced through automated, scalable technical means. Additionally, other measures may serve as an adjunct technological measure, useful in certain regulated or high-trust environments, but not as the primary safeguard.

h) Embedding "responsible AI by design" principles into AI regulation could involve mandating an adversarial robustness toolbox [9] for AI models that are deployed in high-risk areas. Further, these technical measures can be integrated into techno-legal frameworks by mandating regular algorithmic audits by the developers of high-risk AI systems.

I) Agentic Red Teaming methods such as Agent Index and Facts Layer techniques track agents and their attributes for accountability and consistency, while Agentic Firewall enforces real-time control over agent behavior, context, and conversation to ensure secure and responsible interactions [31][32][33].

Furthermore, integrating techno-legal AI governance tools with India's DPI can significantly enhance their effectiveness and
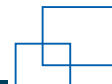
scalability. India's DPI, such as Aadhaar, Digi Locker, and Unified Payments Interface (UPI), provides a secure and interoperable foundation for embedding techno-legal AI governance directly into operational digital systems. DPI's core features: interoperability, consent-based data access, and auditability enable technical compliance tools to be integrated across the AI lifecycle rather than treated as external safeguards. Another key example is the Data Empowerment and Protection Architecture (DEPA), which may enable purpose-limited and consent-driven data sharing through digital contracts, privacy-preserving algorithms, and confidential clean rooms for AI model training [23]. When combined with other technical tools and features such as audit logging, bias evaluation, and machine unlearning, DEPA can function as a techno-legal enforcement layer, translating legal obligations into automated and verifiable technical controls.

Early adoption of DPI- and DEPA-integrated techno-legal approaches may potentially reduce long-term compliance costs and create strategic advantages for startups and enterprises. In addition, test-beds and pilots anchored in DPI ecosystems are essential to validate these tools under real-world conditions before deployment at scale. While integration of such process flow in data sharing and model training may initially require additional time and resources, it enables lawful and secure data access, safety-by-design, reduced regulatory risk, and increased global trust in Indian AI systems. [23]

At present, these technologies function primarily as governance enablers, and have significant potential to strengthen responsible AI deployment. Many tools are still evolving in terms of regulatory standardization, formal legal recognition, and evidentiary status, to independently generate verifiable compliance outcomes. Moreover, as adoption stands, further validation through real-world operational workflows will enhance insight into their scalability, accuracy, and institutional feasibility. This poses an important opportunity for collaborative experimentation between academic institutions, industry stakeholders, and regulators. Such efforts may focus on designing, developing, and deploying techno-legal tools that translate applicable legal requirements into concrete technical workflows capable of producing verifiable, auditable, and enforceable compliance outcomes.
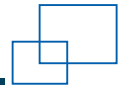
As aforementioned technologies
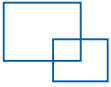
mature, the cost of compliance is likely to decline. However, successful adoption will depend on targeted capacity building, including executive awareness, clear mapping of legal duties to technical safeguards, lifecycle-level training, and organizational structures for AI risk management.

5. Operationalization of India's AI Governance Framework

The success of any policy instrument depends on its effective operationalization. Thus, alongside tool development, it is essential to strengthen the broader AI governance ecosystem comprising Industry, Academia, Government, AI model developers, deployers and AI-users. It is equally essential to identify gaps in the current ecosystem and address them periodically through updates to extent: rules,
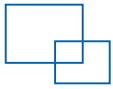
regulations, and guidelines. To achieve this goal, the AI governance guideline report has proposed institutional mechanism such as AI Governance Group (AIGG), Technology & Policy Expert Committee, AI safety Institute, AI Incident Database, and Voluntary frameworks and Vountary Industry commitments, while strongly advocating for the whole-of-the government approach.

The AI Governance Group (AIGG), chaired by the Principal Scientific Advisor to the government of India and comprising representatives from various government ministries, regulators and policy advisory bodies, will enhance coordination among ministries and regulators, addressing the current fragmentation in governance and operational processes. Within the techno-legal governance context, such coordination will help establish uniform standards for responsible AI

I)AI Governance Group (AIGG)

principles, which are explainable, provable, and enabling across sectors. In addition, the AIGG will play a key role in operationalizing the Techno-Legal Governance Framework by:

- Promoting responsible AI innovation and the beneficial deployment of AI in key sectors.

- Studying emerging AI risks, identifying regulatory gaps, and recommending necessary legal amendments.

II) Technology and Policy Expert Committee (TPEC)

To support AIGG, a dedicated TPEC within the nodal ministry (MeitY) is proposed. It would pool multidisciplinary expertise from areas such as Law, Public Policy, Machine Learning, AI safety, cybersecurity, and public administration etc. The TPEC will assist AIGG on matters of national importance in relation to AI policy and governance, including global developments in AI policy and Governance, potential risks and regulatory gaps, and new and emerging capabilities of AI. It will also serve as a link between government, industry, and academia to ensure that tested and reliable methods are adopted in practice.

III) AI Safety Institute (AISI)

The AI Safety Institute (AISI) will act as the primary centre for evaluating, testing, and ensuring the safety of AI systems deployed across sectors. It will also support the IndiaAI mission to develop techno-legal tools to address issues related to content authentication, cybersecurity, bias, etc. Key Functions of the AISI are proposed as follows:

- Research and Development of Safety Tools

- Evaluation of High-Risk Systems

- Development of Safety Frameworks and Toolkits

- Promoting Voluntary Compliance

- Capacity Building and Training

- Global Engagement

The AISI will assist both the AIGG and TPEC by generating risk reports, system evaluations, and compliance reviews that inform policy decisions and regulatory priorities. It will also facilitate cross-border collaboration on safe AI by coordinating with global safety institutes, standards-setting bodies, and international organizations.
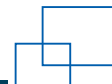
IV) National database of 'AI Incidents'

A national AI Incident Database will keep the records, classify, and analyze AI-related risks and incidents across India, such as safety failures, biased outcomes, security breaches, and misuse. Such reports will be submitted by public bodies, private entities, researchers, and civil society organizations. Within techno-legal framework, this database will enable post-deployment monitoring and accountability through measures such as:

- India-specific risk taxonomy;

- Detection of systemic trends and emerging threats;

- Data-driven audits and targeted regulatory interventions; and

- Evidence-based refinement of technical and legal controls.

This national-level database will draw on global best practices, such as the OECD AI Incident Monitor [14] and its associated framework [17]., but will be adapted to fit India's sectoral realities and governance structures.

VI) Voluntary Commitments and Self-Regulation

Industry-led voluntary practices can play an important role in strengthening the techno-legal framework alongside formal regulations. Such practices may include publishing transparency reports, conducting regular fairness and robustness testing, performing security reviews, and carrying out red-teaming exercises. These steps allow organizations to build familiarity with compliance processes and documentation before requirements become mandatory. Voluntary measures also help identify risks early, improve preparedness, and develop sector-wide capacity for responsible AI deploy-ment. Drawing on examples such as the OECD Framework for the Classification of AI Systems [5] and the EU AI Pact [6], the governance guidelines report encourages voluntary adoption by offering financial, technical, and regulatory incentives to organizations that demonstrate leadership in responsible AI practices. The voluntary industry commitments with provided incentives will promote early adoption of responsible practices and demonstrate accountability even before formal regulations are in place.

Through these measures, the emphasis is on coordination, consistency, and continuous learning and innovation. Coordination through AIGG and TPEC will prevent siloed approaches, consistency in standard and processes to give clarity to businesses and regulators, and continuous learning through AI incident reporting. AISI will provide requisite technical support to keep the framework adaptive to the evolving AI Risks in alignment with national priorities and global norms.
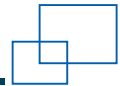
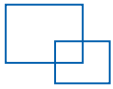# 6. Considerations for Developing Techno-Legal Tools and Framework

wide approach involving different stakeholders across the AI ecosystem and extend well beyond the deployment of technological tools and institutional design. To make this a success, the government may aim to incentivize voluntary adoption through financial, reputational, and regulatory benefits for organizations demonstrating leadership in responsible AI. While this approach offers a strong and promising pathway for AI governance, the continued evolution of India's AI ecosystem and its tools and frameworks bring into focus several critical cross-cutting considerations:

I. A trade-off between privacy and model performance:

A techno-legal framework must carefully balance privacy safeguards with inclusion, equity, and model utility considerations. In practice, privacy, fairness, and model performance often exist in structural tension, strengthening one may weaken another. For instance, to protect privacy, while individuals may legitimately seek the erasure or withdrawal of their data from model training, large-scale or uncoordinated removals can undermine model performance for underrepresented linguistic or

cultural groups. This could potentially lead to unintended demographic exclusions, particularly in a highly diverse nation like India. Conversely, meaningful fairness evaluation may require the use of sensitive attributes, creating additional privacy risks.

Given India's linguistic and demographic diversity, such trade-offs should not be left to implicit engineering decisions. To mitigate these risks, impact-aware data withdrawal mechanisms are needed rather than unconditional erasure. Large-scale unlearning requests should be subject to fairness and representativeness impact assessments, with safeguards triggered where data removal risks degrading performance for underrepresented groups. These trade-offs should be explicitly documented through instruments such as model cards, subjected to stakeholder consultation for high-impact deployments, guided by regulator-defined acceptable boundaries based on risk categories, and periodically revisited as AI systems and use contexts evolve. In this context, the standard setting bodies such as Bureau of Indian Standards (BIS) and Telecommunication Engineering Centre (TEC) with the

recommendations of AIGG and TPEC can develop and issue standards to ensure that privacy protection is implemented without demographic, linguistic and cultural exclusion.

## II. Governance for AI-subject Centric Applications:

It is also important to consider that operationalization of AI governance depends on distinguishing between AI users and AI subjects. Users are the people or organizations that directly use an AI system and can choose how to use it or agree to its terms. Subjects are people about whom AI systems make decisions, often without their knowledge or any meaningful ability to contest outcomes. In India, in many welfare schemes like healthcare, education, and public safety, the most affected people are subjects, not users. The experience built from deploying DPI ecosystem such as Aadhar which distinguish between enrollment agencies (users) and beneficiaries (subjects) may be leveraged for AI governance through techno-legal approach. Moreover, Techno-legal approach can mitigate this by pre-deployment algorithmic impact assessments, clear and proactive disclosure when AI is used in decision-making, human-in-the-loop mechanism at critical points in AI lifecycle, grievance redressal mechanisms and regular audits to assess

and address disparate impacts across demographic groups. The AIGG and TPEC may act as a body responsible for setting clear governance requirements for subject-facing AI applications. The AISI may function as the technical anchor by developing the evaluation methods and test procedures required to operationalize these requirements.

III. Special Consideration for Deepfakes:

In the case of deepfakes, it is recognized that content-level takedowns are structurally insufficient. Deepfake abuse operates through a pipeline, local generation tools, cloud platforms, and bots for amplification, and enabling infrastructure such as computing providers, payments, and model repositories, allowing rapid re-upload and domain migration. A techno-legal approach through content provenance mechanisms, including mandatory disclosure, persistent identifiers, and cryptographic provenance metadata at the point of generation and distribution may mitigate the harms. When combined with infrastructure-level obligations, such as usage logging, repeat-offender detection, and coordinated incident reporting across platforms, and adaptive updates as adaptive techniques evolve, these measures enable early identification, traceability, and disruption of deepfake
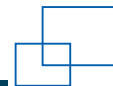
pipelines.

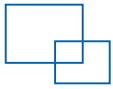## IV. Cross-Border Challenge: Need for Global Alignment

AI models are developed and deployed across multiple jurisdictions, legal standards, and enforcement mechanisms. This creates a challenge - a model trained or governed under one country's framework may not incorporate the safeguards required by another. For India, this means that even if a strong framework is established domestically, many globally accessed AI systems may not embed these protections. To tackle the issue of global alignment, it is essential to identify core features of AI systems that are significant at the global level, and where meaningful convergence is feasible. These may include privacy, security, safety, reliability, explainability or understandability, transparency, accountability, and inclusivity or non-discrimination. Once such features are identified, alignment can be supported by developing techno-legal tools to operationalize these features, alongside parallelly setting rules, laws, standards and frameworks. In this context, a techno-legal approach becomes particularly relevant, as it enables legal requirements to be translated into system-level technical controls that can function across borders. Institutions such as the AI Safety

Institute, through its network of global safety institutes, can further inform AIGG by identifying alignment between national AI governance frameworks and emerging international norms.

V. Building Capacity for AI Governance

Integrating a comprehensive techno-legal framework entails significant economic and capacity-building costs. For the private sector, these include the costs of compliance, such as conducting audits, implementing new security measures, hiring specialized legal and technical talent, and potentially investing in local data infrastructure. For the government, costs include administrative costs for the Committee, Secretariat, AI Incident Database, and building the necessary expertise within regulatory bodies to conduct effective oversight. Several techno-legal measures may require substantial computing and technical capacity, which can raise feasibility and access constraints, particularly for smaller firms and public agencies; therefore, it is important to build capacity, end-user literacy, and technical standardization to support equitable implementation. While these challenges are substantial, they must be weighed against the much higher costs of inaction. The failure to govern AI responsibly could lead to

widespread economic and social harm, including systemic discrimination, large-scale data breaches, and erosion of public trust in institutions. Therefore, it is a strategic choice to secure both economic growth and public trust in AI. Academia, and industry associations can jointly invest in capacity-building initiatives such as interdisciplinary training, open tools for risk assessment, and shared testing environments to strengthen responsible AI deployment.

VI. Balancing Compliance with Flexibility

The benefit of embedding law into technology is that compliance will become easier. But it should also be recognized that technology can support compliance only to the extent that legal obligations, and set benchmarks are integrated into the technological tools and would ignore any aspect that has not been integrated into the technology. Thus, it is prudent that a parallel techno-legal framing, incorporating laws, rules, standards, alongside technological tools development should be done to retain flexibility in compliance.

VII. Balancing Compliance with Cost and Accuracy

Experts have observed that the introduction of techno-legal mea-sures can have system-wide implications, affecting the performance of AI systems, increasing training requirements, and raising concerns about the cost and resource efficiency of AI systems. While such measures can enhance trust, safety, and fairness, their adoption should be proportionate to the level of risk, impact or harm involved and preserve system accuracy. Testing these trade-offs through testbeds and sandboxes is vital to calibrate an appropriate balance based on risk and context.

## VIII. Legal Clarity and Operational Alignment

Evolving laws around data protection, intellectual property, and AI governance directly shape how techno-legal mechanisms are implemented. The ecosystem can contribute by translating emerging legal standards into operational practices through compliance-by-design approaches, shared model documentation templates, and automated audit systems that support transparency and trust. Operationalization also needs to be grounded in India-specific evaluation, since many global benchmarks are calibrated to Western datasets and do not adequately capture Indian conditions, such as multilingual usage, local accents, and skin-tone sensitivity in vision systems.
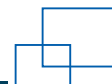
References

1. World Economic Forum, (21 June 2022). "What Is RegTech and What Does It Mean for Policymakers?" *World Economic Forum*.
2. Lemieux, F., A. Behr, C. Kellermann-Bryant, and Z. Mohammed, (1 March 2025) "Cognitive Bias Detection Using Advanced Prompt Engineering." *arXiv*.
3. Ministry of Electronics and Information Technology. (22 December 2023). "Call for expression of interest on Responsible AI". Innovate India – MyGov.
4. IndiaAI. (10 December 2024). "Expression of interest for Safe & Trusted AI projects under IndiaAI Mission."
5. OECD. "OECD Framework for the Classification of AI Systems", (Feb 2022) OECD Digital Economy Papers, no. 323.
6. European Commission, (August 2024) "AI Pact." *Digital Strategy*.
7. Bureacã, E., and I. Aciobãniţei. "A Blockchain-Based Framework for Content Provenance and Authenticity." (June 2024) *16th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, Iasi, Romania, 2024, pp. 1–5. IEEE.
8. Reserve Bank of India, (13 Aug 2025) "Framework for Responsible and Ethical Enablement (FREE) of Artificial Intelligence in the Financial Sector".
9. Nicolae, Maria-Irina, et al, (15 November 2019) "Adversarial Robustness Toolbox v1.0.0". arXiv, arXiv:1807.01069.
10. González-Sendino, Rubén, Emilio Serrano, and Javier Bajo, (June 2024) "Mitigating Bias in Artificial Intelligence: Fair Data Generation via Causal Models for Transparent and Explainable Decision-Making. Future Generation Computer Systems," vol. 155, 2024, pp. 384–401.
11. Government of India, Ministry of Electronics and Information Technology, (2025). Establishment of AI hub in Andhra Pradesh. Inrajya sabha starred question no. *64.
12. The White House, (July 2025) 'Americas AI Action Plan.
13. Ministry of Electronics and Information Technology, Government of India, (November 2025) India AI Governance Guidelines: Enabling Safe and Trusted AI Innovation.
14. OECD.AI. Overview and Methodology of the AI Incidents and Hazards Monitor.
15. Ministry of Electronics and Information Technology. (22 October 2025). *Explanatory note:* Proposed amendments to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 in relation to synthetically generated information [Government explanatory note].
16. Government of India. (11 August 2023). The Digital Personal Data Protection Act, 2023 (No. 22 of 2023). Ministry of Law and Justice (Legislative Department).
17. Perset, K., L. Aranda, and B. Rispal (Feb 2025). Towards a Common Reporting Framework for AI Incidents. OECD Artificial Intelligence Papers, no. 34, Organisation for Economic Co-operation and Development.
18. NASSCOM (Nov 2024). The Developer's Playbook for Responsible AI in India.
19. IndiaAI. (October, 2024). Safe & Trusted AI

20. OECD (Jun 2025). Sharing trustworthy AI models with privacy-enhancing technologies.
21. OECD (Mar 2023). Emerging privacy-enhancing technologies.
22. OWASP (Jan 2025). Gen AI red Teaming Guide
23. *ProductNation*. (31 October 2025). Privacy in the Age of AI: New Frameworks for Data Collaboration-
24. NIST (February 12, 2020), –LINNDUN Privacy Threat Modeling Framework
25. NIST (Jan 2026). NIST AI RMF
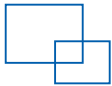
26. Article 29 Data Protection Working Party. (2014). Opinion 05/2014 on Anonymisation Techniques.

27. Department for Science, Innovation & Technology. (2023). UK Artificial Intelligence Regulation Impact Assessment.

28. OWASP Foundation. (n.d.). OWASP Top 10 for Large Language model Applications.

29. MITRE ATLASTM (n.d.)

30. National Cyber Security Centre (n.d.), UK. Guidelines for secure AI system development.

31. Raskar, R., Chari, P., Zinky, J., Lambe, M., Grogan, J. J., Wang, S., Ranjan, R., Singhal, R., Gupta, S., Lincourt, R., Bala, R., Joshi, A., Singh, A., Chopra, A., Stripelis, D., B, B., Kumar, S., & Gorskikh, M. (2025, July 18). Beyond DNS: unlocking the internet of AI agents via the NANDA index and verified AgentFacts. *arXiv*.

32. CSA, (May 2025), Agentic AI Red Teaming Guide.

33. OWASP (2025, April 28). Agentic AI - OWASP lists threats and mitigations.

34. Pathak, S., Shreshtha, S., Singh, R., & Vatsa, M. (2025, July 29). Quantum-Inspired Audio Unlearning: towards Privacy-Preserving Voice Biometrics. arXiv.org.

35. Joshi, H. C., & Kumar, S. (2025). FairGenerate: Enhancing Fairness through Synthetic Data Generation and Two-Fold Biased Labels Removal. *ACM Transactions on Software Engineering and Methodology*, *35*(1), 1–42.

36. Reserve Bank of India. (2025, August 13). Framework for responsible and ethical enablement of artificial intelligence (FREE-AI
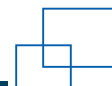
List of Abbreviations

AI	Artificial Intelligence
AIGG	AI Governance Group
AISI	AI Safety Institute
API	Application programming interface
BIS	Bureau of Indian Standards
BNS	Bhartiya Nyaya Sanhita
DEPA	Data Empowerment and Protection Architecture
DIAT	Defence Institute of Advanced Technology
DPDP	Digital Personal Data Protection
DPG	Digital Public Goods
DPI	Digital Public Infrastructure
DPIA	Data Protection Impact Assessments
EoI	Expression of Interest
EU	European Union
FREE	Framework for Responsible and Ethical Enablement
GDPR	General Data Protection Regulation
Gen AI	Generative AI
ICMR	Indian Council of Medical Research
IEC	International Electrotechnical Commission
IIIT	Indian Institute of Information Technology
IIT	Indian Institute of Technologies
IP	Intellectual Property
ISO	International Organization for Standardization
IT	Information Technology
ITES	Information Technology Enabled Services
LLM	Large Language Model
MeitY	Ministry of Electronics and Information Technology
NASSCOM	National Association of Software and Service Companies
NIT	National Institute of Technologies
OECD	Organisation for Economic Co-operation and Development
PET	Privacy-enhancing technologies
RAG	Retrieval-Augmented Generation
RBAC	Role Based Access Control
RBI	Reserve Bank of India
RegTech	Regulatory Technology
SDF	Significant Data Fiduciaries
SEBI	Securities and Exchange Board of India
TEC	Telecommunication Engineering Centre
TPEC	Technology and Policy Expert Committee
UPI	Unified Payments Interface

Office of the Principal Scientific Adviser
to the Government of India