FROM POLICY INTENT TO IMPLEMENTATION:

What Should India's AI Safety Institute Actually Do?

Devansh Gupta

Independent Policy Researcher

May 2025

Abstract

India is at a defining moment in its AI trajectory. Artificial intelligence is no longer an emerging technology on the horizon of Indian governance; it is already being explored and deployed across public service delivery, financial infrastructure, healthcare systems, and population-scale digital platforms. At the same time, India has set out an ambitious vision to become a global AI innovation leader through the IndiaAI Mission. Together, rapid deployment and national ambition create a difficult governance challenge.

This paper argues that India's central AI governance question has changed. The question is no longer whether India needs an AI Safety Institute. The IndiaAI Mission has already answered that. The real question is how such an institute should work in practice: what it should prioritise, what powers it should have, how it should relate to existing regulators, and how it can support innovation while protecting the public.

By examining global AI governance models, including the EU AI Act, the NIST framework, the UK and US AI Safety Institutes, and Singapore's AI Verify, this paper identifies what India can borrow and what it should avoid. Its core argument is that India's problem is distinct. Western safety institutions focus heavily on frontier model risks. India's more immediate challenge is to ensure that AI systems deployed in public services, financial systems, healthcare, and digital infrastructure work safely and fairly for hundreds of millions of people across many languages.

The paper proposes a hub-and-spoke model for the IndiaAI Safety Institute: a central body that sets standards, supported by sectoral and regional cells that handle evaluation, implementation, and local expertise. It recommends five first-priority functions: (1) a risk-tiered classification system for AI applications, (2) mandatory safety audits for high-impact public-sector AI, (3) Indian-language benchmarking for AI fairness, (4) a national AI incident reporting system, and (5) a Safety-as-a-Service program to reduce compliance barriers for startups and smaller organizations. The paper concludes that AI governance should not promise total control over AI systems. Its realistic goals should be harm reduction, institutional accountability, and public trust.

Keywords: AI governance, AI Safety Institute, IndiaAI Mission, Digital Public Infrastructure, multilingual AI, risk-based regulation, AI audits, India

Transparency and AI Assistance Disclosure

This working paper is a conceptual policy analysis based on public sources, comparative governance frameworks, and author-directed research. AI tools were used to assist with drafting, organisation, and language refinement. The author reviewed and revised the final text and is responsible for the paper's framing, claims, conclusions, and any remaining errors.

1. Introduction

Artificial intelligence is no longer a technology India is merely preparing to govern. It is already being explored, deployed, and governed unevenly in real time. AI-driven systems increasingly influence areas such as credit eligibility, hiring, healthcare triage, public welfare access, and online information environments. These are not only future scenarios. They are emerging governance problems that require accountability structures proportionate to the risks they carry.

The scale of this challenge is distinctive. India's digital ecosystem is not only large; it is complex in ways many AI governance frameworks were not built to address. The country's Digital Public Infrastructure (DPI), including Aadhaar, UPI, DigiLocker, and the Unified Health Interface, is among the world's most consequential state-backed digital systems. When AI is integrated into these systems, both benefit and harm can operate at population scale. A biased credit algorithm in India could affect very large numbers of people. A language model that performs poorly in Hindi, Tamil, or smaller regional languages can exclude large parts of the population.

In this context, the Government of India has committed to setting up an AI Safety Institute (AISI) under the Safe and Trusted AI pillar of the IndiaAI Mission. This is real progress. However, current policy documentation, including the India AI Governance Guidelines and the IndiaAI Mission framework, remains mostly at the level of principles. What is still missing is an operational blueprint: a clear and feasible design for what the AISI should do, how it should be structured, which systems it should prioritise, and how it should manage the tensions between safety, innovation, and inclusion.

This paper tries to answer those questions. It does not propose an ideal AI governance system in the abstract. It proposes a realistic and implementation-ready framework for India's AI Safety Institute, shaped by India's institutional context, resource constraints, and governance priorities.

1.1 Scope and Limitations

This paper focuses on the operational mandate and institutional design of the IndiaAI Safety Institute. It does not cover every dimension of AI policy in India. Questions of data protection, intellectual property in AI-generated content, and AI-related labour market transitions are important, but outside

its scope. The paper is a conceptual policy analysis based on publicly available sources and comparative governance models; it is not an empirical study. It also recognises that no governance framework can eliminate AI risks entirely. The pace of technological change, the resource constraints of the Indian state, and the tension between oversight and innovation are treated as real constraints that any workable proposal must address.

2. The Governance Problem: India's AI Trilemma

India is trying to do three things at once with AI — and all three pull in different directions. Before proposing any institutional design, it helps to name this tension clearly, because it shapes everything the AISI can and cannot realistically achieve.

2.1 The Innovation Imperative

India's AI ambitions are substantial and legitimate. The IndiaAI Mission represents a multi-billion rupee investment in compute infrastructure, AI datasets, sovereign AI capabilities, and startup ecosystem support. Becoming a competitive AI economy is strategically necessary in a world where AI capabilities increasingly shape economic and geopolitical power. But governance rules that are too costly or unpredictable can harm these goals, especially for startups and academic researchers.

This is not just theory. The European Union's experience with the AI Act has already demonstrated that compliance frameworks designed for large enterprises can impose disproportionate burdens on smaller actors, effectively concentrating AI development among a small number of well-resourced incumbents. India cannot afford a governance regime that produces this outcome.

2.2 The Safety Imperative

At the same time, AI without guardrails can cause real harm. India already faces concerns around deepfakes in political communication, algorithmic exclusion in welfare and financial systems, language-based inequity in digital services, and AI-enabled fraud. These examples should be understood as risk categories rather than fully documented case studies within this paper. A governance framework needs to address them without overstating what any single institution can solve.

The claim that market incentives and voluntary self-regulation will solve these risks is not supported by evidence. In many AI deployments, affected users do not understand the technology and have little practical means of redress. In those settings, market mechanisms do not provide enough protection. High-impact public-sector AI systems need structured oversight because their users usually have no market alternative.

2.3 The Inclusion Imperative

India's third challenge is one that most countries do not face at this scale: inclusion. India has 22 scheduled languages, hundreds of dialects, and massive variation in literacy and digital access. Many AI systems are strongest in English, reasonably capable in Hindi, and weaker in several other Indian languages. That is not a minor inconvenience. If an AI system mediates access to loans, welfare, or health information and performs worse for people using Odia, Kokborok, Santali, or other underrepresented languages, the result can be exclusion even without discriminatory intent.

No existing global AI governance framework addresses this dimension adequately. The EU AI Act's fairness requirements were designed mainly for European demographic contexts. The NIST AI RMF offers useful process guidance, but it does not fully address multilingual equity at India's scale. India's AI Safety Institute will therefore need to develop or adapt governance tools that are still underdeveloped globally.

2.4 Why the Trilemma Cannot Be Dissolved

These three imperatives are genuinely in tension. Stronger safety requirements increase compliance costs, which can slow innovation. Stronger inclusion requirements need expensive multilingual evaluation, which can burden smaller actors. And governance frameworks strong enough to constrain large platforms may still miss the many smaller systems where AI harms accumulate.

The right response is not to pretend this trilemma can be solved once and for all. The goal should be to design an institution that manages it intelligently: making tradeoffs explicit, building mechanisms for learning and adjustment, and being honest with the public about what it can and cannot achieve. Everything that follows is written with that in mind.

3. What India Can — and Cannot — Learn from Global Models

India's AI Safety Institute need not be designed from scratch. Several mature governance frameworks offer useful lessons. But each also has limits, which makes direct copying risky. This section examines the most relevant models and asks what India can realistically borrow from them.

3.1 The European Union AI Act: Risk Classification Without the Compliance Architecture

The EU AI Act came into force in August 2024 and is currently the most detailed AI law in the world. Its main idea is simple: not every AI system is equally dangerous, so they should not all face the same rules. It puts AI applications into four categories — unacceptable risk, high risk, limited risk, and minimal risk — and applies stricter rules as you go up the ladder. That logic is sound. Focus the heavy oversight where the real harm can happen.

India should borrow the EU's core idea: not all AI systems are equally risky, so not all should face the same level of oversight. But India should not copy the EU's full rulebook. The EU's model assumes companies have large legal teams, money for compliance, and regulators with enough technical capacity to review complex documentation. India does not have that capacity at scale yet. If India copies the EU approach too closely, companies may simply fill out forms without making their systems safer, while smaller startups struggle with costs that large companies can absorb. India would get the burden of regulation without the safety benefits.

3.2 The NIST AI Risk Management Framework: Process Discipline for an Indian Context

The NIST AI Risk Management Framework, published in January 2023, takes a very different approach from the EU. Instead of writing laws, it gives organisations a practical guide for thinking through AI risks. It breaks the process into four steps — Govern, Map, Measure, Manage — which any organisation can follow regardless of its size or sector. Think of it less as a rulebook and more as a structured checklist.

For India, the most useful thing about NIST is not the specific steps but the mindset: governance does not have to be one-size-fits-all. It can be flexible, practical, and scaled to what an organisation can actually handle. India's Safety-as-a-Service idea, covered in Section 6, borrows directly from this. The problem with NIST is that it is entirely voluntary. For AI systems that influence welfare access, credit decisions, healthcare triage, or other high-stakes outcomes, a voluntary checklist is not enough. India needs something stronger for those cases: mandatory rules at the top, flexible guidance below.

3.3 UK and US AI Safety Institutes: Technical Depth Without Deployment Focus

The UK and US AI Safety Institutes have built serious technical expertise in testing the most advanced AI systems in the world — the kind of large, powerful models that are on the cutting edge of what AI can do. They run adversarial tests, publish evaluation methods, and coordinate internationally, including through the 2023 Bletchley Declaration. This is valuable and India should learn from it.

India's AISI should build relationships with both institutions and develop similar technical evaluation capabilities over time. But the UK/US model has a major limitation as a template for India: it focuses mainly on frontier model safety. India's immediate challenge is different. It concerns the deployment of ordinary but consequential AI systems, such as automated decision tools, chatbots, translation services, and credit scoring algorithms, in high-stakes public contexts. A safety institute focused mainly on frontier AI would be well placed to assess future risks, but less prepared to address harms already occurring today.

3.4 Singapore's AI Verify: Practical Tools as a Starting Point

Singapore's AI Verify framework is probably the most directly useful reference point for how India should operate day-to-day. It is a practical testing toolkit — organisations use it to check their AI systems against basic principles like fairness, transparency, and reliability. It is open-source, modular, and designed so that even small teams without dedicated compliance staff can use it.

India should absolutely build on this for its Safety-as-a-Service program and startup sandboxes. The one big difference is scale. Singapore is a city-state. India has thousands of AI developers spread across 28 states, dozens of languages, and every sector imaginable. Singapore never needed regional offices or sector-specific cells. India does.

3.5 The Comparative Synthesis

EU AI Act	Risk-tiered legal structure	High compliance cost; requires mature institutions	Adopt risk classification logic; not the compliance architecture
NIST AI RMF	Scalable, modular process guidance	Voluntary; insufficient for high-impact systems	Use as baseline for SaaS toolkit and lower-risk systems
UK/US AISI	Deep technical evaluation capability	Frontier-model focused; not deployment-oriented	Long-term partnership model; not immediate template
Singapore AI Verify	Practical, accessible testing tools	Designed for small economy; not scalable as-is	Strong foundation for startup safety programs

4. India-Specific AI Risk Landscape

Any governance institution must be designed around the risks it is expected to manage. This section maps five categories of AI risk that are especially important for India, focusing on both the risks themselves and the structural conditions that make them difficult to govern.

4.1 AI in Digital Public Infrastructure: Population-Scale Deployment Risk

India's Digital Public Infrastructure is an extraordinary governance challenge. Aadhaar covers over 1.3 billion people; UPI processes billions of transactions monthly; and the Ayushman Bharat Digital Mission is building digital health infrastructure at national scale. When AI systems are integrated into this infrastructure for fraud detection, eligibility verification, document authentication, or service routing, errors and biases can have serious consequences.

The governance problem here is not only technical. It is also a question of accountability. If a fraud detection algorithm incorrectly flags a legitimate transaction, who is responsible for the error? What recourse does the affected user have? Who monitors for systematic bias across demographic or linguistic groups? In many contexts, the answers to these questions remain unclear, creating both legal uncertainty and practical risk for users who may have no meaningful alternative service.

The AISI should treat AI integrated into DPI systems as a distinct and high-priority governance category, requiring mandatory pre-deployment safety audits, human-in-the-loop oversight for consequential decisions, defined appeals mechanisms, and continuous post-deployment monitoring.

4.2 Multilingual Equity: The Invisible Exclusion Problem

India's linguistic diversity creates a class of AI risk that most global governance frameworks have not taken seriously enough. AI systems, especially large language models, voice assistants, chatbots, and translation tools, often perform better in English than in Indian languages, and better in Hindi than in smaller regional languages. This performance gap is not just inconvenient. If an AI system mediates access to financial services, healthcare information, or welfare benefits, weak performance in underrepresented languages can become a form of exclusion, even if no developer intended it.

The problem is made worse by the lack of evaluation infrastructure. There are no widely accepted benchmark datasets for assessing AI fairness and accuracy across India's 22 scheduled languages, and no standard methodology for testing performance in code-mixed contexts such as Hinglish or Tanglish. The AISI should treat multilingual evaluation infrastructure as a foundational function, not a supplementary one.

4.3 AI-Generated Misinformation and Deepfakes

India is highly exposed to AI-generated misinformation. High social media penetration, political polarisation, and uneven media literacy create conditions in which synthetic media can spread quickly and cause harm. The 2024 general election showed why detection, provenance, and response infrastructure matter, even though this paper does not attempt a full empirical assessment of election-related AI misuse.

The governance challenge here involves multiple actors — social media platforms, telecom providers, electoral authorities, and law enforcement — whose responsibilities need to be clarified and coordinated. The AISI is not meant to be the main regulator of online content. However, it has a distinctive role to play in developing technical standards for deepfake detection and content provenance, supporting research into detection methodology, maintaining an incident database, and providing technical support to electoral and law enforcement authorities.

4.4 AI-Enabled Financial Fraud

AI may lower the cost and increase the sophistication of financial fraud in India. Voice-cloning tools can imitate family members and solicit emergency transfers; AI-generated phishing messages can be harder to distinguish from legitimate bank communication; and automated social engineering can operate at scale. These risks sit at the intersection of AI governance, financial regulation, and cybersecurity, where current institutional arrangements need stronger coordination.

The AISI's role here is coordination rather than primary regulation. India's financial regulators — RBI, SEBI, and IRDAI — already have mandates that cover financial fraud. What is currently lacking is a shared technical language, common risk taxonomies, and coordinated incident reporting between AI governance bodies and financial regulators. The AISI should be designed explicitly to fill this coordination gap.

4.5 Startup Compliance Asymmetry

A less-discussed but important risk is that AI governance could become affordable only for large incumbents. If safety obligations require resources that only well-funded companies have, governance may entrench existing players and slow new competition. That would hurt both innovation and AI safety, since concentrated AI markets do not automatically produce safer systems.

India has a large and growing AI startup ecosystem, much of which operates with limited access to legal or compliance expertise. Any governance framework that does not address the needs of smaller organisations is likely to produce compliance asymmetry by default. The Safety-as-a-Service model proposed in Section 6 is designed to address this risk.

5. Proposed Institutional Design: The Hub-and-Spoke Model

The institutional design proposed here rests on a simple premise: India is too large, too diverse, and too resource-constrained for a single centralised AI governance body to be effective across every domain. A national body trying to evaluate AI systems in healthcare, finance, agriculture, criminal justice, and digital public infrastructure, across 22 languages and 36 states and union territories, would either spread itself too thin or focus only on a few high-visibility cases while most deployments remain ungoverned.

The hub-and-spoke model addresses this by separating functions that need centralisation from those that need local expertise. The hub sets standards, coordinates evaluations, and ensures public accountability. The spokes provide sectoral evaluation, regional testing, and startup support. The hub maintains coherence; the spokes provide reach.

5.1 The Central Hub: Standards, Coordination, and Accountability

The central IndiaAI Safety Institute should be constituted as an independent statutory body with a defined mandate, transparent governance structure, and public accountability mechanisms. Its core functions should include:

- Developing and maintaining national AI risk classification standards, with clear criteria for designating systems as high-impact, limited-impact, or minimal-impact
- Coordinating AI safety evaluations conducted by sectoral and regional cells, and maintaining national consistency in evaluation methodology
- Operating a national AI incident database, publishing periodic public reports on AI risks and harms, and communicating findings to relevant regulators
- Advising Parliament, ministries, and sector regulators on AI governance questions
- Representing India in international AI governance fora, including the AI Safety Institutes Network and relevant OECD and UNESCO working groups
- Managing the Safety-as-a-Service program and startup safety sandbox initiatives

The hub should be governed by a board that includes government, independent technical experts, civil society, and industry, with term limits and conflict-of-interest rules strong enough to reduce the risk of regulatory capture. Its annual reports should be tabled in Parliament, not merely submitted to the Ministry.

5.2 The Sectoral Spokes: Distributed Technical Expertise

Sectoral AI Safety Cells should be established in partnership with existing institutions — universities, research laboratories, sector regulators, and industry bodies — to conduct domain-specific AI evaluation work. Each cell should have defined evaluation responsibilities, operate under the hub's methodological standards, and report findings to the central institute.

Healthcare AI Cell	AIIMS, ICMR, NHA	Evaluate AI diagnostic tools, clinical decision support, and health data systems
Financial AI Cell	RBI, SEBI, IITs	Assess credit scoring algorithms, fraud detection systems, and robo-advisory tools
Education AI Cell	NCERT, IITs, State Universities	Review AI tutoring, assessment, and student-profiling systems
Public Services AI Cell	NIC, MeitY, State IT Departments	Audit AI in welfare delivery, document verification, and citizen services
Indian Language AI Cell	CIIL, IIITs, TDIL	Develop multilingual benchmarks and evaluate language model fairness
Cybersecurity AI Cell	CERT-In, IITs, DRDO	Monitor AI-enabled threats and evaluate defensive AI systems
Startup Safety Cell	iSPIRT, NASSCOM, Startup India	Provide Safety-as-a-Service support and manage regulatory sandboxes

5.3 Regional Cells: Addressing Linguistic and Geographic Diversity

In addition to sectoral cells, the AISI should establish regional cells in at least four zones: North, South, East, and West. These cells should evaluate AI systems in regional languages and work with state governments on AI governance. Without regional presence, AI governance will be weighted toward Hindi and English systems in major cities, while harms affecting speakers of smaller languages may go unseen.

6. Five First-Priority Functions

Institutional design is necessary but not sufficient. An AI Safety Institute without a clear operational agenda will spread itself too thin. This section proposes five first-priority functions that the AISI should aim to deliver within its first 18 months.

6.1 Function One: Risk-Tiered Classification of AI Applications

The AISI's first task should be to develop and publish a comprehensive risk classification framework for AI applications in India. This framework should draw on the EU AI Act's tiered approach while adapting the classification criteria to India's specific deployment context. The classification should be organised around three tiers:

- Tier 1 — High-Impact AI: Systems that make or materially influence consequential decisions affecting individual rights, welfare entitlements, financial access, healthcare outcomes, or public safety. These systems require mandatory pre-deployment safety audits, human oversight provisions, and incident reporting obligations.
- Tier 2 — Moderate-Impact AI: Systems that interact with users in ways that carry reputational, financial, or informational risks but do not directly determine consequential outcomes. These systems require adherence to AISI-published standards and participation in voluntary incident reporting.
- Tier 3 — Low-Impact AI: Systems with limited consequential impact on users. These systems are subject to light-touch oversight through AISI-published guidelines and no mandatory compliance obligations.

The classification framework should be published as a living document, reviewed annually, and accompanied by sector-specific guidance notes that help organisations understand where their specific AI applications fall.

6.2 Function Two: Mandatory Safety Audits for High-Impact AI

Tier 1 systems, especially those deployed by or with government agencies, should face mandatory safety audits before deployment and at defined intervals afterward. These audits should examine accuracy, reliability, fairness across demographic and linguistic groups, explainability, security, human oversight, and redress mechanisms.

Three design principles should guide the audit system. First, audits should be continuous rather than one-time checks. A system that passes before deployment can still become harmful as data, use patterns, or deployment contexts change. Second, audits should be conducted or verified by qualified third parties, not only by the deploying organisation. Self-certification is not enough for high-impact systems. Third, audit summaries for publicly funded AI systems should be disclosed, so civil society and affected communities can scrutinise them.

A critical acknowledgement: audits are imperfect. An audit shows that a system met defined safety criteria at a specific point in time. It does not guarantee safe performance in every future context. The AISI should say this clearly in public communications to avoid the false assurance that a passed audit can create.

6.3 Function Three: Indian-Language Benchmarking

No function of the AISI is more distinctively important than multilingual AI evaluation. Within its first year, the AISI should commission standardised benchmark datasets for assessing AI performance across India's scheduled languages, beginning with languages that have large speaker populations and high rates of digital service use.

These benchmarks should evaluate three dimensions: accuracy (does the AI system perform comparably across languages for equivalent tasks?), safety (does the system produce harmful or discriminatory outputs at different rates in different languages?), and fairness (does the system's performance vary systematically across demographic groups who communicate in different languages?). The benchmarks should be open-source and made freely available to developers, to both reduce evaluation costs and to create a common quality standard across the ecosystem.

This is genuinely hard. India's linguistic diversity — 22 scheduled languages, hundreds of dialects, pervasive code-mixing, and significant regional variation within languages — means that even a well-resourced benchmarking effort will take years to develop comprehensive coverage. The right approach is to be systematic, beginning with the highest-impact languages and extending coverage progressively, rather than to wait until complete coverage is achievable before beginning.

6.4 Function Four: National AI Incident Reporting System

India currently lacks a systematic way to record, analyse, and learn from AI-related harms. A welfare algorithm that wrongly excludes eligible beneficiaries, a deepfake that contributes to public disorder, or a credit scoring system that disadvantages a community may be reported to different authorities, or not reported at all. These examples illustrate the type of incident an AI incident reporting system should capture. Without such a system, India will struggle to identify patterns and design better governance responses.

The AISI should establish a national AI incident reporting system, modelled in part on aviation safety reporting, where structured and non-punitive reporting has improved safety over time. The system should be accessible to affected individuals, civil society organisations, developers, and government agencies. It should give reporters meaningful feedback about how reports are used, protect legitimate confidentiality interests, and avoid becoming a shield for organisations that want to conceal harmful incidents.

6.5 Function Five: Safety-as-a-Service for the Innovation Ecosystem

The Safety-as-a-Service (SaaS) program addresses the compliance asymmetry problem directly. The AISI should provide, free of charge or at heavily subsidised cost, a suite of safety evaluation tools and support services accessible to any organisation developing or deploying AI in India. These should include:

- Open-source bias testing tools, pre-configured for common AI application types
- Audit templates for self-assessment of AI systems against AISI standards
- Indian-language benchmark datasets for testing multilingual AI performance
- A regulatory sandbox program that allows startups to test innovative AI applications under AISI supervision, without full compliance obligations, before seeking market authorisation
- A dedicated advisory service providing compliance guidance to organisations without in-house AI governance expertise

The SaaS program should operate in genuine partnership with industry bodies, academic institutions, and civil society organisations. It should not be a government service delivered to passive recipients, but a shared infrastructure built with the ecosystem it serves. This would improve the quality of the tools and make organisations more likely to use them.

7. Structural Limits and Unavoidable Tradeoffs

This may be the most important section of the paper: what cannot be achieved, no matter how well the AISI is designed. Governance frameworks that are not honest about their limits tend to overclaim, create false assurance, and produce cynicism when they fall short. Being honest about limits is not pessimism. It is what makes governance credible.

7.1 Perfect AI Safety Is Not Achievable

No governance framework eliminates AI harm. AI systems will still produce wrong outputs, biased decisions, and harmful content even after passing every safety check — because the underlying problems are not fully solved yet. Fraudsters will always find new ways to misuse AI faster than detection tools can keep up. And systems that work fine in testing will still fail in the real world in ways nobody predicted. That is not a reason not to govern AI. It is a reason to be honest about what governance can and cannot do.

The right response to this is to design governance around realistic objectives — harm reduction, accountability, and the creation of institutional learning — rather than around the unachievable goal of total safety. An AI Safety Institute that claims to guarantee safe AI will lose public trust when AI harms inevitably occur. An AI Safety Institute that is honest about managing risk rather than eliminating it can maintain credibility even when the systems it oversees fail.

7.2 State Capacity Constraints Are Real and Will Not Be Quickly Resolved

Many Indian regulators simply do not yet have people who understand AI well enough to evaluate it properly. That is not a criticism — this is a new and fast-moving field. But it matters enormously for what governance can actually achieve right now. If you require mandatory audits before you have trained auditors, what you get is paperwork. Forms get filled. Boxes get ticked. Nothing actually gets safer.

This means the AISI's capacity-building function must be treated as a first-order priority. Training regulators, developing audit methods, and building expert networks cannot wait until after the governance framework is operational. Without this investment, the framework may look impressive while achieving little in practice.

7.3 The Political Misuse Risk

AI governance institutions are not immune to political misuse. An AI Safety Institute with broad powers to classify AI systems and impose compliance obligations could, under adverse political conditions, be used to disadvantage politically inconvenient media platforms, suppress AI tools used for government accountability, or justify surveillance infrastructure in the name of safety. These risks are real and have parallels in other regulatory domains.

The answer is not to weaken the AISI until it becomes ineffective. The answer is to build safeguards into its design: transparent decision-making, published criteria for major governance decisions, multi-stakeholder oversight with civil society representation, strong protections for whistleblowers and researchers who document AI harms, and annual reporting to Parliament rather than only to the executive branch.

7.4 The Tradeoffs Are Permanent, Not Transitional

It is tempting to think these tensions will sort themselves out once the institutions mature. They will not. A framework that takes safety seriously will always cost innovators something. A framework that prioritises speed will always leave some people unprotected. That is just the reality. The job of a well-designed institution is not to pretend these conflicts do not exist — it is to make the trade-offs visible, explain them clearly, and let the public hold decision-makers accountable for the choices made.

8. Implementation Roadmap

The roadmap below is ambitious but realistic. It prioritises the foundations of effective governance over symbolic early activity.

Phase 1: Foundations (Months 1–6)

- Establish the AISI's statutory basis and governance structure, including board composition, conflict-of-interest provisions, and accountability mechanisms
- Develop and publish the national AI risk classification framework, with sector-specific guidance notes for healthcare, finance, education, and public services
- Identify partner institutions for the first wave of sectoral and regional cells
- Begin training programs for regulators and government AI practitioners
- Launch public consultation on the AI incident reporting system design

Phase 2: Infrastructure Development (Months 7–18)

- Establish and begin operations of initial AISI Cells in healthcare, financial services, and public digital infrastructure
- Commission the first wave of Indian-language benchmark dataset development, covering the eight most widely-spoken scheduled languages
- Launch the national AI incident reporting system in pilot mode
- Publish the Safety-as-a-Service toolkit and open the first startup sandbox cohort
- Complete mandatory pre-deployment audits for the five highest-impact government AI systems

Phase 3: Maturation and Scale (Months 19–36)

- Expand sectoral cells to cover all designated high-impact sectors
- Publish first annual AI safety report, including incident database summary and audit findings
- Extend Indian-language benchmark coverage to all 22 scheduled languages
- Strengthen coordination protocols with sector regulators — RBI, IRDAI, SEBI, TRAI, and state health departments
- Begin structured engagement with international AI safety institutions on evaluation methodology and incident data sharing

9. Conclusion

India does not have only an AI problem. It has an institutions problem. The risks are real, the principles have been written down, and the political will appears genuine. What is missing is the hard, unglamorous work of building institutions that function: institutions that turn policy documents into oversight, respond when something goes wrong, and remain credible five years from now when the technology has changed.

The hub-and-spoke model proposed in this paper is not the only viable design for India's AI Safety Institute. But it reflects principles that any viable design should respect: India's governance challenges are too diverse for a single centralised institution to manage alone; compliance frameworks must be calibrated to organisational capacity, not just organisational size; inclusion and equity are governance priorities, not side issues; and honest acknowledgement of governance limits is a form of institutional strength.

If India's AISI succeeds, it will not eliminate AI risk. It will build systems that identify emerging harms early, create accountability when harms occur, protect citizens who have no market recourse against AI systems that affect their lives, and support responsible innovation by making safety infrastructure accessible rather than exclusive. These goals are achievable. They are also the ones that matter most.

The decisions made in the next 12–18 months will shape India's AI governance architecture for a generation. They deserve careful, serious attention.

References

Official and Institutional Sources

Government of India, Ministry of Electronics and Information Technology. India AI Governance Guidelines Development Report. Press Information Bureau, 2024. Available at: https://www.pib.gov.in/PressReleasePage.aspx?PRID=2228315

IndiaAI Mission. Safe and Trusted AI. MeitY, 2024. Available at: https://indiaai.gov.in/hub/safe-trusted-ai

IndiaAI Mission. Call for Partnerships as Part of the IndiaAI Safety Institute. MeitY, 2024. Available at: https://indiaai.gov.in/article/call-for-partnerships-as-part-of-the-indiaai-safety-institute

Office of the Principal Scientific Adviser to the Government of India. India Takes the Lead: Establishing the IndiaAI Safety Institute for Responsible AI Innovation. PSA, 2024. Available at: https://www.psa.gov.in

National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0). US Department of Commerce, 2023. Available at: https://www.nist.gov/itl/ai-risk-management-framework

National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework for Generative Artificial Intelligence (AI RMF Generative AI Profile). US Department of Commerce, 2024.

European Commission. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2024.

UK Department for Science, Innovation and Technology. AI Safety Institute: Overview and Objectives. HM Government, 2023. Available at: https://www.gov.uk/government/organisations/ai-safety-institute

Infocomm Media Development Authority (IMDA) and AI Verify Foundation. AI Verify Testing Framework and Toolkit. Singapore Government, 2023. Available at: https://www.imda.gov.sg

OECD. OECD Framework for the Classification of AI Systems. OECD Digital Economy Papers No. 323, 2022. Available at: https://oecd.ai

UNESCO. Recommendation on the Ethics of Artificial Intelligence. UNESCO, 2021. Available at: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Academic and Policy Research

Ada Lovelace Institute. Algorithmic Accountability for the Public Sector: Learning from the UK's Experience. Ada Lovelace Institute, 2022.

Bommasani, R. et al. On the Opportunities and Risks of Foundation Models. Stanford Center for Research on Foundation Models (CRFM), 2021.

Jobin, A., Ienca, M., & Vayena, E. The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence, 1(9), 389–399, 2019.

Raji, I.D., & Buolamwini, J. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. Proceedings of AAAI/ACM AIES, 2019.

Kak, A. (Ed.). Regulating Biometrics: Global Approaches and Urgent Questions. AI Now Institute, 2020.

Carnegie India. Governing India's Emerging AI Ecosystem: A Policy Landscape. Carnegie India, 2023.

Observer Research Foundation. Between Safe and Secure: What Next for India's AI Safety Institute? ORF Issue Brief, 2024.

Gurumurthy, A., & Bhatt, N. Towards an Inclusive Digital Economy: A Framework for Understanding Digital Divides in India. IT for Change, 2022.

Raso, F.A., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. Artificial Intelligence & Human Rights: Opportunities & Risks. Berkman Klein Center for Internet & Society, Harvard University, 2018.

About the Author

Devansh Gupta is an independent policy researcher focused on AI governance, digital public infrastructure, and technology policy in India. The views expressed in this paper are the author's own. AI tools were used to assist with drafting, organisation, and language refinement. The author reviewed and revised the final paper and is responsible for the decision to share it and for any remaining errors.