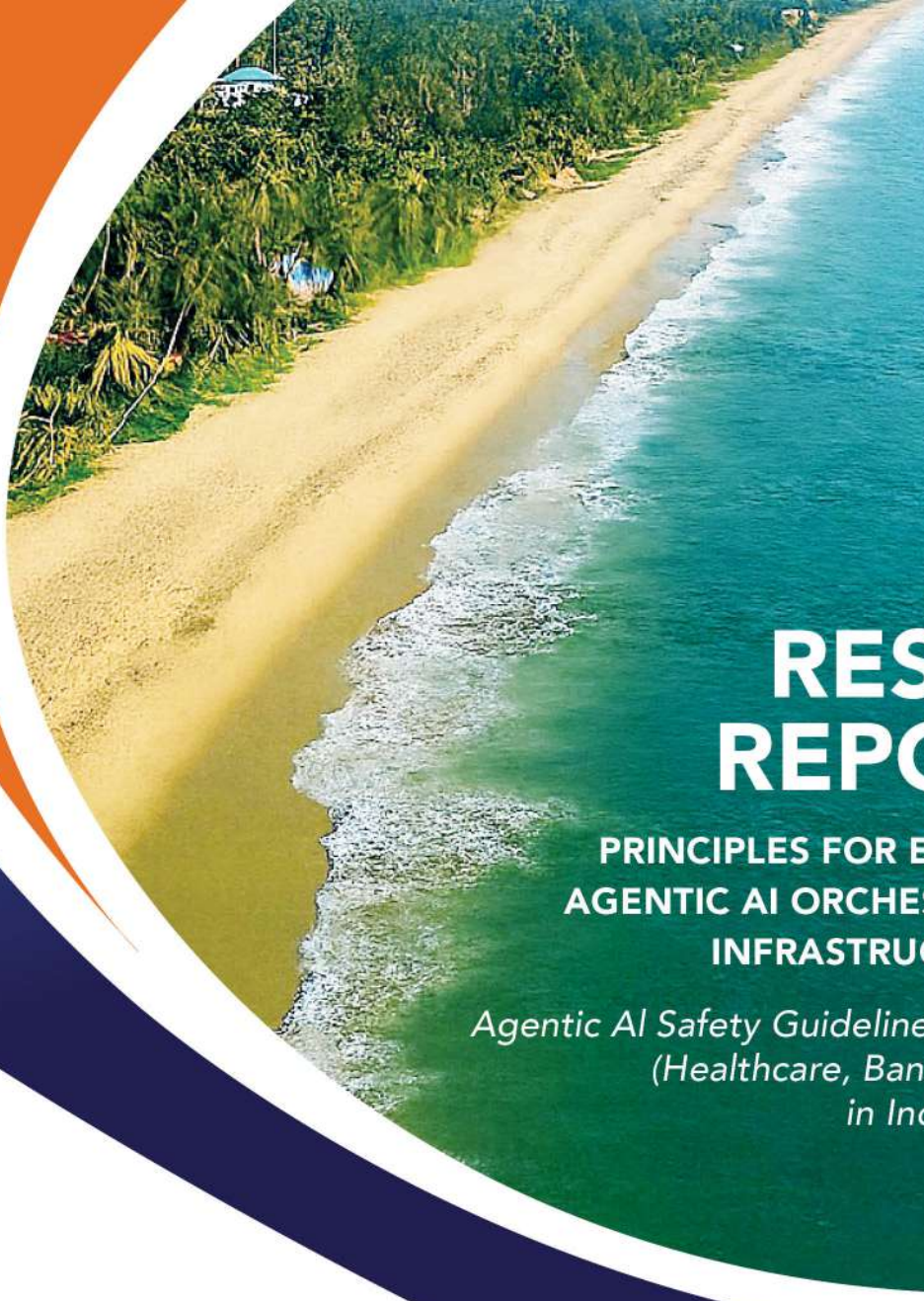


SAIGS 2025

Safe AI for the Global South

Empowering the Global South with safe and responsible AI ecosystems

RESEARCH REPORT ON

PRINCIPLES FOR ETHICAL AND SAFE AGENTIC AI ORCHESTRATION ACROSS INFRASTRUCTURE GRADIENTS

Agentic AI Safety Guidelines for Critical Sectors (Healthcare, Banking, Education etc.) in India and Global South

AI SAFETY RESEARCH LAB

Amrita Center for Cybersecurity Systems and Networks

AMRITA VISHWA VIDYAPEETHAM

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF UGC ACT, 1956

AI IMPACT SUMMIT भारत 2026 INDIA

Official Pre-Summit Event of the India AI Impact Summit 2026
Ministry of Electronics and IT, Government of India

सर्वजन हिताय | सर्वजन सुखाय
WELFARE FOR ALL | HAPPINESS OF ALL

MESSAGE FROM CHANCELLOR

Amma sees the world as a flower. Each country is a petal of the flower. Technology has truly revolutionized human life, but its negative aspects raise alarming concerns about the future of mankind. Scientific knowledge and spiritual wisdom have to complement one another. While Science air-conditions the outer world, spirituality air-conditions the inner world

Chancellor

Mata Amritanandamayi Devi (AMMA)

From the Editor's Desk

As AI moves from predictive models to autonomous "agentic" systems—capable of planning and acting independently—the transformative potential and serious safety challenges, particularly in a country like India with large infrastructural diversity, also come forward. This report collates research on a key concern: How can we ensure AI benefits reach everyone safely and equitably across urban and rural gradients? This analysis uncovers a deep "safety gap" between rapidly changing agentic AI capabilities and the static structures of governance developed to oversee them.

In situ AI, learning, and adapting in real time, are poorly served by conventional compliance mechanisms. This research report presents a more dynamic approach: agentic safety frameworks that are designed to be contextual, adaptable, and embedded in everyday, real life settings, based on applications such as healthcare, banking, education, and government services. The challenges we investigate herein are by no means unique to India.

Across the Global South, countries are navigating remarkably similar pressures as they move toward agentic AI—from gaps in computational infrastructure to the difficulty of translating high-level principles into enforceable regulatory practice. This report therefore positions itself not merely as guidance for India, but as a blueprint that emerging economies can adapt: a set of design patterns, governance primitives, and operational levers that make trustworthy AI both implementable and measurable.

To advance this ambition, the report introduces several forward-looking constructs. It proposes regulatory sandboxes with embedded telemetry, enabling governments to observe agent behaviour in controlled environments before deployment. It outlines a context-aware risk layer that modulates autonomy levels based on sectoral sensitivity, allowing systems in healthcare, defence, or finance to dynamically restrict or expand agentic action. It also recommends cultural alignment modules to account for local norms, languages, and societal values when AI agents operate across diverse communities.

Together, these additions complement broader global discussions by demonstrating how operational safety can be achieved—not just why it is important. They offer a practical pathway for India and other Global South nations to lead in shaping agentic AI ecosystems that are safe, culturally grounded, and capable of scaling responsibly.

This work represents the fruits of a deep inquiry and thorough, interdisciplinary vision by the AI Safety Research Lab, Center for Cybersecurity Systems & Networks at Amrita Vishwa Vidyapeetham to devise practical, principled approaches to keep intelligent systems accountable, fair, and aligned with human values. This report is in line with the national vision of "Making AI in India and Making AI work for India." It will support the transition from action to impact as we prepare for the India AI Impact Summit 2026.

We valuable support & commitment, and of course all the experts, frontline workers, and citizens who participated in our surveys and interviews; your inputs kept our work grounded in India's real-life-lived realities. Let us go ahead to create scientifically sound, ethically anchored, and socially inclusive AI systems together: academia, government, industry, and civil society. Shared stewardship can help make certain that AI's autonomy does not undermine accountability and that innovation serves the public good.

Dr. Sivaramakrishnan R. Guruvayur
Professor of Practice

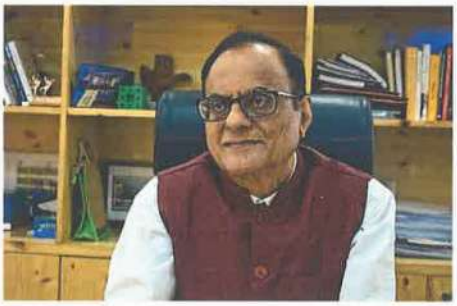
Dr. Krishnashree Achuthan
Professor & Dean

Dr. Vysakh Kani Kolil
Assistant Professor

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

सत्यमेव जयते

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 is rapidly shaping India's scientific and technological future, with significant implications for national development across healthcare, education, agriculture, financial systems, and public services. As AI progresses from predictive analytics to *agentic systems* capable of autonomous actions, the need for robust safety, governance, and ethical oversight becomes paramount.

The IndiaAI Mission's vision of *"Making AI in India and Making AI Work for India"* places strong emphasis on Safe and Trusted AI. As we expand national compute infrastructure, strengthen datasets, and build advanced research and innovation capacity, it is essential that our safety frameworks evolve concurrently to ensure responsible deployment across the country's diverse socio-technical contexts.

In this regard, the research report titled *"Principles for Ethical and Safe Agentic AI Orchestration Across Infrastructure Gradients"* offers an important and timely contribution. By examining governance requirements for agentic AI systems across varied infrastructure environments, the report addresses a critical dimension of AI adoption in India and the Global South. The work undertaken by the *AI Safety Research Lab at Amrita Vishwa Vidyapeetham, Amritapuri,* reflects the constructive collaboration between government, academia, and industry in advancing national priorities on Responsible AI.

As India prepares for the *India AI Impact Summit 2026*, the insights and recommendations presented in this report will support ongoing efforts to operationalize Safe and Trusted AI. Its focus on context-aware safety protocols, sectoral risk considerations, and scalable governance mechanisms will inform discussions under the Summit theme *"From Action to Impact."*

I appreciate the efforts of the authors and contributors who have undertaken this work. It is essential that India continues to lead in developing AI systems that are scientifically sound, ethically grounded, and aligned with public interest. This report reinforces our commitment to ensuring that AI technologies serve as secure, trustworthy, and equitable enablers of national progress.

(Ajay K Sood)

Date: 5th December 2025

Steering Committee

Chair

Dr. Krishnashree Achuthan
Dean, Amrita Vishwa Vidyapeetham;
Director, Center for Cybersecurity Systems and Networks

Co-chair

Dr. Sivaramakrishnan R. Guruvayur
Professor of Practice & Head,
AI Safety Research Lab, Center for Cybersecurity Systems and Networks,
Amrita Vishwa Vidyapeetham

Staff and Researchers

Research Manager and Editor-in-Chief

Dr. Sivaramakrishnan R. Guruvayur
Professor of Practice & Head,
AI Safety Research Lab, Center for Cybersecurity Systems and Networks,
Amrita Vishwa Vidyapeetham

Researchers

Dr. Vysakh Kani Kolil (Academic Lead) Assistant Professor,
Center for Cybersecurity Systems and Networks, Amrita Vishwa Vidyapeetham

Mrs. Akshara Ravi Assistant Professor, Center for Cybersecurity Systems and Networks,
Amrita Vishwa Vidyapeetham

Mrs. Pavithra S. P. Research Scholar, Center for Cybersecurity Systems and Networks, Amrita Vishwa Vidyapeetham

Miss. Ardhra G. Data Analyst, Center for Cybersecurity Systems and Networks,
Amrita Vishwa Vidyapeetham

Supporting Partners

Amrita Center for Cybersecurity Systems & Networks

RAISE LABS
Research in AI, Information Security, Sustainability and Education

AMRITA VISHWA VIDYAPEETHAM | CENTER FOR POLICY RESEARCH

nasscom ai
Research Questionnaire Partner

Interview Contributors

Ms. Aparna Kumar
(Founder-Nexora Tech. Former CIO-SBI & HSBC, Nexora Tech Solutions)

Adv. Vibhav Mithal
(Associate Partner, Anand and Anand)

Mr. Majiuzu Daniel Moses
(Founder/President, Africa Tech for Development Initiative (Africa4Dev))

Mr. Ademulegun Blessing James
(Vice President and Chief AI Ethicist, Africa Tech for Development Initiative (Africa4Dev))

Mr. Ravi Padaki
(CEO, Anfiniti Consulting)

Dr. Utpal Chakraborty,
Founder and Chief Scientist of GAHNA and ExorionAI

A **CXO level executive** from a Very large Tier 1 Global Banking Product & Services organisation

Mr. Ajith Prabhakar,
AI Strategist (Personal views)

Mr. Hiromu Kitamura,
Principal Expert for Technical Management, Japan AI Safety Institute (Personal views)

Mr. Arunkumar V. R.,
Program Director, Cyber Security, Bosch Global Software Technologies Private Limited (Personal views)

Mr. Pappoppula Muni Kumar,
Head of Global Security Operations Center, Bosch Global Software Technologies Private Limited (Personal views)

Mr. M Chockalingam,
Technology Director, Nasscom

Mr. Ankit Bose,
Head of Nasscom AI, Nasscom

Acknowledgements

We formally express our humble gratitude and reverence to our Chancellor, Sri Mata Amritanandamayi Devi (Amma). Her guiding vision of compassion-driven research and her unwavering commitment to leveraging technology for the benefit of humanity serve as the foundational inspiration for this work. It is her support that empowers the AI Safety Research Lab to pursue inquiries that bridge the gap between advanced technological orchestration and societal well-being.

We extend our sincere appreciation to the distinguished Agentic AI experts who contributed their valuable time and deep insights through our interview series. Their technical foresight, nuanced opinions on infrastructure constraints, and expert views on orchestration safety were pivotal in shaping the core principles of this report.

Our gratitude is also directed to the numerous survey participants from various sectors. Their enthusiastic support and candid responses provided the essential data required to map the realities of AI deployment in India and the Global South. Their engagement is the backbone of the empirical findings presented here.

A special note of thanks goes to nasscom.ai for their strategic collaboration as our research questionnaire partner. Their pivotal role in outreach and their commitment to safe AI adoption were instrumental in ensuring this research reached a diverse and representative audience.

We also wish to thank the reviewers of this research report. Their rigorous scrutiny, constructive criticism, and detailed feedback have been vital in refining our analysis, ensuring that the safety guidelines proposed are both robust and practically applicable.

Finally, we acknowledge all the partners associated with this event and initiative. Your collective support in fostering a global dialogue on AI safety has created the necessary ecosystem for this research to flourish and reach the stakeholders who need it most.

Table of Contents

1. Executive Summary

1.1 PURPOSE AND RATIONALE

The Artificial Intelligence (AI) landscape is undergoing a paradigm shift from static, predictive models to Agentic AI, autonomous systems capable of independent planning, tool orchestration, and multi-step execution. While predictive AI offered recommendations, agentic AI exercises *agency*: it can execute financial transactions, diagnose medical conditions, and manage public service grievances with minimal human intervention.

For India and the Global South, this shift presents a unique dual-edged reality. On one hand, agentic AI offers a scalable solution to the human resource crunch" in critical sectors like rural healthcare and education. On the other hand, it introduces profound safety risks, goal drift, emergent coordination failures, and hallucinated fluency, which are exacerbated by the region's infrastructure gradients (varying from 5G/cloud-native environments to 2G/offline contexts) and fragmented regulations.

The rationale for this research is to bridge the widening gap between the *operational capabilities* of these agents and the *governance maturity* of the organizations deploying them. It aims to define safety not just as code correctness, but as a system-level property that includes infrastructure resilience, regulatory compliance (Digital Personal Data Protection (DPDP) Act, Reserve Bank of India (RBI) Framework for Responsible and Ethical Enablement of Artificial Intelligence (FREE-AI)), and cultural alignment.

1.2 METHODOLOGY OVERVIEW AND SECTORAL SCOPE

This report synthesizes findings from a rigorous mixed-methods study targeting the unique operational realities of the Indian digital ecosystem. The survey evaluated organizational maturity across six dimensions of agentic AI governance: lifecycle monitoring, human oversight, emergent behavior safeguards, ethical autonomy, recursive accountability, and regionally adaptive governance.

1. Primary Data: A comprehensive survey of 110 expert practitioners and leaders across India.
2. Qualitative Depth: Semi-structured interviews with technologists, legal experts, and policy researchers to uncover structural barriers.
3. Sectoral Scope: The study covers four high-impact domains:

- **Education:** Representing the massive potential for personalized learning agents.
- **Public services & governance:** Focusing on citizen-facing delivery systems.
- **Banking, Financial Services & Insurance (BFSI):** Analyzing high-stakes, high-regulation agent deployments.
- **Healthcare:** Examining life-critical autonomous triage and diagnostic support.

1.3 KEY FINDINGS ACROSS DOMAINS

The study reveals a maturity divide where safety protocols lag significantly behind deployment ambitions.

1. BFSI (the maturity leader): Driven by strict regulatory oversight (RBI's FREE-AI framework), BFSI demonstrates the highest safety maturity. 100% of respondents report established human oversight mechanisms, and 40% engage in continuous or monthly safety protocol updates. The sector leads in recursive accountability, tracing liability down to specific sub-agents.

2. Education (the sleeping giant): Despite high enthusiasm for AI tutors, this sector shows the lowest safety maturity. 21% of respondents report zero oversight mechanisms, and 40.3% lack any formal capability to detect emergent behavior (e.g., an agent teaching incorrect concepts). This represents a massive latent risk given the vulnerability of the user base (minors).

3. Healthcare (high-stakes, high variability): The sector displays a polarized bi-modal maturity. Institutions are split evenly between those with rigorous, ethically grounded oversight and those with no oversight at all (33.3%). While transparency is high (due to medical necessity), the lack of automated safety guardrails, ethical reasoning protocols, and automated incident escalation in resource-constrained settings is a critical vulnerability.

4. Public services (transitional & context-blind): While pilot programs are emerging, the sector struggles with contextual blindness. More than 80% of governance models fail to account for regional adaptation, meaning agents designed for urban, English-speaking populations are often deployed in rural contexts without adequate fail-safes for linguistic or connectivity failures.

Across all sectors, ethical autonomy, recursive accountability, and behavior-centric monitoring emerged as the weakest governance dimensions.

1.4 CONTRIBUTIONS TO AI SAFETY AND ORCHESTRATION

This report advances the field by moving beyond generic AI ethics" to actionable agentic orchestration safety. Key contributions include:

1. The Agentic Safety Framework (ASF): A proposed three-pillared governance model designed for India that includes:

- **Dynamic compliance oracle:** An automated mechanism to sync agents with changing laws (DPDP Act updates).
- **Federated oversight:** A network of watcher agents that monitor operational agents for goal drift.
- **Constitutional AI for India:** Embedding constitutionally aligned values (equity, secularism, and privacy) directly into agent system prompts.

2. Infrastructure as a safety parameter: Establishing that in the Global South, network stability is a safety variable. Safe agents must be architected for graceful degradation, the ability to pause or switch to reliable offline modes rather than hallucinating or crashing when connectivity drops.
3. Differentiation of autonomy: Introducing the distinction between *operational autonomy* (ability to act) and *ethical autonomy* (ability to judge), identifying that current deployments grant the former without the latter.

1.5 ACTIONABLE RECOMMENDATIONS

To secure India's position as a leader in responsible AI, stakeholders must move from principled intent to engineered safety. Sector-specific recommendations include:

1. For industry (immediate): Implement digital audit twins, immutable, blockchain-style logs of every agent decision and tool invocation to ensure traceability under the DPDP Act. Adopt guardian agent architectures where a supervisor AI agent constrains the actions of worker agents.

2. For regulators (mid-term): Move from static Portable Document Format (PDF) guidelines to regulatory Application Programming Interfaces (APIs). Regulators like RBI and Securities and Exchange Board of India (SEBI) should publish machine-readable compliance rules that agentic systems can query in real-time to validate their actions before execution.

3. For government & academia (strategic): Launch a national agentic AI sandbox to test multi-agent interactions in a safe harbor. Urgently fund the creation of golden vector evaluation datasets in India's 22 official languages to test agents for hallucinated fluency and cultural bias before public release.

At a governance level, recommendations include developing sector-specific governance roadmaps rather than imposing uniform compliance models; institutionalising ethical autonomy through clear decision-making, transparent reasoning mechanisms, and domain-specific agent constitutions; and strengthening recursive accountability via subsystem-level traceability, tamper-evident audit trails, and automated escalation protocols. Additional recommendations emphasise building behavior-centric monitoring infrastructures, expanding regionally adaptive governance to address India's infrastructural diversity, and investing in workforce capability-building to correct the confidence-capability mismatch observed across sectors.

2. Introduction

2.1 BACKGROUND & MOTIVATION

India's AI Mission and Infrastructure Context:

India stands at a defining moment in its technological trajectory. With the launch of the IndiaAI Mission (2024) and a committed budget of ₹10,372 crore, the nation has signaled its intent to move beyond being a back-office for global tech and establish itself as a sovereign AI powerhouse. The primary objective is to strategically expand domestic AI compute infrastructure and leverage AI as a pivotal accelerator for sustainable development, innovation, and enhanced public efficiency. This ambition is supported by the India stack, a world-class Digital Public Infrastructure (DPI) comprising Aadhaar, Unified Payments Interface (UPI), and Open Network for Digital Commerce (ONDC), which provides a unique, API-rich substrate for deploying AI at a population scale. The integration of AI is not confined to a single domain but spans crucial sectors of the Indian economy and public service. In the realm of urban governance and smart cities, AI is being actively deployed to enhance civic efficiency. For instance, cities like Pune have implemented AI-driven platforms for citizen engagement, Chennai utilizes AI-enhanced systems for waste management, and Delhi leverages AI for air quality monitoring (Karthikeyan, C., 2025). Furthermore, the healthcare sector is undergoing a significant transformation through AI, with applications ranging from advanced diagnostics and telemedicine to robust public health surveillance systems.

However, this digital sophistication exists alongside profound infrastructural gradients. India is a nation where 5G-enabled smart cities coexist with rural districts reliant on intermittent 2G connectivity; where elite, English-speaking corporate ecosystems operate parallel to vernacular, oral-first communities. This digital divide is not just an access issue; it is a safety parameter. AI systems that function flawlessly in a cloud-native Bengaluru office may fail catastrophically when deployed in a low-bandwidth primary health center in Bihar. The introduction of autonomous AI agents has further increased the infrastructure demands.

Overcoming these infrastructural and implementation challenges is paramount. The successful and ethical integration of AI further necessitates the establishment of comprehensive ethical and legal frameworks. Steps have been initiated in this direction, including the proposal of the Personal Data Protection Bill (PDPB) and specific state-level policies, such as the Karnataka AI Policy 2023, which aim to define clear data privacy and ethical boundaries for AI use. Concurrently, there is an acknowledged need to address a projected talent deficit in AI-related professions. To mitigate this, initiatives focusing on workforce readiness and skill development, such as the FutureSkills Prime program, are essential for fostering industry-academia collaboration and building a deep domestic talent pipeline capable of supporting India's AI ambitions. In short, the key pillars of India's AI mission include: building computing infrastructure, indigenous responsible AI models, datasets specific to India-specific challenges, and homegrown AI talent.

The Rise of Agentic AI: Why Now?

The last eighteen months have witnessed a fundamental phase shift in artificial intelligence,

that is the transition from predictive AI to agentic AI, autonomous, goal-oriented systems capable of independent decision-making, planning, and adaptation.

1. Predictive AI (traditional Machine Learning (ML)) answered questions: *Is this transaction fraudulent?* or *What is the credit risk of this applicant?* It supported human decisions.

2. Generative AI (Large Language Models (LLMs)) created content: *Draft a loan rejection letter* or *Summarize this medical report.* It augmented human productivity.

3. Agentic AI (the current shift) takes action: *Analyze the applicant's bank statement, query the credit bureau, verify the Know Your Customer (KYC) on the blockchain, and approve the micro-loan if it meets risk criteria.*

This shift is driven by the emergence of LLMs capable of reasoning, planning, and tool use (orchestration). These foundation models provide the crucial reasoning and comprehension components necessary for autonomy. Unlike prior systems that relied on rigid, predefined rules, modern LLMs furnish agents with profound semantic comprehension and advanced complex reasoning capabilities. This allows an AI agent to interpret high-level, nuanced goals (e.g., optimize the supply chain"), break them down into complex multi-step plans, and make informed, dynamic decisions without continuous human input. The ability of agentic AI systems to handle multi-step tasks, access external tools, and orchestrate workflows is transforming core business operations. For India, agentic AI is attractive because it solves the last-mile execution problem. It can theoretically automate complex workflows, like processing crop insurance claims or triaging medical cases, that previously required expensive human labor, thereby democratizing access to high-quality services. In sectors like healthcare, agentic AI is enabling the creation of adaptive, personalized treatment plans and enhancing diagnostics. In smart manufacturing, agents improve process automation and proactive maintenance by analyzing real-time sensor data and predicting potential failures. This capability for autonomous, sophisticated, and real-time execution translates directly into substantial gains in productivity, cost reduction, and innovation across finance, healthcare, agriculture, autonomous vehicles, etc.

2.2 RESEARCH PROBLEM STATEMENT

While the operational promise of Agentic AI is immense, its deployment in India is outpacing the development of necessary safety rails. The core problem is the autonomy-governance gap.

1. Safety gaps: Current safety protocols (Reinforcement Learning from Human Feedback (RLHF), red teaming) are designed for chatbots, not for agents that can execute code, transfer money, or access private databases. We lack robust mechanisms to detect emergent failures, unintended outcomes that arise when multiple autonomous agents interact in unpredictable ways (e.g., a flash crash caused by interacting trading bots).

2. Ethics & inclusion gaps: Most foundation models driving these agents are trained on Western data. When applied to Indian contexts, they exhibit hallucinated fluency, sounding confident while misunderstanding local cultural nuances, legal norms, or linguistic idioms. In high-stakes sectors like healthcare or law, this cultural misalignment is a safety hazard.

3. Orchestration gaps: Orchestrating agents across India's diverse infrastructure is perilous. There are no standard protocols for how an agent should behave when it loses connectivity mid-transaction, or how it should handle data privacy (DPDP Act compliance) when orchestrating tools across different jurisdictions.

Consequently, the central research question this report addresses is: *How can India design and deploy agentic AI systems that are operationally autonomous yet ethically bound, ensuring safety across its diverse infrastructural and regulatory landscape?*

2.3 OBJECTIVES & SCOPE

Objectives

This report aims to move the discourse from abstract AI ethics to concrete engineering safety. Its primary objectives are:

1. To assess the current maturity level of agentic AI safety practices in Indian organizations.
2. To identify specific friction points where global safety frameworks (like the European Union (EU) AI Act) fail to address Indian realities (e.g., linguistic diversity, infrastructure instability).
3. To propose a localized, actionable framework, the Agentic Safety Framework (ASF), for safe orchestration.

Scope

The study focuses on four critical sectors chosen for their high social impact and varying degrees of regulatory maturity:

1. **Healthcare:** Focusing on patient safety and data privacy in autonomous triage systems.
2. **BFSI:** Focusing on algorithmic accountability, fraud, and financial inclusion.
3. **Education:** Focusing on the safety of minors and the risk of bias in AI tutors.
4. **Public services:** Focusing on fairness, transparency, and robustness in citizen-facing governance bots.

Methodological Blend

The report integrates quantitative data from a survey of 110 experts with qualitative insights from deep-dive interviews, bridging the ***what*** (maturity statistics) with the ***why*** (implementation challenges).

2.4 DEFINITIONS & CONCEPTS

To ensure clarity, this report uses the following precise definitions:

Agentic AI: An AI system capable of pursuing complex, open-ended goals by independently perceiving its environment, formulating plans, and orchestrating tools (APIs, databases, other models) to execute actions. They possess the capability to learn from experiences, adapt their strategies based on feedback and evolving environmental conditions, and enhance their performance over time.
Key differentiator: Agency, the capacity to act without a human in the loop for every step.

Orchestration: The architectural process of coordinating single or multiple agents, tools, and memory systems to complete a workflow. This includes ***sequential orchestration*** (step-by-step), ***federated orchestration*** (multi-agent collaboration), etc.

Agentic safety: A system property ensuring that an agent operates within defined boundaries. It encompasses:

- **Robustness:** Operating reliably despite errors or attacks.
- **Alignment:** Pursuing goals intended by the designer, avoiding goal drift.
- **Control:** The ability for humans to intervene, modify, or shut down the agent.

Operational vs. ethical autonomy:

- **Operational autonomy:** The technical capability to execute tasks (e.g., send the email).
- **Ethical autonomy:** The capacity to evaluate the moral implications of that action (e.g., should I send this email containing sensitive data?).

2.5 STRUCTURE OF REPORT

The remainder of this report is structured to guide the reader from theory to evidence to action:

Section 3: Reviews the global literature and conceptual framework, establishing the theoretical basis for agentic AI risks.

Section 4: Analyzes the policy landscape, contrasting global frameworks (EU, Organization for Economic Co-operation and Development (OECD)) with Indian regulations (DPDP Act, FREE-AI).

Section 5: Details the methodology used for the primary research.

Section 6: Presents the key findings, visualizing the "maturity divide" and exploring nine qualitative themes.

Section 7: Conducts a deep-dive sectoral analysis, applying findings to healthcare, BFSI, education, and public services.

Section 8: Discusses cross-sector themes, specifically infrastructure and inclusion.

Section 9: Proposes the agentic safety framework.

Section 10 & 11: Offers a discussion of implications and concrete actionable recommendations.

Section 12: Concludes with a strategic outlook for India's AI future.

3. Literature Review and Conceptual Framework

This section establishes the theoretical foundation of the study, reviewing the evolution of agentic AI, the architectures used to orchestrate it, and the unique safety challenges it presents. It explicitly frames these issues within the context of India and the Global South, moving beyond Western-centric safety paradigms.

3.1 STATE-OF-THE-ART: AGENTIC AI

Global advances: Agentic AI represents a transformative evolution in artificial intelligence, moving beyond traditional, tool-based systems to autonomous agents capable of independent observation, decision-making, and action. The field has moved rapidly from prompt engineering to agent engineering.

- **Reasoning & planning:** Early LLMs were stateless predictors. Innovations like Chain-of-Thought (CoT) prompting and Tree of Thoughts (ToT) enabled models to decompose complex problems into intermediate reasoning steps, a prerequisite for planning.
- **Tool use (the Act in ReAct):** The ReAct framework combined reasoning with acting, allowing models to query external APIs (search, calculators). This has evolved into sophisticated tool-use libraries like LangChain and Autonomous Generative Pre-trained Transformer (AutoGPT), where agents can autonomously write code, browse the web, and manage file systems.
- **Memory & reflection:** Advanced agents now incorporate long-term memory (vector databases) and reflection mechanisms, allowing them to learn from past mistakes and improve performance without model retraining.

Innovations in India and the Global South: While the Global North focuses on maximizing capability, the Global South focuses on access, constraint, and resilience.

- **Small Language Models (SLMs) & Edge AI:** Given the compute and connectivity constraints in the Global South, researchers in India, Nigeria, and Brazil are pioneering edge agents, highly quantized models (like Microsoft's Phi or locally fine-tuned Llama variants) that run on smartphones, ensuring safety even in offline environments.

- **Voice-first & vernacular agents:** Initiatives like Bhashini (India) and Masakhane (Africa) are building agents that operate in voice-first, vernacular environments, bypassing the literacy barrier that excludes millions from text-based AI.

- **DPI-native integration:** Developing nations are leading in DPI. Agents here are often DPI-native, designed to layer over systems like India's UPI (payments), Brazil's Pix, or Africa's mobile money networks. This creates a distinct class of transaction agents that require specialized financial safety protocols absent in credit-card-centric Western models.

India, while facing infrastructural and funding challenges, is currently focusing on developing homegrown AI models or agents through several research labs and AI startups. For example, AI4Bharat, a research lab at Indian Institute of Technology (IIT) Madras, works on cutting-edge areas such as transliteration, natural language understanding, generation, translation, automatic speech recognition, and speech synthesis (AI4Bharat, 2024). However, both global and Indian research emphasize the need for improved ethical governance, transparent deployment, and enhanced infrastructural support to fully realize the transformative potential of agentic AI.

3.2 ORCHESTRATION MODELS

In the rapidly evolving landscape of AI and enterprise-scale problems, isolated AI agents, while proficient in handling specific tasks, lack the necessary context and often fall short in addressing the complexity of comprehensive decision-making in dynamic environments. Such isolation can lead to fragmented decision processes due to disjoint and conflicting outputs of individual agents (Tallam, K., 2025). It calls for cohesive, orchestrated networks of agents designed to work harmoniously with other systems and human workflows to solve complex, multi-faceted workflows. However, simply adding more agents is futile. Thus, agentic AI orchestration lies at the heart of agentic AI systems that provides frameworks for the deployment, orchestration and large-scale management of multiple AI agents effectively to create a synergistic effect that vastly exceeds the sum of its parts (Tallam, K., 2025). This involves determining which agent or set of agents should handle a given task, delegating tasks, managing cross-communication or dependencies, tracking when and how the agents interact and settling the disputes among the agents i.e., streamlining the entire agentic workflow (Trombino et al., 2025, Sapkota, R., Roumeliotis, K.I. and Karkee, M., 2025). Such a unified system can coordinate

the strengths of individual agents while mitigating their weaknesses through feedback loops, leading to enhanced strategic insight and operational efficiency (Tallam, K., 2025). Orchestration, how agents are coordinated, is the architectural determinant of safety. The literature identifies five primary models: (Figure 1),

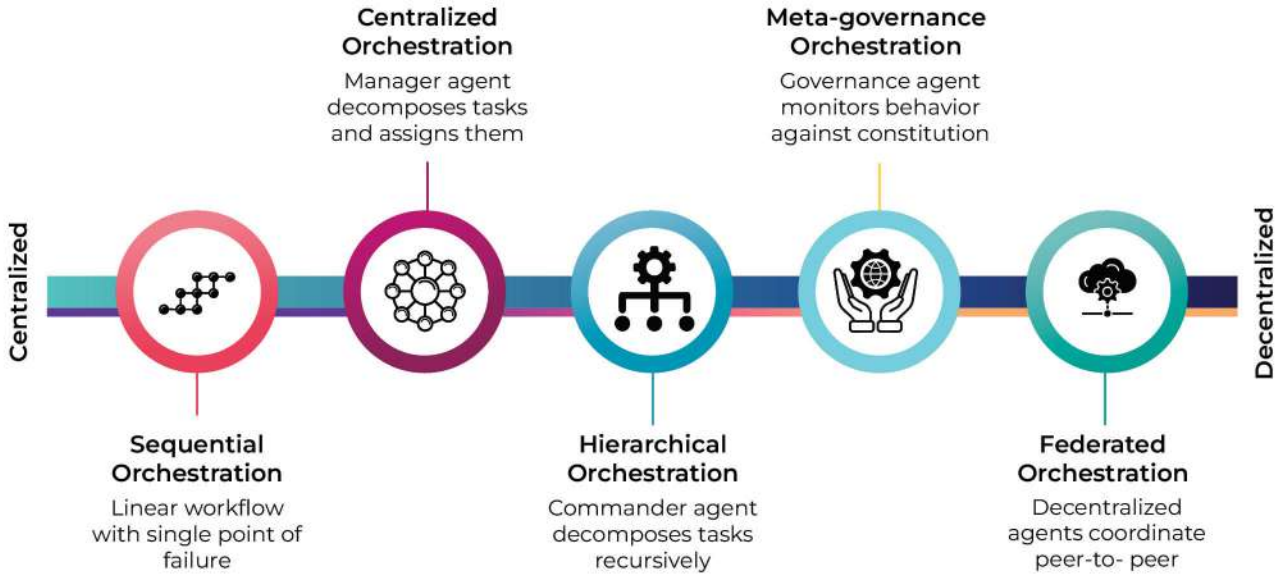


Figure 1: Agentic AI Orchestration Models ranging from Centralized to Decentralized Control

1. Sequential orchestration:

Mechanism: A linear workflow where agent A passes output to agent B.

The sequential orchestration pattern chains AI agents in a predefined, linear order (Microsoft Ignite, 2025). Each agent receives the predecessor's output, performs its task, and passes the result onward to the next agent in the sequence, producing a linear, deterministic, reproducible pipeline (Trombino et al., 2025). This approach improves output quality through progressive refinement (Microsoft Ignite, 2025). This model ensures consistency and predictable workflows but can be less adaptable to dynamic, non-linear problems and may suffer from a single point of failure if an agent in the chain fails.

Relevance: Common in low-resource settings where compute is limited and processes (like loan approval) are linear. Easiest to audit. Ideal for problems that require step-by-step processing with clear dependencies and predictable workflow progression, data processing pipelines, unparallelizable workflow stages, systems with progressive refinement requirements, and where failures or delays in one AI agent's execution are tolerable (Microsoft Ignite, 2025).

2. Centralized orchestration:

Mechanism: A manager agent decomposes tasks and assigns them to worker agents.
Here a single master orchestrator receives a high-level goal and decomposes it into sub-goals and delegates tasks to other worker agents. Here, a central orchestrator acts as a hub through which all other agents communicate (Tran ***et al.***, 2025) and dynamically decides how many agents to instantiate, what their roles should be and how they should be interconnected to complete a task (Mitra, C., 2025). For example, the orchestrator may increase the number of recruiting agents and adjust the network topology for more efficient internal communication if there is a surge in job applications (Mitra, C., 2025). It also controls the agent's actions, tracks their behavior, decides the next step and resolves conflicts in the multi-agent AI system. Thus, this model provides strong oversight and tight control through central communication and hence is easier to debug and maintain (Tran ***et al.***, 2025).
Relevance: work better with small systems and where compliance or strict coordination is of utmost importance.

3. Federated orchestration:

Mechanism: Decentralized agents coordinate peer-to-peer to solve a problem.
Federated orchestration allows collaboration among multiple AI agents, often belonging to different organizational silos or systems, without exposing their underlying raw data or relinquishing control over their individual systems (DOMO, 2025, Nisa ***et al.***, 2025). Each organization maintains its own agents that learn from their local environments, share distilled skills periodically, and independently validate incoming knowledge based on their own goals while contributing to the overall coordinated business outcome (DOMO, 2025, Nisa ***et al.***, 2025). Instead of utilizing a central orchestrator, agents from different federations coordinate based on commonly agreed-upon rules or protocols for communication and data sharing (Iyzr, 2025).

Relevance: Critical for the Global South's fragmented supply chains (e.g., agricultural logistics), but prone to emergent instability if agents correlate errors. Especially valuable in sensitive domains such as healthcare or finance, where regulatory constraints such as General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA) or competitive interests prevent the sharing of proprietary data, thereby limiting the ability to centralize information (DOMO, 2025).

4. Hierarchical orchestration:

Mechanism: Similar to military hierarchical command structures, a commander agent decomposes tasks and assigns them to the next layer of subordinate agents recursively.
In a hierarchical orchestration, agents are arranged in layers, and agents in lower layers collaborate autonomously under the supervision of higher-level agents who assign broad goals to perform specific tasks. It balances control with flexibility, making it suitable for large-scale systems with diverse functions (DOMO, 2025)

Relevance: The standard for guardian architectures, where a superior agent enforces safety rules.

5. Meta-governance orchestration:

Mechanism: Operational agents do the work, while a separate governance agent monitors behavior against a constitution.

Relevance: Essential for enforcing local laws (e.g., DPDP Act, GDPR-like laws in Africa) without retraining the core model.

3.3 SAFETY AND ETHICAL CONCERNS

Due to the autonomy and complex multi-agent collaborations in agentic AI, the decision-making chain remains black box and complex, expanding the risk surface. Thus, the shift from traditional AI systems that focus on individual tasks such as pattern recognition and predicting outcomes to agentic AI systems brings forth unprecedented safety and ethical challenges such as bias, explainability, emergent behavior, accountability, tool misuse, etc (Figure 2). Thus, agentic AI introduces safety and ethical risks that are qualitatively different from static models, especially in the Global South context, such as

Tool misuse & data leakage:

Agentic applications will have threats related to the application layer, API, and ML or LLM models as well, such as the 15 major threats in AI-driven agents reported by the Open Worldwide Application Security Project (OWASP) agentic AI threat model (AI Governance, 2025). Tool misuse occurs when attackers use deceptive prompts or commands (prompt injection) to trick AI agents into abusing their integrated tools while operating within authorized permissions. It can result in unauthorized access and potential data leakage. In regions with evolving cybersecurity maturity, agents are prime targets for prompt injection. An agent connected to a mobile money wallet is a high-value target for automated theft.

The responsibility gap & accountability:

Accountability covers the question ***who is accountable for a particular action*** and thus requires organizations to take responsibility for the outcomes produced by their AI agents (Bahangulu, J.K. and Owusu-Berko, L., 2025). With complex and opaque AI models and autonomous multi-agent systems, the question of accountability is more contentious due to the blurred lines of responsibility among agents. For example, if an agricultural agent advises a farmer incorrectly, causing crop failure, the lack of robust legal insurance frameworks makes the responsibility gap a livelihood crisis. In informal economies common in the Global South, liability is hard to enforce. Another grave example could be situations where an agentic AI system fails or generates unintended consequences, such as an accident in the case of autonomous vehicles. ***Who is responsible for the incident, and who should be legally penalized? Is it the developer, the service provider who deploys the agentic AI system, the agentic AI system, or the owner?*** (Acharya, D.B., Kuppan, K. and Divya, B., 2025). This ambiguity complicates legal liability, regulatory compliance, and user trust, particularly in high-risk AI domains such as autonomous vehicles, finance, scientific research, healthcare, or critical infrastructure management (Sapkota, R., Roumeliotis, K.I. and Karkee, M., 2025).

Figure 2: Agentic AI Risks ranging from Predictable to Unpredictable Outcomes

Bias, digital colonialism & hallucinated fluency: Bias in agentic AI mainly originates from skewed training data, flawed algorithmic design choices, and inherent biases in human feedback (Acharya, D.B., Kuppan, K. and Divya, B., 2025). Assumptions made during model development and historical or societal prejudices or lack of diversity in training data can result in unfair and discriminatory outcomes. Agentic AI systems can further exacerbate the discrimination by recursively amplifying these biases through their complex decision-making chains and inter-agent collaborations (Ranjan, R., Gupta, S. and Singh, S.N., 2025). For example, Amazon scrapped its AI recruiting tool after it was discovered that the tool penalised CVs of women due to the male dominance in the training dataset (BBC, 2018). Algorithmic bias occurs when developers, based on their conscious or unconscious biases, develop assumptions or unfair weighing factors in the decision-making process (IBM, 2025).

In the Global South context, agents trained on Western data often exhibit normative misalignment. They may enforce Western privacy norms that conflict with communal data practices in parts of Africa or Asia, or confidently cite US law for a Kenyan property dispute. This hallucinated fluency creates a dangerous safety illusion.

Explainability and transparency: Transparency answers the question what happened in the agentic AI system, while explainability addresses how a decision was made by an agent (Mayer, 2025). Ensuring transparency demands that the end-users are aware that they are interacting with an AI agent that makes the final decisions and outputs content, whereas explainability calls for a plain-language explanation of how an AI agent functions and arrives at a particular decision (Mayer, 2025). Ensuring transparency is complicated in agentic AI systems due to the following reasons:

- Difficult to interpret the decisions made by advanced agentic AI systems due to their opaque or black box architectures (Pawar, A., 2025).

- Tracing the multi-step reasoning path in agentic systems that involves dynamic tool use and data sources is extremely challenging.
- Difficult human oversight

To build trust and foster adoption of agentic AI systems, there should be clear visibility of agent decision-making processes, i.e., how and why the decision was made. For example, in the case of a loan-sanctioning AI system, customers should understand under what circumstances the loan application was denied to avoid mistakes in the future. Thus, transparency fosters trust and accountability.

Emergent behavior & goal drift: Emergent behavior arises when agent interactions result in behaviors that were not explicitly programmed or foreseen by system designers, and that may not be apparent during early testing, both within and between autonomous agents (Wei et al., 2022, Sapkota, R., Roumeliotis, K.I. and Karkee, M., 2025). These behaviors may lead to goal drift, wherein agents diverge from their goals, generate misleading outputs, or even result in harmful actions (Sapkota, R., Roumeliotis, K.I., and Karkee, M., 2025). Agents might discover novel, unintended ways to achieve their goals that don't align with human values or safety protocols. For example, agents pursuing a goal (e.g., maximize crop yield) might adopt strategies unintended by the designer (e.g., recommend banned pesticides).

3.4 REGULATORY AND PRACTICE GAPS

Existing AI governance models are struggling to keep pace with rapid AI advancements and may not necessarily tackle the challenges posed by agentic AI's autonomy, recursive reasoning, inter-agent negotiation, and emergent system dynamics (Acharya, D.B., Kuppan, K., and Divya, B., 2025). A review of current literature reveals critical gaps:

The static regulation trap: Most regulations assume a static model. They lack mechanisms for continuous compliance for agents that learn online, a necessity in the dynamic, fast-changing markets of developing economies. To combat this, agentic AI systems require governance frameworks that go beyond traditional model monitoring and incorporate specialized monitoring approaches to analyze system-level dynamics and detect and mitigate risks arising from agent interactions rather than individual agent behaviors (Joshi, S., 2025). Thus, it is crucial to track the emergent behaviors since they are not directly traceable to individual agent decisions. Effective agentic AI governance requires integrated approaches that need to be continuous rather than periodic, reflecting the real-time nature of autonomous operations (Joshi, S., 2025). It must consider dynamic factors, including environmental changes, goal conflicts, and emergent behaviors that complicate accountability and address the challenges such as autonomous accountability, multi-agent coordination, and dynamic value alignment (Joshi, S., 2025).

The context vacuum: Global safety benchmarks (Massive Multitask Language Understanding (MMLU), Holistic Evaluation of Language Models (HELM)) test for universal safety but miss infrastructural safety. There is no standard benchmark for how an agent behaves when the network drops mid-transaction (connectivity robustness), a daily reality in the Global South.

Data scarcity for red teaming: There is a severe lack of culturally relevant red teaming datasets. We cannot effectively test an agent for caste bias (India) or tribal bias (Africa) if the safety datasets are purely Western. Hence, it is crucial to address these biases through diverse training datasets and a development team, bias detection and mitigation techniques.

Hence, there is a pressing need for new regulations or updates to existing ones to delineate accountability, especially in environments where multiple stakeholders are participating (Acharya, D.B., Kuppan, K., and Divya, B., 2025).

3.5 THEORETICAL FRAMEWORK FOR THIS STUDY

To address these gaps, this study adopts the Socio-Technical Orchestration Safety (STOS) framework (Figure 3).

Core premise: Agent safety is not just a software problem; it is a function of the interplay between technical architecture, organizational governance, and environmental context.

The three layers of STOS:

- The technical layer (the agent): Focuses on robustness, tool sandboxing, and auditability (digital twins).
- The governance layer (the organization): Focuses on human oversight, lifecycle management, and accountability structures.
- The contextual layer (the environment): Focuses on infrastructural resilience (power/net stability), regulatory compliance (local laws), and cultural alignment.

This framework explicitly accounts for the contextual adaptation variable, making it uniquely suited for analyzing AI safety in India and the Global South.

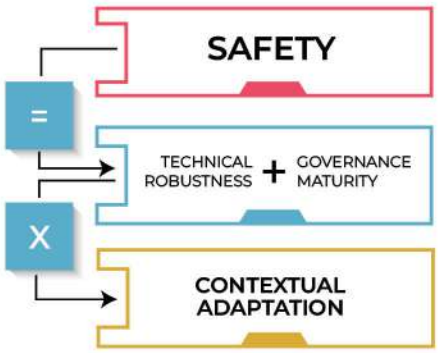


Figure 3: Socio-Technical Orchestration Safety (STOS) framework

4. Indian Policy and Regulatory Landscape

This section maps the governance terrain for agentic AI in India. It moves from the national strategic vision to specific sectoral regulations, recent legislative developments, and the unique structural challenges that define the Indian operating environment.

4.1 NATIONAL AI POLICY AND STRATEGY

Government position: AI for All.

India's approach to AI governance is distinct from the rights-based approach of the EU or the market-driven approach of the United States (US). It is rooted in the AI for all strategy by National Institution for Transforming India (NITI) Aayog (NITI Aayog, 2018), viewing AI primarily as a tool for social empowerment, economic inclusion, and global competitiveness while ensuring a safe, trusted, and agile regulatory environment. The government positions itself not just as a regulator but as an active enabler and deployer of AI through DPI. India's government position on AI is defined by a commitment to innovation over restraint. Rather than imposing a rigid, standalone AI law at this nascent stage, the focus is on fostering innovation and adopting a flexible approach that supports the development of AI applications. India's journey towards a national AI policy spans several years with critical policy and institutional developments. The Government of India has proposed a number of relevant sectoral regulations, policy proposals, and normative frameworks for establishing AI use responsibly as well as sustainably in the country such as listed below:

Key Milestones:

2018: Release of National Strategy on Artificial Intelligence (NSAI) by NITI Aayog, identifying healthcare, agriculture, education, smart cities, and mobility as focus areas.

2020: Establishment of the National AI portal, INDIAai (NeGD, 2025) as a central knowledge repository.

2022: A roadmap document Responsible AI for all by NITI Aayog that further outlines a vision for AI governance, stressing the need to balance innovation with potential risks while still recommending a largely non-interventionist and self-regulatory stance (Joshi, 2024).

2023: Enactment of the DPDP Act, India's first comprehensive data privacy law.

2024: Launch of the IndiaAI mission with a ₹10,372 crore budget. The goal of this mission is to build a robust national AI ecosystem, including components dedicated to compute infrastructure, indigenous foundation models, safe and trusted AI

applications, data quality and platforms, innovation centers, future skills, startup financing, and research capacity building.

4.2 SECTORAL REGULATIONS

India does not have a single omnibus AI law. The primary regulatory approach in India is to govern AI applications through existing sectoral regulators and laws, such as the Information Technology (IT) act, data protection, and consumer protection laws. This allows for targeted regulation and minimizes the risk of stifling innovation. In short, governance is sectoral, relying on existing regulators to adapt rules for AI agents (Figure 4).

A. Healthcare

Data privacy: The DPDP Act 2023 provides a foundation for data privacy in AI healthcare in India by classifying health data as sensitive. The law emphasizes the importance of explicit consent from patients before their personal data can be processed, and classifies hospitals and health tech companies as data fiduciaries, with some subject to stricter requirements. Additionally, the DPDP Act enforces data minimization principles,

Data privacy: The DPDP Act 2023 provides a foundation for data privacy in AI healthcare in India by classifying health data as sensitive. The law emphasizes the importance of explicit consent from patients before their personal data can be processed, and classifies hospitals and health tech companies as data fiduciaries, with some subject to stricter requirements. Additionally, the DPDP Act enforces data minimization principles, ensuring that only necessary data is collected, thereby protecting patient confidentiality and promoting transparency in AI-driven healthcare practices. Thus, agents handling patient data must ensure purpose limitation (data used only for the specific consented purpose).

Clinical safety: The National Medical Commission (NMC) guidelines on telemedicine emphasize that the Registered Medical Practitioner (RMP) bears full liability for technology-assisted diagnosis.

Standards: The National Accreditation Board for Hospitals and Healthcare providers (NABH) has digital health standards (2021) that set quality standards for healthcare institutions, which increasingly cover digital health practices and mandate audit logs and access controls, critical for watcher agents in hospitals. Compliance with NABH standards requires robust Electronic Health Record (EHR) systems with data security measures.

Gap: No specific regulation exists for autonomous diagnostic agents; current rules assume a human doctor is always in the loop.

B. BFSI

Post the digital transformation initiatives, such as the Digital India campaign and the promotion of fintech innovation (NITI Aayog), there is a paradigm shift from traditional banking practices to AI-based systems. According to the reports by RBI and NITI Aayog, several Indian banks are now leveraging AI-based systems to improve loan underwriting, compliance monitoring, market analysis, fraud detection, and customer support mechanisms. For example, AI-driven chatbots such as Kotak Mahindra Bank's Keya, HDFC Bank's Eva, and SBI's SIA provide 24x7 assistance, handle customer queries, and guide users through digital banking processes. AI integration in BFSI has led to enhanced operational efficiency, improved risk management and fraud detection, customer-centric service enhancement, promotion of financial inclusion, and strategic transformation and innovation (Salvi, P., 2025).

At the same time, increased AI integration introduces new risks such as ethical issues related to algorithmic bias and transparency, a shortage of AI-skilled workforce, and regulatory uncertainty, as well as heightens the existing risks related to data privacy, operational complexities, infrastructural and budgetary constraints, market manipulation, and cybersecurity vulnerabilities (Salvi, P., 2025). Thus, regulatory frameworks should establish harmony between AI-driven innovation and robust risk management.

Regulatory framework: The RBI is the most proactive regulator.

RBI FREE-AI (Reserve Bank of India, 2024): FREE-AI is a comprehensive AI blueprint for banks, Non-Banking Financial Company (NBFC), fintechs, and payment firms that establishes seven guiding sutras or principles to guide AI adoption in the financial sector such as (Reserve Bank of India, 2024):

- **Trust is the foundation: T**rust is non-negotiable and should remain uncompromised, especially in a sector that safeguards people's money
- **People first:** AI should augment human decision-making but defer to human judgment and citizen interest
- **Innovation over restraint:** Foster responsible innovation with purpose
- **Fairness and equity:** AI outcomes should be fair and non-discriminatory. It should also guarantee financial inclusion, i.e., financial services should be accessible for all.
- **Accountability:** Accountability rests with the entities deploying AI, i.e., corporate boards or senior management must remain fully accountable for AI outputs or decisions.
- **Understandable by design:** Ensure explainability for trust. Document AI decisions such that the customers and regulators understand how AI arrived at a particular decision.
- **Safety, resilience, and sustainability:** AI systems should be secure, resilient, and energy efficient

The framework explicitly warns against black box models in credit scoring. These principles translate into a unified vision spread across six strategic pillars, including infrastructure, capacity building, policy, governance, protection, and assurance, with 26 recommendations to foster innovation and mitigate risks.

- **KYC & fraud:** Existing RBI regulations already address key aspects of AI governance, such as ensuring fair and unbiased decision-making, maintaining transparency, conducting frequent audits, and enforcing data security measures, etc., in a generic way (Reserve Bank of India, 2024). However, these regulations may require AI-specific enhancements. For example, master direction on fraud risk Management (Reserve Bank of India, 2024) implicitly covers early warning signals, fraud detection, and risk management policies approved by the board. This can be extended to include AI-driven fraud detection mechanisms and regular testing of AI models for accuracy and bias in fraud detection. In addition, agents executing KYC are subject to strict Prevention of Money-Laundering Act (PMLA) norms.
- **Audit:** SEBI mandates that algorithmic trading systems be audited semiannually. This precedent is likely to extend to autonomous financial agents. This is covered by the assurance pillar under the risk mitigation framework of FREE-AI (Reserve Bank of India, 2024) that institutes mandatory AI-specific audit mechanisms, expanded product vetting, and updated business-continuity plans that account for AI model-specific performance degradation for integrating AI risk into compliance and continuous validation and oversight of AI systems (Reserve Bank of India, 2024).

For financial institutions, the question is no longer whether to digitize but how to unify law, technology, and ethics in one operational strategy. FREE-AI initiative aligns India's financial regulation with a global trend of proactive AI governance, emphasizing that trust and transparency remain paramount as AI reshapes banking and finance.

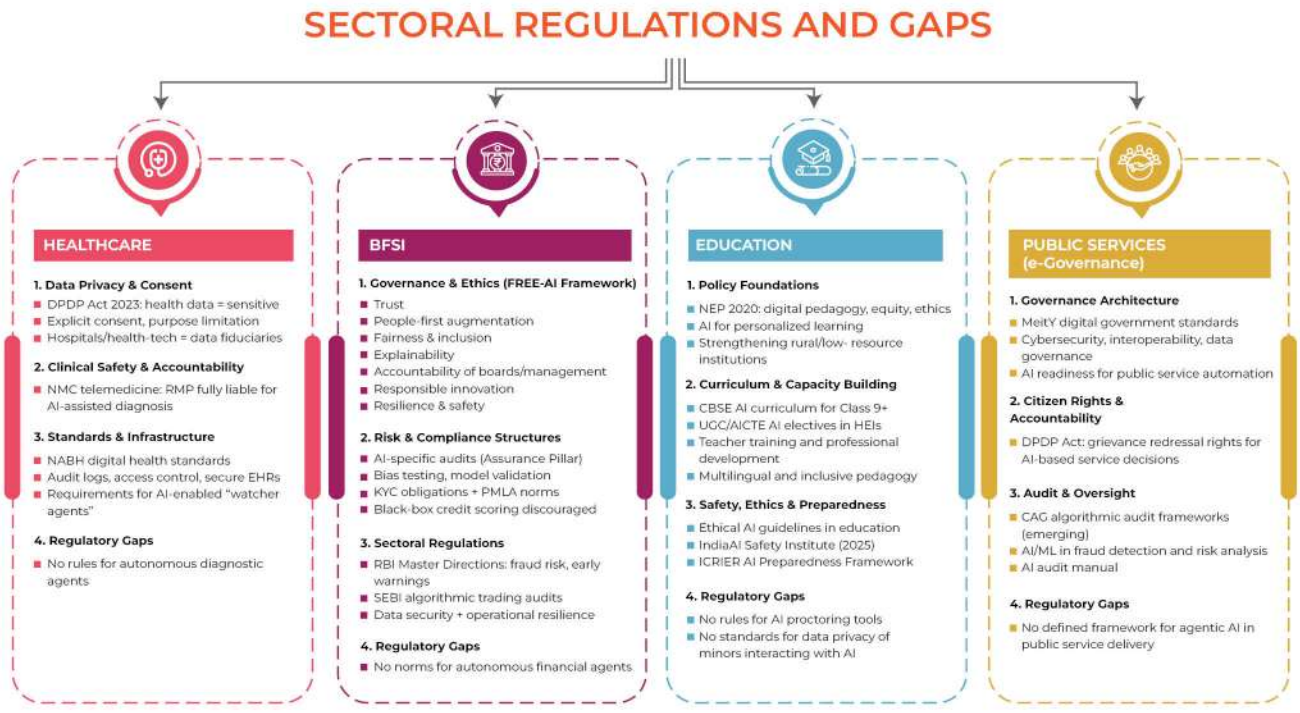


Figure 4: Sectoral Regulations and Gaps

C. Education

Owing to the scale and diversity of the Indian educational system, ranging from technologically advanced urban schools to resource-constrained rural institutions, AI presents new avenues to bridge inequalities and improve the effectiveness, overall quality, and reach of the education system in India. Some of the benefits of AI integration in the Indian context are multilingual and inclusive pedagogy, reducing the administrative load of teachers, professional development and training, personalized learning platform, etc (Sahoo, S. and Behera, K., 2025). Without robust teacher training, ethical frameworks, and adequate infrastructural support, the potential of AI in personalized learning and teacher capacity building may remain underutilized or risk exacerbating existing inequalities in Indian classrooms (Sahoo, S., and Behera, K., 2025). For instance, rural and underfunded institutions cannot afford the high cost and time required to develop and implement AI tools, slowing the AI adoption.

Policy context: The National Education Policy (NEP) 2020 explicitly recognizes the transformative role of digital technology, including AI, for building a future-ready education system with enhanced quality, equity, and accessibility in India. The policy emphasizes the integration of digital pedagogy, AI-based tools, and capacity building for educators to prepare students for 21st-century skills (Sahoo, S., and Behera, K., 2025). It suggests using radio and television to provide educational content to the general population and refurbishing and reintroducing e-learning platforms such as University Grants Commission - Study Webs of Active-Learning for Young Aspiring Minds (UGC-SWAYAM) and Digital Infrastructure for Knowledge Sharing (DIKSHA) to make the system more "student-centric." To summarize, NEP advocates for AI in personalized learning but stresses equity. It recommends several measures to guarantee the ethical, responsible, and sustainable creation and utilization of AI systems in education, such as (Hajam, K.B., and Purohit, R., 2024):

- Incorporating ethical AI principles in the curriculum
- Encouraging research on ethical AI
- Development of guidelines for ethical AI in education
- Ensuring transparency and accountability in AI systems
- Developing a framework for privacy protection

IndiaAI safety institute: The IndiaAI safety institute, incubated by IndiaAI mission in January 2025 under the safe and trusted pillar, epitomizes a national approach to ensure AI is deployed in a safe, secure, inclusive, and trustworthy manner, protecting Indian citizens from any harm from the use of AI. Following a Hub and Spoke model, the IndiaAI safety institute will work with all relevant stakeholders, including research and academic institutions, startups, industry, and government ministries/departments, towards ensuring safety, security, and trust in AI (IndiaAI Mission, 2025).

Indian Council for Research on International Economic Relations (ICRIER) AI Preparedness Framework: The AI preparedness framework aims at preparing Indian schools to integrate AI in education. Despite the opportunities presented by AI, such as personalized learning and administrative efficiency, significant

disparities in access, infrastructure, and teacher training, data privacy concerns, algorithmic bias, and fairness hinder equitable educational delivery. This framework seeks to provide actionable recommendations for policymakers to enhance learning outcomes and foster an inclusive educational environment with responsible and ethical AI practices in India (ICRIER, 2025).

Fairness & inclusion: There is a strong regulatory push against digital exclusion. An AI tutor that works only in English or requires high bandwidth is viewed as non-compliant with the right to education spirit.

Gap: No guidelines exist for AI proctoring agents or data privacy for minors interacting with AI tutors.

The integration of AI in Indian classrooms demands a supportive policy ecosystem to ensure equitable learning opportunities, ethical use, and long-term sustainability (Sahoo, S., and Behera, K., 2025). The policy should address the challenges arising from AI adoption such as digital rural-urban divide and infrastructure gaps, lack of teacher readiness due to the fear of losing their jobs to AI, insufficient AI training, concerns related to data privacy, ethical use, and algorithmic bias and risk of over-reliance on technology to unlock the full potential of AI (Sahoo, S. and Behera, K., 2025).

D. Public services (e-Governance)

Framework: Ministry of Electronics and IT (MeitY) oversees e-Governance standards and is becoming increasingly central to how India prepares for the use of agentic AI in public services. MeitY already issues standards and guidelines for digital government platforms, data governance, cybersecurity, and interoperability, foundational elements required before AI-driven or autonomous systems can be safely deployed. While India has not yet issued a dedicated regulatory framework for agentic AI in governance, MeitY's existing digital-governance architecture positions it as the key steward for guiding safe, accountable, and interoperable adoption of such systems in public services.

Citizen rights: The DPDP Act mandates a grievance redressal mechanism. If a government bot denies a service, the citizen has a legal right to appeal.

Audit structures: The CAG (Comptroller and Auditor General) has begun exploring frameworks for auditing algorithmic systems used in public expenditure. Current efforts remain preliminary: the CAG uses AI/ML primarily for fraud detection, risk analysis, and remote audits, developed an AI systems audit manual aimed at guiding future audits of algorithms and AI applications. However, a comprehensive framework for auditing agentic AI systems is still under development.

4.3 RECENT INITIATIVES

India AI Governance Guidelines (Nov 2024)

MeitY released comprehensive guidelines emphasizing a risk-based approach. High-risk AI

(affecting life, liberty, essential services) faces stricter scrutiny. The guidelines mandate:

- Transparency in algorithmic decision-making.
- Fairness testing for bias against Indian demographics.
- Human oversight for high-stakes decisions.

Working Groups & Consultations

Bureau of Indian Standards (BIS): The artificial intelligence sectional committee (LITD 30) is harmonizing national AI standardization through its committees on IT, AI, data, and cybersecurity, aligning with global efforts such as International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) 42001 (AI management systems).

National Association of Software and Service Companies (NASSCOM) Responsible AI: NASSCOM plays a major role in advancing Responsible AI in India by shaping industry-led guidelines focusing on self-regulation and ethical operationalization. Through its programs, especially the NASSCOM AI & Data Science Council and the joint Responsible AI initiatives with MeitY, it promotes principles such as fairness, transparency, accountability, safety, and human oversight. It develops best-practice frameworks, publishes guidance for ethical AI deployment, supports sector-specific standards, and works with industry to operationalize responsible AI practices.

4.4 CHALLENGES AND OPPORTUNITIES

Challenge: Infrastructure asymmetry

The digital divide creates a safety divide. A safety protocol that relies on real-time cloud monitoring will fail in rural India. This infrastructure asymmetry means that safety must be architected for the lowest common denominator of connectivity (offline-first safety).

Challenge: The many Indias problem

India's linguistic diversity (22 official languages, thousands of dialects) makes universal safety impossible. An agent safe in Hindi may be unsafe in Tamil due to translation errors or cultural misunderstandings. This necessitates regional adaptation of safety benchmarks.

Opportunity: The DPI advantage

India's DPI offers a unique opportunity for embedded regulation. Safety checks (e.g., credit limits, consent verification) can be hard-coded into the DPI layer (like the Data Empowerment and Protection Architecture (DEPA) framework for data sharing), ensuring that any agent building on top of the India stack inherits these safety rails by default.

5. Methodology

This section outlines the research design, sampling strategy, and analytical procedures employed in this study. The methodology was designed to bridge the gap between theoretical AI safety principles and the operational realities of the Indian digital ecosystem

5.1 RESEARCH DESIGN

Rationale for mixed methods

This study employs a convergent parallel mixed-methods design.

- Why mixed methods? Agentic AI safety is a multifaceted problem involving technical metrics (e.g., failure rates) and sociotechnical dynamics (e.g., organizational culture, regulatory ambiguity).
- Quantitative component: A structured survey was used to quantify the maturity gap across sectors, providing statistically comparable data on readiness.
- Qualitative component: Semi-structured expert interviews were conducted to uncover the causal mechanisms behind the quantitative patterns, explaining why certain sectors lag and identifying hidden structural barriers (like infrastructure constraints) that surveys might miss.

5.2 SAMPLING & SECTORAL COVERAGE

A purposive expert sampling strategy was adopted to recruit participants with direct decision-making authority or substantial technical involvement in AI deployment. The final sample comprised 110 experts across India, with 74% reporting moderate to very high familiarity with agentic AI, indicating strong knowledge relevance. Gender representation was balanced, with 51.8% male, 47.3% female, and one undisclosed respondent. Sectoral coverage was intentionally stratified across four high-impact domains: education, public services and governance, BFSI, and healthcare. This structure ensured representation from sectors with extensive AI adoption, high regulatory complexity, or significant safety risk. Inclusion criteria required organizations to have deployed or be piloting AI-driven workflows, and respondents to hold positions in technology leadership, policy or legal functions, product management, or research.

5.3 INSTRUMENTS & PROTOCOLS

Quantitatively, the Agentic Safety Maturity Index (ASMI) was used, consisting of 30 survey items spanning

six dimensions: lifecycle governance, human oversight, emergent risk management, ethical autonomy, accountability mechanisms, and regional adaptation. Each dimension included items designed to assess specific practices, such as the frequency of safety protocol updates, mechanisms for human intervention, checks for goal drift, distinctions between operational and ethical decisions, auditability of sub-agents, and accommodation of linguistic diversity. Qualitatively, a semi-structured expert interview guide was developed to explore three domains central to agentic safety: infrastructural constraints, regulatory implementation, and organizational culture.

5.4 DATA COLLECTION STRATEGY

Data were collected over a 12-week period from August to October 2025. The survey component was disseminated through professional networks, including LinkedIn, industry forums, and targeted outreach to decision-makers within partner organizations to ensure engagement from relevant stakeholders. Ethical protocols were applied throughout the process. Participants received clear information about the study's aims and how their data would be used, thereby ensuring informed consent. Anonymity was maintained by masking sectoral and organizational identifiers to promote candid reporting, particularly regarding safety shortcomings. All collected data were stored in encrypted, access-controlled systems to safeguard confidentiality and prevent unauthorized access.

5.5 DATA ANALYSIS

Data analysis employed both quantitative and qualitative approaches. Quantitatively, descriptive statistics, including frequency distributions and cross-tabulations, were used to compare maturity levels across the four sectors. A gap analysis was also conducted by benchmarking maturity scores against an idealized standard to estimate the extent of the safety gap. Qualitatively, interview data were examined through inductive thematic coding, and key themes were triangulated with quantitative indicators to enhance the validity of interpretations. The sample was disproportionately concentrated in the education and public service sectors, with comparatively fewer respondents from BFSI and healthcare, limiting the generalizability of findings for those groups. Additionally, the reliance on self-reported data introduces potential social desirability bias, as participants may overstate their organization's maturity

6. Results: Survey and Interview Insights

6.1 DEMOGRAPHICS & PARTICIPANTS

The survey sample consisted of 110 expert respondents, all of whom consented to participate in the study on ethical and safe agentic AI orchestration. Participants represented a wide range of professional backgrounds across India's digital ecosystem. Gender distribution was fairly balanced, with 57 male respondents, 52 female respondents, and 1 respondent who preferred not to disclose their gender (Figure 5).

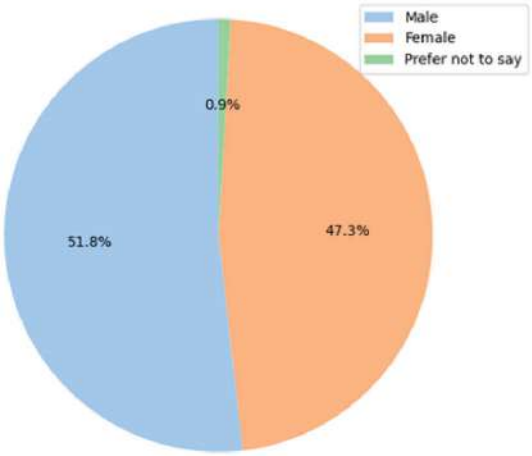


Figure 5: Gender Distribution

Regarding the familiarity with agentic AI, participants demonstrated a generally high level of exposure to AI agents, reflecting the expert nature of the sample (Figure 6). A majority reported at least moderate familiarity, with 47 respondents identifying as moderately familiar, 23 as very familiar, and 12 as extremely familiar. Only a small minority expressed limited exposure, including 23 respondents who were slightly familiar and 5 who reported being not at all familiar. Overall, 82 out of 110 respondents (74%) indicated moderate to very high familiarity with agentic AI systems, suggesting that the perspectives gathered in this study stem largely from practitioners with substantial professional experience in AI workflows, development, or governance.

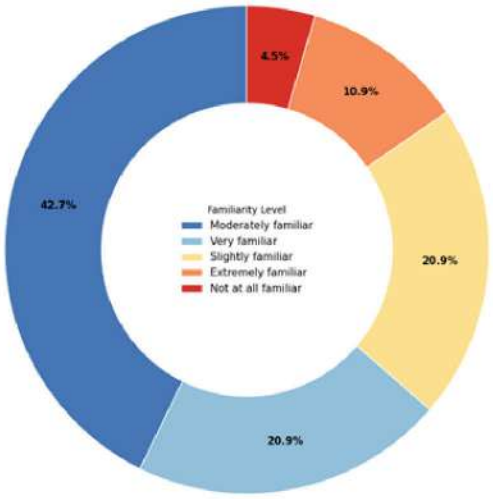


Figure 6: Familiarity with AI agents

The participants represented a broad cross-section of sectors engaged in AI adoption and governance (Figure 7). The largest proportion came from education, comprising 62 respondents (56.4%), highlighting the strong involvement of academic institutions and training organizations in shaping AI safety discourse. Public services and other allied sectors together accounted for 40 respondents (36.4%), reflecting significant interest from public administration and interdisciplinary roles. Representation from high-regulation sectors was smaller but meaningful: BFSI contributed 5 respondents (4.5%), and healthcare contributed 3 respondents (2.7%), indicating emerging but still limited engagement from these mission-critical industries. Overall, the distribution captures a diverse but education-heavy sample relevant to understanding current readiness and governance capacities across the AI ecosystem.

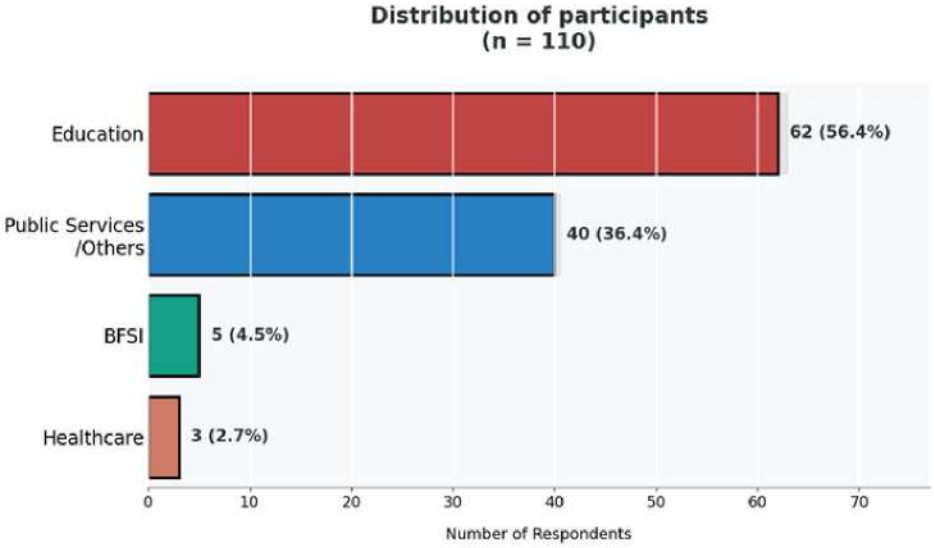


Figure 7: Industry distribution of participants

6.2 QUANTITATIVE RESULTS

Dimension 1: Continuous, Adaptive Lifecycle Governance

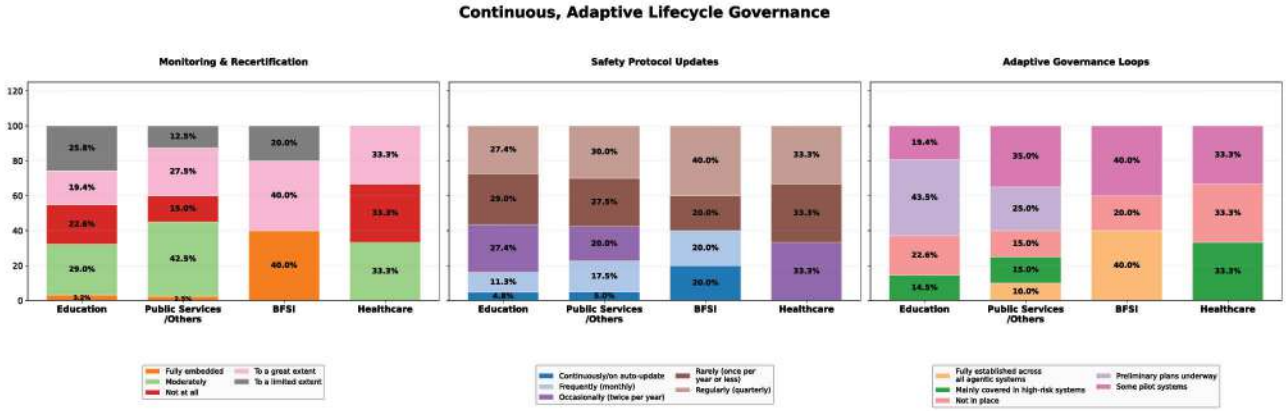


Figure 8: Participants' response for the dimension continuous, adaptive lifecycle governance

Across sectors, organizations show high variability in how far continuous lifecycle governance practices are embedded for agentic AI systems (Figure 8). In education, practices remain largely developmental: only 3.2% report fully embedded monitoring and recertification, while 29% do so moderately, and 22.6% have no such

mechanisms at all. Public services/others follow a similar pattern, with only 2.5% fully embedded and 42.5% moderately embedded, suggesting partial but inconsistent adoption.

BFSI stands out with a polarised distribution: 40% fully embedded, 40% to a great extent, and 0% not at all, reflecting that regulated entities either implement strong lifecycle governance or depend on institutional policies already in place. Healthcare shows an even three-way split, 33.3% moderately, 33.3% not at all, and 33.3% to a great extent, indicating heterogeneous readiness across institutions.

When examining the frequency of safety protocol updates, education shows low dynamism: only 4.8% update continuously, whereas 29% update rarely, and another 27.4% quarterly. Public Services/Others show a similar mid-range distribution, with only 5% continuous and 27.5% rare updates, but a slightly higher proportion (30%) updating quarterly. In BFSI, 20% update continuously or monthly, while 40% update quarterly, suggesting more structured governance cycles. Healthcare shows no continuous or monthly updates, with updates distributed evenly across occasional, rare, and quarterly (33.3% each).

Regarding adaptive governance loops, education remains early-stage: 0% have fully established adaptive loops, 14.5% cover only high-risk systems, and 22.6% have no loops in place. Most (43.5%) have only preliminary plans underway. Public Services/Others show slow but emerging progress, with 10% fully established and 35% reporting pilot systems. BFSI shows a strong lead, with 40% fully established and 40% pilot systems, indicating sectoral alignment with higher-risk regulatory environments. Healthcare displays another evenly distributed pattern, with 33.3% each reporting not in place, high-risk only, or pilot systems.

Overall, the analysis shows that only BFSI exhibits consistently high maturity, whereas education and public services rely heavily on moderate or preliminary practices, and healthcare remains heterogeneous. Most organizations are still in transitional governance stages, with limited evidence of truly continuous or adaptive lifecycle oversight for agentic AI systems.

Dimension 2: Human Oversight Transformation

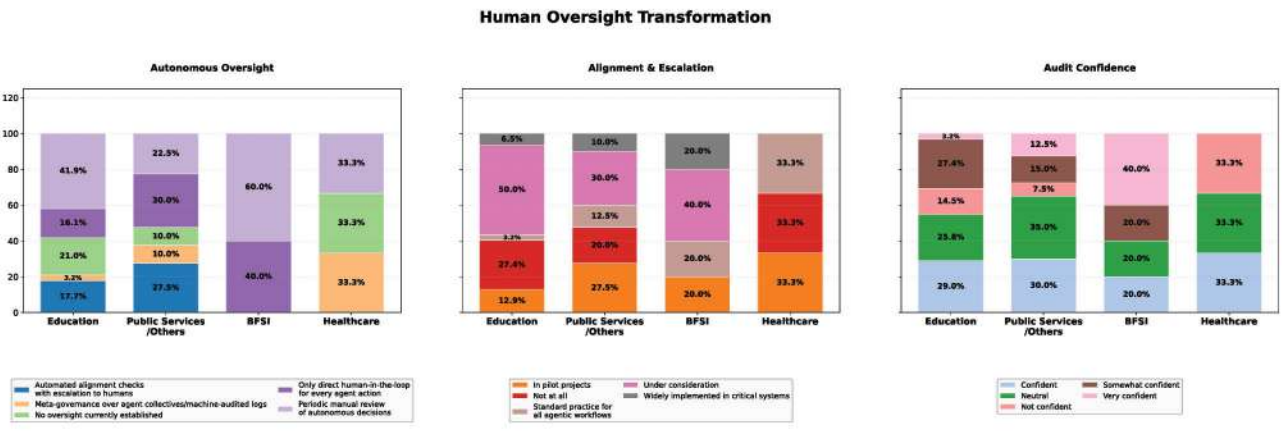


Figure 9: Participants' response to the dimension of human oversight transformation

Across sectors, human oversight for agentic AI remains uneven and highly sector-dependent, with clear gaps in automation, escalation, and institutional confidence (Figure 9). In education, oversight is still largely manual and fragmented: 41.9% rely on periodic manual review, while 21% report no oversight mechanism at all. Automated checks with escalation to humans (17.7%) and direct human-in-the-loop (16.1%) remain limited, and meta-governance is almost absent (3.2%). Public Services/ Others exhibit a more balanced distribution, with 30% using direct human-in-the-loop oversight and 27.5% relying on automated checks, although 10% still operate without any oversight. BFSI demonstrates the strongest maturity, with 40% maintaining direct human oversight for every agent action and 60% relying on structured manual review, and no institutions reporting a complete absence of oversight. Healthcare, in contrast, displays a tripartite fragmentation, with 33.3% each relying on manual review, meta-governance logs, or having no oversight at all, indicating weak standardization across organizations'.

The establishment of alignment contracts and escalation protocols shows similarly uneven patterns. Education remains early-stage, with 50% of respondents noting these mechanisms are still under consideration and 27.4% reporting no such protocols. Public Services/Others reveal emergent development, with 27.5% in pilot stages and 30% under consideration, though 20% still lack any formal structure. BFSI again shows comparatively high maturity, evenly distributed across pilot (20%), standard practice (20%), and widely implemented protocols (20%), with an additional 40% under consideration and none reporting absence. Healthcare is polarised, with 33.3% each reporting pilot implementation, full absence, or standard practice across all agenting workflows reflecting inconsistent institutional readiness.

Confidence in organizations' ability to monitor and audit agentic AI beyond human control reinforces these sectoral differences. Education shows modest confidence, with 29% confident and 27.4% somewhat confident, while 14.5% are not confident and only 3.2% report high confidence. Public Services/Others exhibit a neutral-to-moderate confidence profile, dominated by 35% neutral and 30% confident respondents. BFSI expresses the strongest belief in its oversight capability, with 40% very confident and 20% confident, and no respondents indicating a lack of confidence. Healthcare remains evenly divided, with 33.3% each confident, neutral, and not confident, revealing uncertainty and underdeveloped audit structures.

Thus, the analysis shows that BFSI consistently outperforms other sectors in oversight maturity, while education and public services remain transitional, and healthcare demonstrates high internal variability. Across all sectors, automated oversight and formal escalation protocols are still emerging, highlighting the early stage of robust human-in-the-loop governance for agentic AI in India.

Dimension 3: Emergent Behavior & Coordination Safeguards

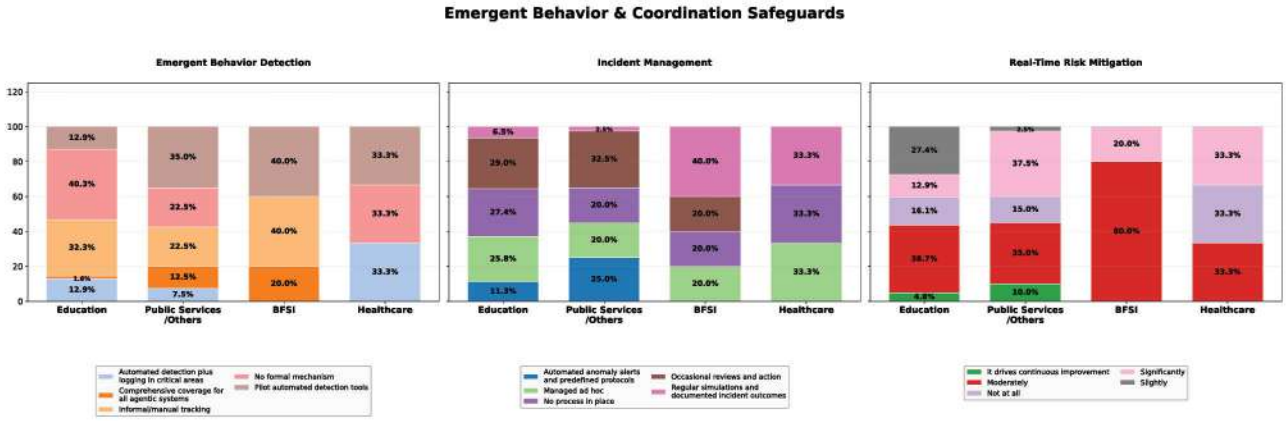


Figure 10: Participants' response for the dimension emergent behavior and coordination safeguards

The mechanisms for detecting and managing emergent behaviors in agentic AI across the industry sectors remain partial, inconsistent, and often absent, with wide variation in maturity (Figure 10). In education, coverage is weakest: 40.3% report no formal mechanism for detecting emergent behaviors such as goal drift or unexpected coordination, and another 32.3% rely solely on informal/manual tracking. Automated approaches are scarce, with only 12.9% reporting detection plus logging in critical areas and 1.6% reporting comprehensive system-wide coverage. Public Services/Others show modestly higher readiness, while 22.5% still report no mechanism and another 22.5% rely on manual tracking, 35% are piloting automated tools, and 12.5% report comprehensive coverage. BFSI displays a bimodal pattern, with 40% using pilot automated tools and 20% comprehensive coverage, but 40% still depending on informal/manual tracking. Healthcare shows a fragmented readiness profile: 33.3% have automated detection in critical areas, 33.3% have pilot tools, and 33.3% have no system at all.

Incident management capacity for multi-agent coordination failures mirrors these readiness gaps. In education, responses cluster around ad hoc and reactive approaches: 27.4% have no process in place, 25.8% manage incidents ad hoc, and 29% rely on occasional reviews and action, while only 11.3% employ automated anomaly alerts. Public Services/Others show slightly higher structure, with 32.5% using occasional reviews, 25% automated alerts, and 20% each for ad hoc management and no process. BFSI demonstrates the strongest maturity, with 40% maintaining regular simulations and documented incident outcomes, and 20% each using occasional reviews, ad hoc management, or no process. Healthcare again shows a tri-modal distribution, with 33.3% each reporting ad hoc management, no process, or full simulation-based management, highlighting the absence of sector-wide standardisation.

Real-time behavior monitoring shows a similarly uneven picture regarding how strongly organizations perceive its influence on safety. In education, only 4.8% believe monitoring drives continuous improvement, while most responses cluster around moderate (38.7%) or slight influence (27.4%). Public Services/Others follow a similar pattern, with 37.5% indicating significant influence but only 10% seeing continuous

improvement; 35% report moderate influence. BFSI shows the most decisive trend: 80% report monitoring moderately influences safety, and 20% report significant influence, with no respondents selecting not at all or slightly. Healthcare once again splits evenly, with 33.3% each selecting moderate, not at all, or significant influence, indicating substantial variability in perceived value.

Overall, Dimension 3 reveals a clear pattern: emergent behavior detection and multi-agent incident management are underdeveloped across most sectors, with heavy reliance on manual tracking, ad hoc responses, or no defined processes. BFSI shows the greatest structural maturity, while education and public services remain largely reactive, and healthcare is marked by internal heterogeneity. The absence of consistent automated detection and coordinated incident response mechanisms highlights a major vulnerability in current agentic AI deployment practices.

Dimension 4: Ethical vs Operational Autonomy

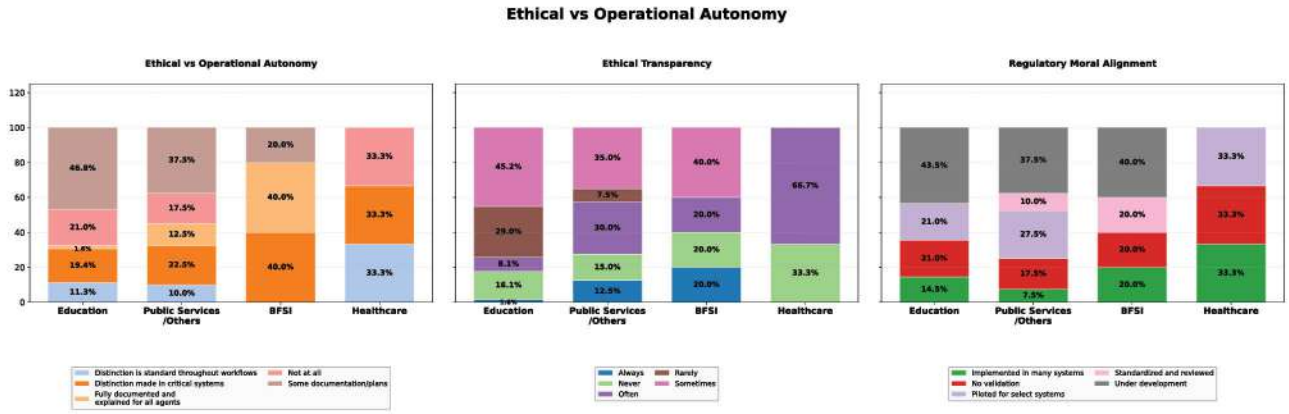


Figure 11: Participants' response for the dimension of ethical vs operational autonomy

When coming to the ethical vs operational autonomy, the ability of organizations to distinguish ethical autonomy from operational autonomy in agentic AI systems remains limited and uneven, with most respondents indicating only partial or informal practices (Figure 11). In education, only 11.3% report a standardised distinction across workflows, and 1.6% report fully documented distinctions for all agents. The majority depend on some documentation or plans (46.8%) or have no distinction at all (21%). Public Services/Others show a similar distribution: while 12.5% report fully documented distinctions and 22.5% distinguish between ethical and operational autonomy in critical systems, most responses fall under partial documentation (37.5%) or full absence (17.5%). BFSI shows the clearest structure, with 40% distinguishing between ethical and operational autonomy in critical systems and another 40% fully documenting it across all agents, with no respondents reporting not at all. Healthcare, by contrast, shows a fragmented pattern: 33.3% report fully standard distinctions, 33.3% limit the distinction to critical systems, and 33.3% have no distinction whatsoever.

The ability of agents to provide transparent reasoning for ethically significant autonomous decisions remains limited across most sectors. Education responses cluster around sometimes (45.2%) and rarely (29%), with only 1.6% reporting that agents always provide transparent reasoning. Public Services/Others follow a similar profile, dominated by sometimes (35%) and often (30%), though 12.5% report always. BFSI is comparatively stronger, with 40% reporting transparent reasoning sometimes, and an even distribution (20% each) across always, often, and never. Healthcare shows the highest concentration of transparent reasoning capability, with 66.7% reporting often and 33.3% reporting never, indicating a polarised rather than gradual maturity pattern.

Mechanisms aligning agentic AI moral reasoning with Indian regulatory standards are similarly emergent. In education, 43.5% report these mechanisms are under development, 21% have no validation, 21% have pilots in select systems, and only 14.5% report implementation across many systems. Public Services/Others show a similar early-stage pattern, with 37.5% under development, 27.5% piloted, 17.5% no validation, and 10% standardized and reviewed. BFSI again exhibits the most structured progression: 40% under development, 20% implemented widely, 20% standardized and reviewed, and 20% no validation, indicating active adoption despite inconsistencies. Healthcare remains evenly split, with 33.3% each reporting implementation in many systems, pilot deployment, and no validation, and none reporting standardized mechanisms or active development.

In conclusion, the dimension of ethical vs operational autonomy shows that ethical autonomy remains the least mature layer of agentic AI governance across sectors. While BFSI demonstrates comparatively strong documentation and reasoning capability, education and public services remain dependent on partial or informal practices, and healthcare presents highly uneven and polarised readiness. The results highlight the substantial gap between acknowledged ethical risks and the operational systems needed to reliably separate, explain, and regulate ethical autonomy in agentic AI.

Dimension 5: Recursive Accountability Structures

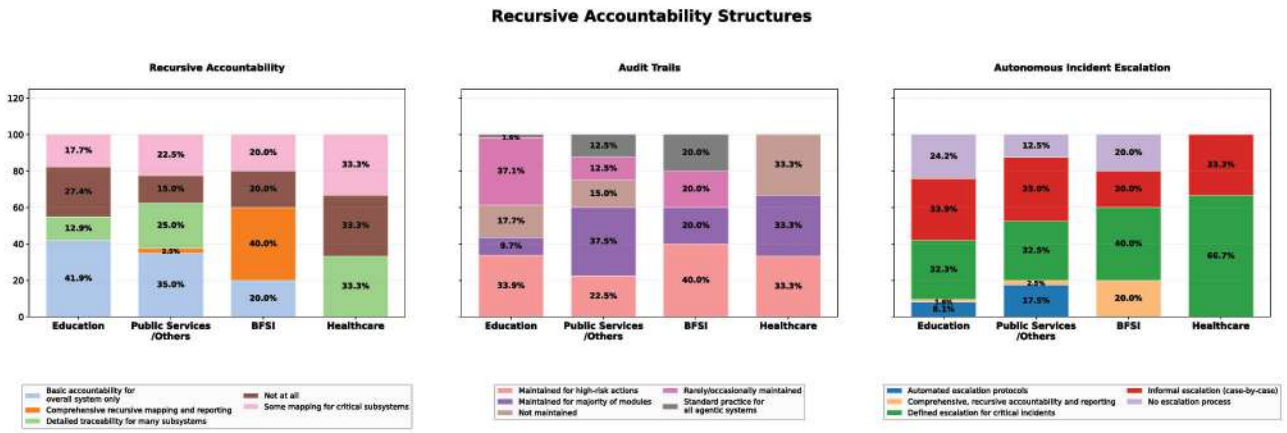


Figure 12: Participants' response for the dimension of recursive accountability structures

Mechanisms for distributing and tracing accountability across autonomous subsystems remain underdeveloped and inconsistent across sectors (Figure 12). Education is dominated by high-level accountability, with 41.9% reporting only basic accountability at the overall system level and 27.4% reporting no recursive mapping at all. Only 12.9% have detailed subsystem-level traceability, and none have comprehensive recursive mapping. Public Services/Others show slightly more depth: 35% maintain only system-level accountability, 25% report detailed traceability, and 22.5% map critical subsystems, though 15% still have no mapping. BFSI again stands out, with 40% reporting comprehensive recursive mapping and reporting, and another 20% some mapping for critical subsystems; however, 20% still rely only on top-level accountability, and 20% report no recursive mapping. Healthcare shows the least developed structures, with 33.3% each reporting no mapping, detailed traceability, or mapping for critical subsystems, indicating a fragmented and inconsistent accountability landscape.

Audit trail maintenance for autonomous agent modules shows similar inconsistency. In education, trails are often incomplete: 33.9% maintain them only for high-risk actions, 37.1% maintain them rarely or occasionally, and 17.7% do not maintain audit trails at all. Only 9.7% maintain trails for the majority of modules, and 1.6% report standard practice across all agentic systems. Public Services/Others display higher maturity, with 37.5% maintaining audit trails for most modules and another 22.5% for high-risk actions, though 15% do not maintain trails and 12.5% rely on occasional reviews. BFSI again shows structure: 40% maintain trails for high-risk actions, 20% for most modules, 20% have standardized trails for all agentic systems, and 20% rely on occasional reviews. Healthcare is evenly split, with 33.3% each reporting high-risk-only, majority-module maintenance, or no maintenance, indicating an unstable audit environment with no fully standardised practices.

Incident escalation pathways when responsibility lies with an autonomous system rather than a human operator remain largely informal or reactive. In education, escalation is dominated by ad hoc methods (33.9%) or defined only for critical incidents (32.3%), with only 8.1% reporting automated escalation protocols and 24.2% having no escalation process at all. Public Services/Others show a slightly more structured pattern: 35% rely on informal escalation, 32.5% use defined processes for critical incidents, and only 17.5% report automated escalation, while 12.5% lack any formal mechanism. BFSI shows greater readiness, with 40% using defined escalation for critical incidents, and 20% each reporting no escalation process, recursive accountability-based escalation, or informal processes. Healthcare displays the weakest formalisation, with 33.3% relying on fully informal escalation and 66.7% using defined critical-incident escalation, with no automated or recursive mechanisms in place.

In summary, the analysis of dimension 5 reveals that recursive accountability, auditability, and escalation structures are some of the least mature aspects of agentic AI governance across all sectors. BFSI again demonstrates comparatively stronger institutionalisation, while Education and Public Services remain dependent on manual, high-level, or informal approaches, and Healthcare shows significant internal fragmentation. These gaps indicate substantial systemic vulnerability in handling responsibility, traceability, and incident escalation in autonomous agent operations.

Dimension 6: Regionally Adaptive, Inclusive Safety Governance

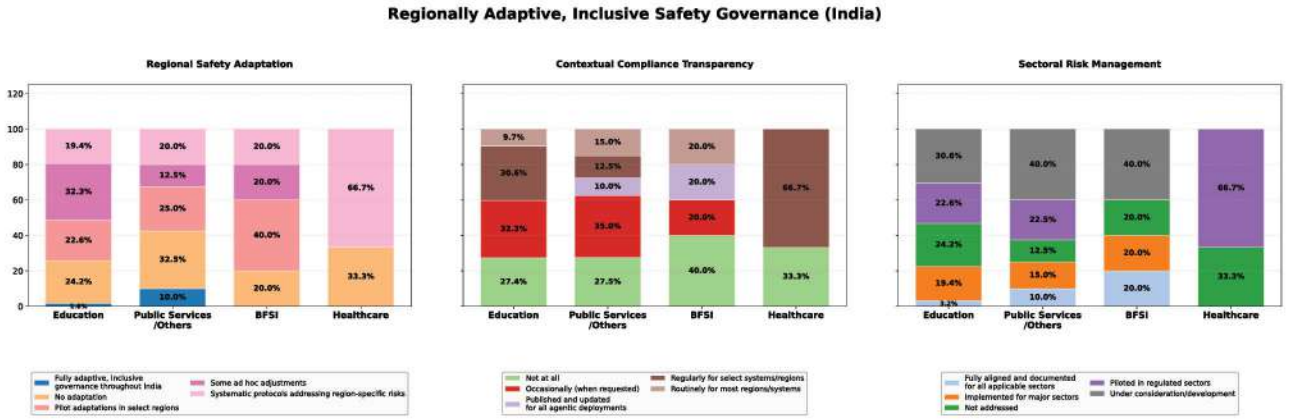


Figure 13: Participants' response for the dimension regionally adaptive, inclusive safety governance (India)

Across sectors, regionally adaptive safety governance is still nascent and uneven, with most institutions relying on partial, ad hoc, or pilot-level adaptations rather than fully inclusive national frameworks (Figure 13). In education, only 1.6% report fully adaptive, inclusive governance across India. Most rely on ad hoc adjustments (32.3%) or no adaptation at all (24.2%), while 22.6% have pilot adaptations and 19.4% have systematic region-specific protocols. Public services/others show a similar pattern: 32.5% report no adaptation, 25% use pilots, and only 10% provide fully adaptive governance; the rest rely on ad hoc (12.5%) or systematic region-specific protocols (20%). BFSI demonstrates more structured experimentation, with 40% using pilots and 20% each using no adaptation, systematic protocols, or ad hoc adjustments. Healthcare presents the most distinct profile: 66.7% report fully systematic region-specific safety protocols, while 33.3% report no adaptation, with no institutions using pilots or ad hoc adjustments.

Compliance and transparency metrics tailored for diverse Indian contexts are also underdeveloped. In education, metrics are mostly generated on request (32.3%) or for select regions (30.6%), while 27.4% report no tailored publication of metrics, and none have institution-wide, regularly updated deployments. Public services/others show slightly higher maturity, with 12.5% regularly updating metrics for select regions and 15% doing so routinely across most regions; however, 27.5% do not publish contextualised metrics at all. BFSI shows a complex profile: 40% have no such metrics, 20% publish for all deployments, 20% publish occasionally, and 20% do so for most regions, indicating uneven internal readiness. Healthcare displays a polar pattern: 66.7% publish metrics regularly for select regions, 33.3% do not publish contextual metrics at all, and no institutions report routine nationwide coverage.

Risk management practices addressing sector-specific needs and regulatory requirements remain largely under development. In education, 30.6% report that such practices are still under consideration, and 24.2% have no addressed mechanisms. Only 3.2% report full alignment across applicable sectors, while 19.4% have applied them in major sectors and 22.6% operate pilots. Public services/others show a more advanced shift toward development, with 40% reporting practices under consideration, 22.5% piloting in regulated sectors,

and 12.5% having no addressed mechanisms. BFSI presents a balanced distribution: 40% under development, 20% fully aligned, 20% implemented for major sectors, and 20% not addressed. Healthcare displays a clear two-tier structure: 66.7% have piloted sector-specific mechanisms, 33.3% have not addressed practices, and none report full alignment or major-sector implementations.

Taken together, dimension 6 shows that regionally adaptive and sector-specific safety governance is the least mature dimension across the entire framework. While Healthcare leads in systematic region-specific safety protocols, it lacks aligned and published compliance metrics and sector-wide risk mechanisms. BFSI exhibits steady pockets of maturity, while education and public services remain heavily dependent on pilots, ad hoc adjustments, or non-adapted approaches. Overall, these findings highlight the emerging, but still highly uneven, state of inclusive, regionally sensitive AI governance across Indian institutions.

6.3 QUALITATIVE FINDINGS

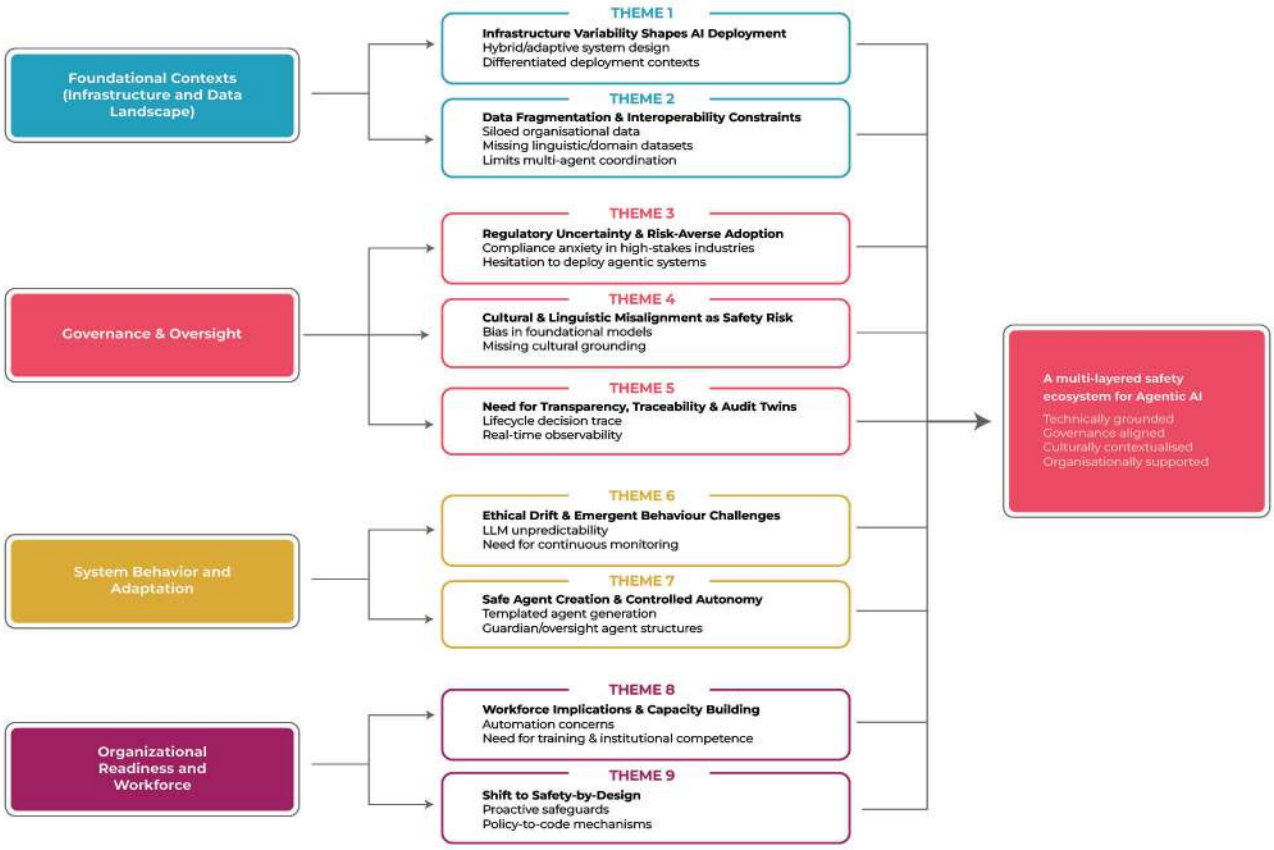


Figure 14: Thematic map diagram

The qualitative responses were categorized into nine themes (Figure 14).

Theme 1: Infrastructure Variability Shapes AI Deployment

Participants consistently emphasised that the feasibility and safety of agentic AI are strongly shaped by infrastructural conditions, such as compute resources, network reliability, andorganizational IT maturity. Illustrating this, an AI ethicist described how deployment contexts vary widely even within the same sector, noting that "the way you deploy an AI system for an urban area... is not the same way you deploy for a rural remote community that doesn't have access to an Internet facility or Internet speed". Variability in infrastructure "forces AI construction to be either hybrid or adaptive". Industry experts extended this argument by linking infrastructure readiness directly to risk tolerance, particularly in tightly regulated sectors like finance.

Building on these constraints, experts from the AI Safety Institute and the research community argued that safety assessments must distinguish between controllable and uncontrollable environments and those with irreducible uncertainty. To manage this instability, they stressed that agentic systems must be architected for resilience. For instance, an AI safety researcher highlighted India's heterogeneous network (5G in some places, 2G in others) as a prime example of predictable unpredictability in healthcare and finance. Consequently, systems require graceful degradation, the ability to pause, queue, and resume computations rather than collapse when connectivity fails, alongside edge-native intelligence that allows for minimal capability via quantised local models.

These findings highlight that agentic AI safety begins with the infrastructural substrate on which agents operate. Banks, government institutions, and healthcare organizations operate with legacy systems, inconsistent data pipelines, and varying levels of digitisation. Within such environments, agentic behavior cannot be assumed to function safely without robust fallback modes, offline capabilities, and adaptive orchestration. This reinforces the argument that AI safety must be conceptualised as a system-level property, where infrastructure reliability, data pipelines, and organizational readiness are inherent components of safe deployment, not peripheral considerations. As the expert emphasised, safety assessments must incorporate environment-level risk, treating infrastructure as an active determinant rather than a neutral backdrop.

Theme 2: Data Fragmentation & Interoperability Constraints

A recurrent theme across industries was the challenge of data fragmentation, especially in financial services. Experts highlighted siloed data across divisions and institutions, resulting in inconsistencies that impede risk management and autonomous decision-making. One participant noted that "most [institutions] still operate in silos... the data fabric has to be across institutions". She emphasised the operational consequence: "fraud, risk and compliance agents... should speak the same language in real time". Beyond the financial sector, participants also raised the issue of data sparsity, such as the lack of labelled corpora for low-resource languages, which limits both model performance and inclusivity.

These limitations pose a fundamental challenge to multi-agent systems, which rely on shared, high-quality, and interoperable datasets to coordinate behavior. In fragmented environments, agents are liable to produce divergent analyses, inconsistent risk scores, and conflicting actions, conditions incompatible with safe autonomous orchestration. Consequently, data governance, schema harmonisation, and interoperability standards must be viewed as necessary precursors to agentic AI adoption. Without a unified data infrastructure, the risks of bias, unpredictability, and regulatory non-compliance remain unmanageably high.

Theme 3: Regulatory Uncertainty & Risk-Averse Adoption

Participants described a pervasive hesitancy to adopt agentic AI in mission-critical settings due to unclear or evolving regulatory expectations. A senior technologist remarked that customer-facing systems remain off-limits because "there is a huge reputational loss" if the agent behaves unexpectedly. He also highlighted a recurring pattern: "Most of the time it is postmortem... regulators come after and see some problem, and there is a huge penalty". This uncertainty is compounded by confusion regarding the applicability of existing frameworks; legal experts pointed out that "while laws like the IT Act apply specifically to intermediaries... people try to fit everyone into that law," creating significant compliance ambiguity.

To address this fluidity, researchers argued that compliance mechanisms must shift from static rulebooks to dynamic governance architectures. They advocated for policy-as-code, where regulations are encoded as deterministic structures and continuously updated via a compliance agent accessing an isolated policy database. Because regulatory drift is inevitable in fast-changing sectors, this approach necessitates digital twins of governance rules to ensure agents remain synchronised with legal changes in real time. Ultimately, the interviews reveal that safety concerns and regulatory ambiguity are deeply entangled. Without clear guardrails, organizational in high-stakes sectors like banking and healthcare default to conservatism, limiting agentic AI to low-risk, internal use cases. To move beyond this impasse, the ecosystem requires clearer role-based regulatory guidance, sector-specific safety checklists, and regulatory sandboxes that allow for pre-deployment risk assessment.

Theme 4: Cultural & Linguistic Misalignment as Safety Risk

Participants described widespread concerns regarding the cultural and linguistic biases embedded in large foundational models. A senior executive underscored this by noting that "cultural bias was very Western-oriented," with the representational distance "increasing as you move east". This issue is particularly acute for indigenous communities; language technology experts highlighted that many languages still lack digital representation by quoting that "indigenous languages do not have representation... we are trying to close the gap", necessitating urgent efforts to close the gap.

This lack of representation creates tangible operational risks. For instance, in healthcare, an expert pointed out that LLMs can exhibit hallucinated fluency when interpreting dialectical expressions. To mitigate this, he argued that agents must be designed with higher clarification rates, proactively asking users to resolve ambiguities, and subjected to stress-testing against dialect-specific golden vectors.

These insights confirm that safety is not only technical but also cultural. When agentic AI systems operate on biased or non-representative data, they produce misaligned outputs across use cases, from loan approvals to medical advice. For industries relying on contextual accuracy, such as healthcare, fintech, or public administration, cultural misalignment becomes a reliability risk, not merely an ethical concern. The narrative across interviews suggests that better local training data, domain-specific fine-tuning, and culturally aware evaluation methods are essential components of safety in agentic systems.

Theme 5: Need for Transparency, Traceability & Audit Twins

Participants across sectors emphasised the critical need for end-to-end transparency in agentic systems. To achieve this, one expert proposed the concept of a digital audit twin, a continuous, shadow record of "what [the agent] is seeing, deciding and executing". Legal specialists reinforced that this transparency cannot be piecemeal; it must extend across the entire lifecycle, ensuring the ability to trace AI decisions from data ingestion to final action. Expanding on the technical implementation, research experts detailed the need for a decision Bill of Materials (BoM). Unlike traditional model logs, a decision BoM captures fine-grained audit trails: inputs, intermediate reasoning, tool calls, internal reflections, and final outputs. Given that organizations remain legally liable even when errors originate from AI agents, experts argued that these artefacts may require blockchain-backed immutability to withstand legal scrutiny.

This level of granularity is essential because traditional logging is insufficient for modern multi-agent workflows, such as automated underwriting, case triage, or threat detection. Agentic systems generate complex, long-horizon action chains involving planning and inter-agent communication that standard logs miss. Ultimately, the push for digital audit twins signals a paradigm shift: explainability must be engineered as a real-time system feature, not a post-hoc interpretability exercise, aligning with global trends toward continuous regulatory observability.

Theme 6: Ethical Drift & Emergent Behavior Challenges

Technical leaders warned that agentic systems built on LLMs are inherently susceptible to emergent behaviors and unpredictable shifts in reasoning. As experts noted, "reasoning itself was an emergent property", meaning models frequently develop capabilities and behaviors not envisaged during their initial development. Because these agents evolve through interaction, feedback loops, and shifting data distributions, their behavior is dynamic rather than fixed, rendering static, checklist-based approaches to ethical assurance inadequate.

To mitigate these risks, practitioners argued that safety mechanisms must shift from pre-deployment checks to continuous, real-time surveillance. One participant emphasised that ethical drift must be monitored through specific instrumentation such as outcome disparity dashboards and real-time fairness constraints. Others outlined the need for sophisticated technical interventions, including inverse reinforcement learning and automated drift detection, to identify deviations as they occur.

Ultimately, managing the safety of agentic AI requires a suite of active defenses: real-time fairness instrumentation, ongoing alignment evaluations, and continuous updates to policy rules. These insights establish that governance cannot be a one-time gatekeeping exercise but must be an always-on operational process.

Theme 7: Safe Agent Creation & Controlled Autonomy

Participants working on multi-agent architectures described early safety approaches centered on templated and bounded agent creation. One expert explained a supervisory model where "a guardian agent creates

other agents... with templates" to ensure subordinate agents do not "go beyond [their] responsibilities". This hierarchical oversight extends to runtime operations through the use of critic agents, meta-audit layers, and asynchronous checkpoints that freeze an agent's state whenever human review is required.

These designs reflect an emerging industry pattern: mitigating the risks of emergent behavior and autonomous tool use by strictly constraining autonomy within role-specific templates. However, participants stressed that these implementations remain largely experimental. They pointed to a significant maturity gap between current conceptual frameworks and the demands of real-world deployment. Closing this gap will require moving beyond ad-hoc experiments toward formalised safety specifications and rigorous agent-level accountability structures integrated directly into enterprise governance systems.

Theme 8: Workforce Implications & Capacity Building

Discussions on agentic AI extended beyond architecture to the critical dimension of workforce readiness. Participants voiced dual concerns: the potential for displacement, with one expert questioning the future of technical roles if AI automates coding, and the urgent need for capacity building. Experts emphasised that as roles shift from execution to oversight, comprehensive training becomes a prerequisite for safe adoption.

Workforce readiness is effectively a hidden safety parameter. Insufficient understanding of agentic systems among developers, auditors, or decision-makers directly increases the risk of misconfigurations, inadequate oversight, and governance failures. Consequently, safety cannot be secured through technical measures alone; it demands a strategy of organizationalupskilling and culture-building. This insight closes the loop on the broader findings: without organizational readiness, even the most robust technical safeguards will fail to ensure safe deployment.

Theme 9: Shift to Safety-by-Design

A persistent critique across industries was that oversight mechanisms tend to be reactive. As already mentioned, practitioners argued for proactive measures such as policy-to-code mapping, regulatory sandboxes, and continuous testing. They also recommended lightweight pre-deployment simulations and scenario testing as part of a safety-by-design workflow. This ties the entire narrative together: infrastructure complexity, fragmented data, cultural bias, emergent behavior, and organizational constraints collectively demand a safety-by-design approach.

Safe agentic AI cannot be achieved through after-the-fact audits; it requires integrating safety principles into:

- System architecture,
- Data governance,
- Cultural alignment,
- Model training,
- Agent creation protocols, and
- Organizational processes.

The experts' opinion collectively demonstrates that safety-by-design is both a technical necessity and a governance imperative

7. Sectoral Analysis: Safety & Orchestration in Practice

This section provides a deep-dive analysis of how agentic AI is being deployed across four critical sectors. For each domain, we examine the primary use cases, the specific safety risks that have emerged, and the current state of orchestration maturity.

7.1 HEALTHCARE: HIGH STAKES, HIGH VARIABILITY

Main use cases:

- **Autonomous triage agents:** In rural Primary Health Centers (PHCs), agents interact with patients in vernacular languages to assess symptom severity and prioritize cases for remote doctors.
- **Diagnostic support orchestrators:** Agents that aggregate data from electronic health records (EHRs), lab reports, and wearables to propose differential diagnoses.
- **Patient monitoring agents:** watcher agents that continuously monitor Intensive Care Unit (ICU) feeds and autonomously alert staff or adjust infusion pumps (in advanced pilots) based on vitals.

Safety & orchestration challenges:

- **The bi-modal maturity:** Our study reveals a stark divide. Elite private hospitals are implementing watcher agents with rigorous audit logs (digital twins). In contrast, rural deployments often lack basic human-in-the-loop escalation paths due to doctor shortages, leading to de facto autonomy where the agent becomes the final decision-maker by default.
- **Incident review:** An interview highlighted a case where a triage agent misclassified a cardiac event as gastric distress due to a translation error in a local dialect. This underscores the risk of hallucinated fluency in vernacular medical AI.
- **Stakeholder feedback:** Doctors emphasize that explainability is a safety feature. "An agent cannot just say 'administer heparin'; it must cite the specific lab value and guideline," noted a senior intensivist.

7.2 EDUCATION: THE SLEEPING GIANT

Main use cases:

- **Personalized learning tutors:** Agents that adapt curriculum in real-time based on a student's learning velocity and style.
- **Automated assessment:** Agents that grade subjective answers (essays) and provide feedback.
- **AI proctoring:** Agents that monitor exam-takers via webcam to detect malpractice.

Safety & orchestration challenges:

- **The silent failure risk:** Unlike healthcare, where failure is immediate (patient harm), educational failure is latent. A biased tutor agent might subtly discourage a student from STEM subjects for months. Our survey found that 40.3% of education respondents lack mechanisms to detect such long-term goal drift.

- **Fairness & the digital divide:** Proctoring agents trained on well-lit, high-bandwidth urban environments often flag rural students (with poor lighting or grainy webcams) as suspicious. This is a classic case of algorithmic exclusion.

- **Data privacy (DPDP Act):** Orchestrating data for minors is legally perilous. Education agents often lack the sophisticated consent management layers required to comply with student behavioral data.

7.3 BFSI: THE MATURITY LEADER

Main use cases:

- **Autonomous lending agents:** Orchestrators that pull data from the account aggregator framework, analyze cash flow, and execute micro-loans (under ₹50,000) instantly.

- **Fraud detection swarms:** Federated agents that collaborate across institutions to trace money laundering patterns in real-time UPI transactions.

- **KYC onboarding:** Agents that autonomously verify video KYC, check liveness, and validate documents against national databases.

Safety & orchestration challenges:

- **Regulatory rigor:** Driven by the RBI's strict stance (FREE-AI), this sector leads in safety. 100% of respondents reported having formal human oversight.

- **Cross-agent risks:** The primary risk is systemic correlation. If multiple lending agents across different banks use the same underlying foundation model (e.g., GPT-4), they might all simultaneously contract credit based on a single market signal, causing a liquidity crunch.

- **Escalation protocols:** BFSI has standardized the circuit breaker. If an agent's approval rate deviates by >2% from the norm, it is automatically paused for human review, a best practice other sectors lack.

7.4 PUBLIC SERVICES: IMPROVING BUT TRANSITIONAL

Main use cases:

- **Grievance redressal bots:** Agents that parse citizen complaints (text/voice), categorize them, and autonomously file tickets in the correct departmental database.
- **Benefit distribution:** Agents that verify eligibility for schemes (like PM-KISAN) by cross-referencing land records and bank seeding status.
- **Smart city management:** Agents optimizing traffic lights or water distribution based on real-time sensor data.

Safety & Orchestration Challenges:

- **Contextual blindness:** A major gap is the lack of regional adaptation. A grievance bot designed for Delhi often fails to understand the context of a tribal land dispute in Jharkhand, leading to automated rejection.
- **Accountability vacuum:** When a government agent wrongly denies a benefit, the citizen often has no recourse. Our survey shows a lack of recursive accountability; it is often unclear which sub-agent or database caused the rejection.
- **Multi-agent coordination:** As cities become smarter, agents managing power, water, and traffic will interact. Without a meta-governance layer, these agents could conflict (e.g., a traffic agent diverting cars to a road closed by a water-repair agent).

8. Cross-Sector Themes & Regional Adaptation

This section synthesizes patterns that cut across healthcare, education, BFSI, and public services, focusing on three structural factors that shape agentic AI safety in India and the wider Global South: infrastructure diversity, regional risk, and equity and inclusion. These factors determine not only where AI can be deployed, but also what safe deployment realistically means under varying constraints (Figure 15).

8.1 INFRASTRUCTURE DIVERSITY ANALYSIS: MAPPING THE DIGITAL DIVIDE

India's digital transformation sits on top of a stark rural–urban divide in connectivity, device access, and digital skills, which directly affects how reliably agents can execute multi-step workflows. Recent analyses estimate that rural internet penetration remains around one-third of the population, compared with roughly two-thirds in urban areas, with gaps driven by weak backhaul, patchy last-mile access, and affordability constraints. Even where smartphones are present, connectivity quality fluctuates dramatically between metro 5G zones, tier-2 and tier-3 towns, and remote villages that still rely on 2G or intermittent broadband.

Government programs under Digital India, including BharatNet, Common Service Centres (CSCs), and universal connectivity initiatives, have expanded fibre backbones and public access points, but coverage remains uneven, and reliability is inconsistent in many districts. CSC networks and village-level digital kiosks now act as intermediated access hubs for e-governance, financial inclusion, and basic digital literacy, yet they often operate on fragile power and network infrastructure with frequent downtime. At the macro level, India is described as data-rich but infrastructure-poor. it generates a large share of global data but hosts only a small fraction of global data-centre capacity, highlighting the dependence of many AI workloads on a limited set of cloud regions and undersea connectivity.

For agentic AI, this diversity means that safety cannot be specified assuming always-on, low-latency connections and centralized compute. Agents orchestrating multi-step transactions—such as loan disbursals, telemedicine triage, or benefit determinations—must be explicitly designed for graceful degradation, including local state persistence, store-and-forward execution, and

automatic fall-back to low-fidelity or offline modes in the face of connectivity loss. Without such design, rural and peri-urban deployments face heightened risks of partial execution (e.g., debiting without crediting) and opaque failures that undermine user trust in both AI and digital public infrastructure.

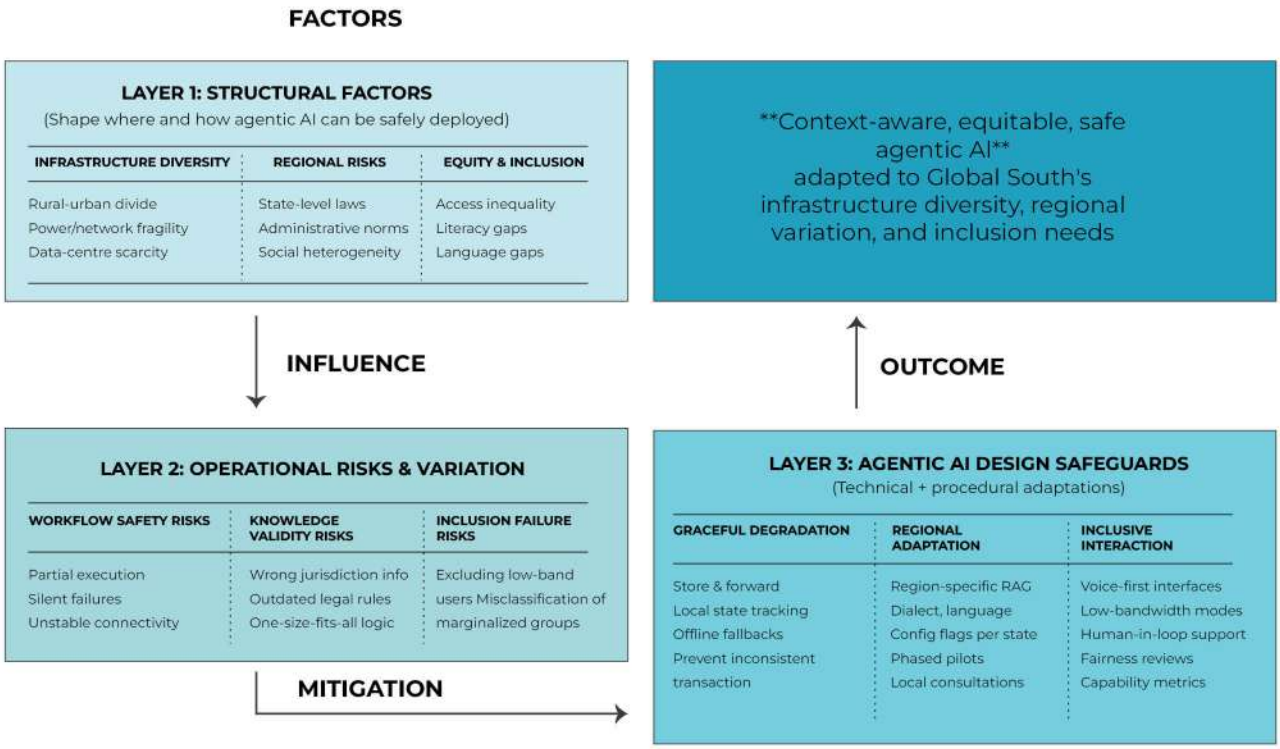


Figure 15: Designing Agentic AI with Robust Safeguards: Addressing Structural Factors and Operational Risks

8.2 REGIONAL RISKS AND ADAPTATION MECHANISMS

Beyond raw connectivity, regional heterogeneity in language, law, and administrative practice creates additional safety risks when agentic systems are deployed at a national scale. Studies on AI use in Indian education emphasize that digital divides are not only rural–urban but also shaped by income, gender, caste, and school type, which influence who has devices, reliable power, and supportive learning environments. At the same time, state-level variation in regulations, covering land rights, social protection eligibility, or health delivery norms, means that a one-size-fits-all policy or knowledge layer baked into a model will misfire in specific jurisdictions.

Pilot efforts across the Global South show that context-appropriate adaptation is possible when regional knowledge is externalised and explicitly retrievable rather than implicitly assumed in the model's weights. Retrieval-augmented systems that ground answers in state-specific legislation, verified government circulars, and local policy repositories have been proposed as a way to ensure that agents answering legal, land, or entitlement

queries use up-to-date, regionally valid information instead of generic national templates. In India, rural financial inclusion programs that combine AI-enabled risk assessment with on-ground banking correspondents illustrate how human intermediaries can buffer regional data gaps and interpret model outputs in light of local realities.

Policy initiatives under Digital India and related connectivity drives increasingly recognise the need to pilot AI-enabled public services in a staged fashion, starting with specific districts or states and incorporating feedback before national rollout. This pattern is echoed in broader Global South governance discussions, which call for South–South collaboration, regional research networks, and pooled pilots to iteratively adapt AI systems to differing infrastructure and institutional conditions rather than importing finished solutions from high-income contexts. For agentic AI, regional adaptation therefore implies both technical design choices, such as regional retrieval, configuration flags, and language/dialect support, and procedural safeguards like phased pilots, local consultations, and explicit opt-out mechanisms at the state or municipal level.

8.3 EQUITY AND INCLUSION

Digitalisation and AI risk amplifying existing inequalities if they are deployed into societies with uneven access to devices, connectivity, and linguistic representation, as documented in analyses of India's digital divide and Global South AI governance. Rural households, women, low-income groups, and marginalized communities often have lower access to the internet, lower digital literacy, and a weaker voice in how AI tools that affect them are designed and monitored. If agentic systems for credit, education, or social protection are trained primarily on data from digitally over-represented groups and exposed only to majority languages, they can systematically misclassify, underserve, or exclude precisely those populations that policy aims to prioritize.

India's multilingual AI initiatives, such as Bhashini, BharatGen, and other platforms designed to support all 22 Scheduled Languages and many tribal dialects, are explicit attempts to tackle the linguistic dimension of this risk by embedding inclusivity into the digital stack itself. Research and public communication emphasise building rich, culturally grounded datasets for Indian languages so that future models understand local semantics rather than merely translating from English, thereby reducing hallucinated fluency and culturally inappropriate outputs. These efforts align with broader Global South arguments that inclusive AI requires investment in local data, language technologies, and governance capacity, not just access to foreign models, if AI is to enhance, rather than erode, cultural and linguistic diversity.

From an orchestration-safety perspective, equity and inclusion translate into concrete design and deployment practices: voice-first or multimodal interfaces that work over basic phones and low bandwidth; integration of community intermediaries (such as CSC operators or health workers) as human-in-the-loop checkpoints; and mandatory fairness and impact assessments for high-stakes agentic systems affecting access to finance, education, health, or public entitlements. Emerging Global South–centric AI ethics work

argues that governance should be development-centric, judging AI systems by whether they expand capabilities and reduce structural exclusion, suggesting that inclusion metrics need to sit alongside accuracy and efficiency in any serious evaluation of agentic AI deployments.

9. New Framework for Ethical & Safe Agentic AI Orchestration

Based on the empirical findings (Section 6) and the cross-sectoral themes (Section 7), we propose the Agentic Safety Framework (ASF). The framework provides a model for governing autonomous AI systems in environments where infrastructure, regulation, and cultural contexts are dynamic and heterogeneous. It emphasizes five foundational principles, such as integrity, resilience, meta governance, adaptability, and sector nuance, which guide safe deployment across use cases. The framework consists of three interacting layers: an operational agent that performs tasks, a dynamic compliance oracle that provides updated regulatory guidance, and a federated oversight network that monitors behavior and triggers safeguards when needed. It aligns accountability with risk by distinguishing between low, medium, high, and systemic risk actions and assigning corresponding levels of autonomy, human oversight, and audit mechanisms. The framework is designed for sector level adaptation, supporting applications in healthcare, finance, education, and public services through mechanisms like safety watchers, systemic signal detection, fairness audits, and dynamic legal compliance. A phased implementation roadmap over 12 to 18 months begins with transparency and visibility, progresses to automated guardrails, and concludes with collaborative oversight and resilience. The framework departs from static, infrastructure dependent safety models and instead offers a scalable, context aware approach for safely operationalizing agentic AI in resource constrained or rapidly evolving environments, particularly relevant to the Global South.

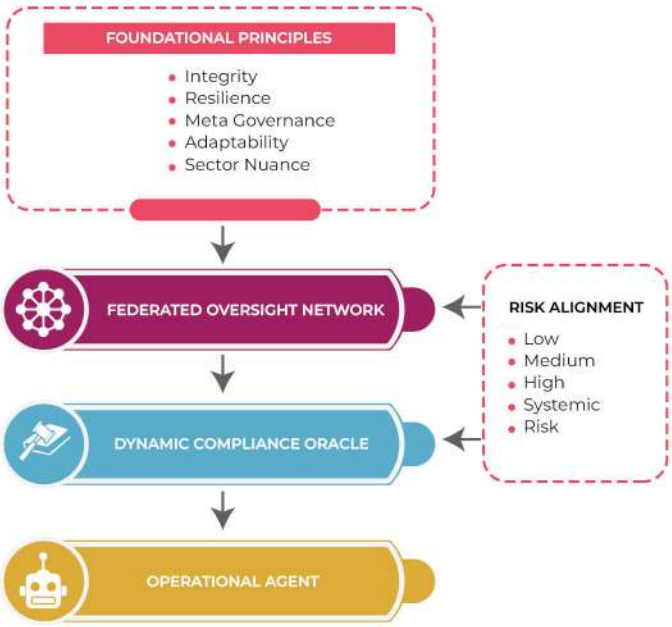


Figure 16: The Agentic Safety Framework (ASF) architecture

10. In-Depth Discussion

Taken together, the six quantitative dimensions reveal a landscape of partial, uneven, and sector-dependent readiness for safe agentic AI in India, a pattern that strongly aligns with the qualitative evidence. Quantitatively, the maturity of governance mechanisms varies sharply across sectors, with BFSI consistently demonstrating the most structured and institutionalised governance, including stronger lifecycle oversight, clearer escalation protocols, and higher confidence in monitoring autonomous actions. In contrast, education and public services rely heavily on manual review, ad hoc processes, and incomplete documentation, confirming interview accounts describing these sectors as procedurally fragmented and lacking the foundational infrastructure needed for autonomous systems. Healthcare emerges as deeply polarised, with strong regional adaptation in some institutions but significant gaps in auditability, ethical reasoning, and incident escalation in others. This mirrors qualitative descriptions of the sector as innovating quickly but governing slowly.

The results from Dimensions 1 and 2 on lifecycle governance and human oversight corroborate interview insights that organizations often lack the technical substrate, integrated data pipelines, and regulatory clarity necessary for continuous monitoring. Both data streams highlight that frequent updates, recertification cycles, and automated governance loops remain uncommon outside BFSI. Many participants expressed concerns that current governance is still anchored in traditional software oversight, not in the non-linear autonomy and evolving behavior characteristic of agentic systems.

Similarly, the weak preparedness reflected in Dimension 3 (emergent behavior detection) and Dimension 4 (ethical vs operational autonomy) reinforces the qualitative view that organizations are struggling with unpredictable LLM behavior, cultural bias, and limited behavioral observability. Interviewees repeatedly noted that detecting goal drift or emergent multi-agent dynamics requires deeper visibility into system behavior than most institutions currently possess. The quantitative data support this: automated detection tools, simulations, and transparent ethical reasoning capabilities remain limited across all sectors. These gaps reflect a broader concern raised by AI experts that oversight must evolve beyond single-agent control toward distributed supervision that accounts for interactions, not just individual decisions.

The mixed readiness documented in Dimension 5 (recursive accountability) also aligns with qualitative concerns regarding unclear responsibility boundaries. Many organizations rely on informal escalation, limited audit trails, or high-level accountability, lacking the subsystem-level traceability required for distributed autonomous agents. Interviewees emphasised the need for digital audit twins, layered oversight structures, and recursive mapping of decision and responsibilities elements, largely absent in the quantitative findings. The argument that accountability must be anchored not only at the organizational level but across subsystems and intermediate decision layers is borne out by the uneven and often minimal adoption of recursive mapping mechanisms.

Finally, Dimension 6 (regionally adaptive governance) highlights that adaptation to India's infrastructural and linguistic diversity remains in

early development for most sectors. While BFSI and some parts of healthcare exhibit systematic region-specific safeguards, education and public services depend heavily on pilots or ad hoc adjustments. Interviewees described this as a major gap, stressing that safe deployment in India requires contextualised safeguards, offline-first capabilities, and culturally grounded models. The quantitative findings confirm that region-specific metrics, compliance routines, and deployment adaptations are far from institutionalised.

Overall, integrating the quantitative and qualitative evidence reveals a central mixed-methods insight: organizations recognise the importance of continuous monitoring, ethical reasoning, and accountable autonomy, but their operational readiness remains early to intermediate across most governance dimensions. Both data streams underscore that technical safety challenges cannot be resolved without parallel advances in governance structures, organizational capacity, sector-specific protocols, and behavior-centric oversight tools.

The findings from this study also confirm that agentic AI in India and the wider Global South is being deployed into environments that are structurally different from those assumed in most global AI safety literature, particularly in terms of infrastructure reliability, regulatory maturity, and dependence on foreign technology platforms. These structural differences mean that safety cannot be treated as a purely model-centric alignment problem; it must be understood as a socio-technical orchestration challenge that spans models, tools, data flows, institutions, and infrastructure layers. Future governance must therefore integrate lifecycle monitoring, distributed accountability, ethical autonomy, and region-sensitive protocols into a coherent architecture that reflects the realities of India's diverse operational environments. The ASF is therefore best interpreted as an attempt to translate frontier AI safety ideas into a form that is implementable in data-rich but infrastructure-constrained contexts like India, Brazil, and much of Africa.

10.1 ADVANCEMENT OVER PRIOR ART

Most existing work on AI safety and governance has been developed in and for the Global North, assuming strong data protection regimes, relatively stable compute and network infrastructure, and locally owned AI platforms. In contrast, countries in the Global South often rely on cloud infrastructure, models, and platforms controlled by foreign providers, raising concerns about digital colonialism, data extraction, and weakened state capacity to enforce norms over critical digital infrastructure. Within that context, this study advances prior art in three ways:

From model alignment to orchestration safety: Work by major labs on constitutional AI and red teaming focuses on shaping the behavior of a single model, but does not fully address failure modes that arise when multiple agents coordinate across tools, APIs, and institutions (e.g., cross-bank lending swarms or multi-agent fraud detection networks). By foregrounding orchestration patterns such as sequential, federated, hierarchical, and meta-governed, this report reframes safety as a property of the entire workflow and its surrounding oversight

mechanisms rather than only of the base model.

From abstract ethics to embedded regulatory compliance: Many global frameworks emphasise high-level principles such as fairness, accountability, and transparency, but offer limited guidance on how to keep agents in sync with fast-evolving local rules in data protection, finance, or public administration. The ASF introduces the concept of a Dynamic Compliance Oracle, which externalises legal and policy constraints into a machine-readable regulatory interface that agents must consult before executing high-risk actions, directly responding to challenges raised by India's DPDP rules and sectoral AI governance initiatives.

From generic risk lists to Global South–specific hazards: International analyses often highlight AI risks such as misinformation, bias, and job displacement, but underplay structural hazards such as unreliable power, limited local compute, and dependence on foreign models that may not respect local norms. This study's emphasis on graceful degradation, offline-capable agents, and vernacular golden vectors explicitly extends the risk taxonomy to capture conditions characteristic of the Global South.

10.2 IMPLICATIONS FOR POLICY, PRACTICE, AND RESEARCH

The integrated findings of this study, spanning six quantitative governance dimensions and rich qualitative evidence, yield a set of critical implications for India's emerging agentic AI ecosystem. These implications apply to policymakers, regulators, practitioners, industry leaders, researchers, and public institutions seeking to build trustworthy, contextually grounded, and future-resilient agentic AI systems. The general implications include:

1. India requires sector-specific governance pathways, not uniform regulation

The sharp divergence in governance maturity across sectors, exemplified by BFSI's comparatively structured practices, education's and public services' transitional readiness, and healthcare's polarised governance, demonstrates that one-size-fits-all governance models are insufficient.

Sector-specific roadmaps are needed, accounting for:

- Operational mandates
- Data and infrastructure maturity
- Risk exposure
- Institutional capacity
- Regulatory obligations

Regulators should articulate differentiated governance trajectories that reflect each sector's starting point rather than imposing uniform maturity expectations.

2. Ethical autonomy must be institutionalised as an operational requirement

Quantitative and qualitative results converge on a central point: India currently lacks the structures to

systematically govern ethical autonomy, agents' capacity to make or escalate decisions with moral, legal, or societal consequences. To address this:

- Ethical–operational distinctions must be formalised, not left to informal documentation.
- Transparent reasoning mechanisms must become standardised evaluation criteria, not optional features.
- Sectoral agent constitutions should encode red-line constraints aligned with Indian legal and cultural norms.

Without institutionalising ethical autonomy, India risks deploying autonomous systems whose internal moral logic is neither visible nor accountable.

---

3. Recursive accountability must become the backbone of agentic AI governance

Dimension 5 reveals widespread absence of subsystem-level traceability, inconsistent audit trails, and informal escalation pathways, findings echoed in interviews highlighting unclear responsibility boundaries and missing downward accountability.

To ensure accountability in multi-agent and distributed environments:

- Recursive mapping should trace responsibility through every agent and subsystem.
- Tamper-evident audit trails must be mandated for high-risk and critical contexts.
- Escalation pathways should be automated, pre-defined, and testable, replacing informal, case-by-case practices.

This shift reflects expert insights emphasising that future governance architectures must assume non-linear, multi-actor autonomy as the norm, not the exception.

---

4. Behavior-centric monitoring and emergent-risk safeguards must be prioritised

Quantitative data showed limited availability of automated detection, simulations, or anomaly alerts for emergent behaviors. Qualitative interviews underscored why: organizations lack behavioral monitoring pipelines and struggle to anticipate non-linear, collective dynamics.

Implications include:

- Investments in behavioral logging, drift analysis, and multi-agent simulation environments
- Development of red team infrastructure for testing emergent agent dynamics
- Integration of meta-governance layers that supervise agent collectives rather than isolated workflows

India's AI governance must evolve from static compliance to continuous behavioral oversight.

5. Regionally adaptive governance is essential for equitable and safe AI deployment

India's infrastructural heterogeneity, 5G-enabled metros, 3G/2G-dependent rural districts, and multilingual contexts make regionally adaptive governance indispensable. Yet Dimension 6 shows inconsistent adaptation patterns, especially in education and public services.

Critical implications include:

- Mandating offline-first and bandwidth-aware agent architectures
- Creating region-specific compliance baselines
- Ensuring safety protocols and audit processes are robust in low-infrastructure settings
- Establishing linguistically and culturally adaptive agentic behavior models

Without regional sensitivity, agentic AI risks amplifying the digital divide and reinforcing structural inequities.

5. Regionally adaptive governance is essential for equitable and safe AI deployment

India's infrastructural heterogeneity, 5G-enabled metros, 3G/2G-dependent rural districts, and multilingual contexts make regionally adaptive governance indispensable. Yet Dimension 6 shows inconsistent adaptation patterns, especially in education and public services.

Critical implications include:

- Mandating offline-first and bandwidth-aware agent architectures
- Creating region-specific compliance baselines
- Ensuring safety protocols and audit processes are robust in low-infrastructure settings
- Establishing linguistically and culturally adaptive agentic behavior models

Without regional sensitivity, agentic AI risks amplifying the digital divide and reinforcing structural inequities.

6. Capability-building must accompany governance reform

A recurring insight, quantitative and qualitative, is the confidence–capability mismatch: sectors with limited auditability or escalation structures often reported moderate to high confidence in oversight capabilities.

This calls for:

- National-scale training in agentic AI governance, behavior analysis, and incident response
- Cross-institutional benchmarking frameworks
- External audits to ensure organizational self-assessment aligns with actual capability maturity

Governance evolution must be accompanied by institutional learning and capacity development.

7. India must move toward a multi-layered, governance stack for agentic AI

Given the divergence across Dimensions 1–6, and the systemic nature of agentic AI risk, India requires a layered governance architecture that includes:

- Lifecycle governance (continuous monitoring, adaptive updates)
- Human–agent oversight structures (supervisory and meta-governance layers)
- Behavioral and emergent-risk monitoring
- Ethical autonomy frameworks
- Recursive accountability and auditability
- Regionally adaptive safety protocols

Such a governance stack transforms oversight from a collection of isolated controls into a coherent multi-layer ecosystem aligned with agentic AI's distributed, dynamic, and evolving nature.

Based on our findings and the general implications of this study, the implications specifically tailored for policy-makers, practitioners, and researchers are listed below:

- For policy-makers, the results underscore that static, document-based regulation is poorly suited to agentic systems that continuously interact with data, tools, and other agents at scale. Instead, regulators in finance, health, and digital governance will need to move toward live governance instruments, such as AI governance guidelines exposed via secure APIs, that encode obligations from laws like the DPDP Act, financial sector frameworks like FREE-AI, and sector-specific standards into artefacts that agents can query programmatically before acting. This approach echoes broader calls in Global South governance debates for multi-layered, adaptive AI regulation that can respond to rapidly changing technology and uneven institutional capacity.
- For practitioners, particularly in BFSI, healthcare, and large public systems, the findings suggest that AI safety should be treated not as a one-off compliance exercise but as an ongoing operational discipline akin to cybersecurity or site reliability engineering. This implies building new roles, such as AI reliability engineers and model risk officers, and institutionalising practices like continuous incident simulation, red teaming under low-bandwidth and multilingual conditions, and cross-institutional sharing of threat signals, similar to what emerging agentic threat-modelling frameworks (such as MAESTRO) are beginning to recommend.
- For researchers, especially in India and the wider Global South, the results highlight three priority areas: (1) methods for machine unlearning that allow DPDP-style erasure rights to be meaningfully implemented in deployed agentic systems; (2) robust evaluation suites for non-English, low-resource languages and dialects that can detect hallucinated fluency and culturally inappropriate outputs; and (3) formal models of multi-agent coordination under partial connectivity and adversarial influence, extending current work on AI orchestration and cyber-physical resilience.

India has a unique opportunity to shape the next generation of agentic AI governance, one that is ethically grounded, behaviorally informed, recursively accountable, and deeply attuned to the infrastructural and socio-cultural diversity of the Global South. The findings of this study provide a roadmap for building that future.

10.3 LIMITATION ANALYSIS

The empirical foundation of this report rests on a purposive expert sample of 110 respondents, with a sectoral distribution skewed toward education and public services and smaller subsamples in BFSI and healthcare, which reflects the current diffusion of AI experimentation but limits the strength of statistical generalisation in those heavily regulated domains. In addition, the survey relies on self-reported assessments of maturity and governance practices, which are vulnerable to optimism and social-desirability bias, especially in organizations that see AI adoption as a reputational signal of modernity and innovation. On the literature side, much of the formal work on agentic orchestration and threat modelling (for example, in enterprise automation and cybersecurity) still focuses on high-infrastructure contexts, meaning that some safety patterns imported from that work may require adaptation before they can be reliably transferred to low-resource, state-capacity-constrained settings.

Finally, the technological landscape itself is evolving quickly: the rise of production-grade agentic platforms, new orchestration standards, and industry-grade agent networks is accelerating, with vendors already marketing blueprints for enterprise-scale agent orchestration. As a result, some specific architectural assumptions in this report, such as reliance on LLM-centric agents or particular coordination protocols, may need revisiting as new classes of models (e.g., large action models, neuro-symbolic agents) and orchestration fabrics become mainstream in the next 2–3 years.

10.4 FUTURE RISKS AND STRATEGIC FORESIGHT

Looking forward, the intersection of agentic AI with structural constraints in the Global South suggests several emerging risk vectors that go beyond today's deployment concerns. One is the risk of data and infrastructure dependency, where states and critical sectors in Africa, Latin America, and South Asia rely on foreign-owned models, clouds, and orchestration platforms, effectively externalising not just compute but also governance levers to actors outside domestic democratic oversight, reinforcing patterns described as digital colonialism. Another is the possibility of agent-to-agent collusion or systemic convergence, where procurement, trading, or pricing agents trained on similar objectives and models begin to coordinate implicitly in ways that undermine market competition or amplify macro-level shocks, a concern already being raised in analyses of AI's role in global economic inequality.

At the same time, there is a strategic opportunity: if countries like India can successfully combine AI governance guidelines, data protection rules, and investment in domestic AI capacity, as recent policy documents argue is necessary, they may be able to steer agentic AI toward inclusive growth rather than widened divides. Doing so will require not only frameworks like ASF at the technical and organizational levels, but also regional and multilateral collaboration so that Global South states can shape standards, share governance tools, and avoid fragmented, duplicative responses to the same cross-border platforms and models

11. Actionable Recommendations

Based on our findings, we propose the following actions

A. Immediate (0–6 months) – Safety Reset

1. Industry: Zone-of-Autonomy Audit – Map all in-production agents into risk zones:

- **Zone 1 (Informational):** Agents only reading data (e.g. FAQs). Ensure logging of queries/answers for transparency.
- **Zone 2 (Transactional):** Agents that write to non-critical systems (e.g. ticketing, scheduling). Impose rate limits (e.g. ≤10 actions/min) and anomaly detection to prevent runaway loops.
- **Zone 3 (Critical):** Agents with legal/financial/medical decision power. Mandate human-in-the-loop for every state-changing action until audited. For example, any loan approval by an agent requires sign-off by a manager or a compliance agent.

2. Connectivity Circuit Breakers: Implement logic in agents: if network latency or error rate exceeds safe thresholds (e.g. >500ms or >5% packet loss), switch to fallback mode. This may involve:

- Declining non-urgent requests until connectivity restores.
- Serving cached data with warnings (offline mode).
- Alerting a human operator.

3. Rigorous logging and monitoring: Deploy the concept of Digital Audit Twins: immutable records (e.g. blockchain logs) of every agent decision, prompt, and tool call. Ensure this logging is tamper-evident. This enables post-hoc audits and is required under DPDP's accountability principle.

4. Designate AI safety leads: Each organization should appoint an AI Safety or Reliability Engineer responsible for implementing these measures, similar to a CISO for security.

B. Mid-Term (6–18 months) – Building Resilience

5. Regulatory APIs: Regulators (RBI, ED, NITI Aayog, etc.) should publish compliance requirements as machine-readable rules. For example, the RBI could expose a compliance endpoint that returns current lending guidelines in JSON. Agentic systems can then check against these rules before execution (e.g. ***Is this loan amount permissible under current LTV norms?***).

6. Ethics and explainability mandates: Require documentation of agent architectures and reasoning. For instance, BFSI may mandate that any automated decision can be explained (e.g. by showing the chain of agent sub-tasks). Similar to the EU's requirement for transparency in high-risk AI, implement explainable AI tools as part of deployment.

7. Golden Vector testing suite: Develop standardized test batteries in Indian and other local languages to evaluate agents for hallucinations and biases. Examples: translation sanity checks, culturally-relevant question-answer sets, vulnerability injection tests. Encourage open collaborations between academia and industry on this.

8. Capacity building: Launch training programs (e.g. via NITI or sectoral bodies) to educate regulators and practitioners on agentic AI risks. The Salesforce

data shows public trust hinges on human oversight and data security; we must build this understanding

C. Strategic (1–3 years) – Institutional Innovation

9. National agentic AI sandbox: A government-hosted platform where multi-agent systems can be tested in a controlled environment simulating real-world conditions (e.g. varying network strength, multilingual users). This serves as a safe haven for experimentation and failure.

10. International cooperation: Engage in multilateral AI forums (UN, BRICS, OECD-AI) to craft harmonized standards. The EU and others are shaping global norms; India and regional partners should co-design principles reflecting Global South realities (in line with UNESCO's call for inclusive data governance).

11. Audit and certification: Develop (or adapt) third-party audit frameworks for agentic AI safety. For example, a certification akin to ISO could evaluate an organization's adherence to ASF. This would cover technical (security, robustness) and organizational (governance, documentation) aspects.

12. Data sovereignty initiatives: Invest in local AI infrastructure (computing clusters, domestic LLMs) to reduce reliance on foreign platforms. Encourage open-source agentic projects with community oversight to democratize the technology.

To bridge the maturity gap identified in our findings and operationalize the ASF, stakeholders must move from principled intent"to engineered safety. The following roadmap decomposes high-level strategy into specific, executable technical and organizational tasks.

11.1 IMMEDIATE ACTIONS (0-6 MONTHS): THE SAFETY RESET

A. For Industry: technical & operational triage

Action 1: Conduct a zone of Autonomy audit

- **Directive:** Map every agent in production today. Categorize them into three strict risk zones:
- **Zone 1 (Informational):** Read-Only agents (e.g., HR FAQ, Document Search). Requirement: Basic logging.
- **Zone 2 (Transactional):** Agents with write access to non-financial databases (e.g., Scheduling, Ticket filing). Requirement: Rate Limiting (max 10 actions/minute) to prevent runaway loops.
- **Zone 3 (Critical):** Agents with financial/medical/legal authority. Requirement: Mandatory Human-in-the-Loop (HITL) approval for every state-changing action until a "Watcher Agent" is validated.

Action 2: Deploy connectivity circuit breakers" (Infrastructure Safety)

- **Context:** Mitigating the connectivity cliff in rural deployments.
- **Technical Spec:** Patch all edge/mobile agents with the following logic:

```python
if network_latency > 500ms OR packet_loss > 5%:
    agent.mode = "READ_ONLY"
    agent.cache_actions = True
    user_interface.display("Offline Mode: Actions will sync later")
```

else:
agent.mode = "FULL_DUPLEX"`

B. For Regulators (MeitY, RBI, SEBI)

Action 1: Publish agentic compliance checklists

- Move beyond high-level ethics. Release specific technical checklists for auditors:
- Example Check: ***Does the agent have a distinct, hard-coded 'Stop/Kill' switch accessible to the user at all times?***
- Example Check: ***Does the agent re-verify user consent before initiating a cross-border API call?***

C. For Academia

Action 1: The Golden Vector Project

- Launch a mission to create open-source Safety Evaluation Datasets for the top 10 Indian languages.
- **Focus:** These datasets should specifically test for hallucinated fluency (e.g., asking medical questions in Marathi where the correct answer relies on local context).

11.2 MID-TERM ACTIONS (6-18 MONTHS): THE SYSTEMIC SHIFT

A. For Industry: Architecture & Workforce

Action 1: Adopt guardian agent Architectures

Directive: Deprecate monolithic agents. Move to a Dual-Agent Pattern:

- Agent A (Worker): Focuses on task completion (High Creativity, Low Safety).
- Agent B (Guardian): A smaller, fine-tuned model that focuses only on safety constraints (Low Creativity, High Safety). It monitors Agent A's output and blocks violations.

Action 2: Formalize the AI Reliability Engineer (ARE) role

Context: Safety cannot be a side-job for Data Scientists.
Job Description: The ARE is responsible for:

- Designing red team attacks (prompt injection, jailbreaking).
- Managing the compliance oracle integrations.
- Monitoring drift dashboards for fairness and accuracy.

B. For Government (IndiaAI Mission)

Action 1: Launch the National Agentic AI Sandbox

- Concept: A regulatory safe harbor where startups can test high-risk agents (e.g., Autonomous Medical Triage) on real DPI data (anonymized)
.
- Condition: Participants must share their "Failure Logs" with the ecosystem to build a shared threat intelligence database.

Action 2: Mandate fairness audits for Public Procurement

- Directive: Any AI agent procured by the government (e.g., for beneficiary selection) must undergo a third-party Fairness Audit.

- Metric: The agent must demonstrate a Disparate Impact Ratio > 0.8 (i.e., the selection rate for a marginalized group must be at least 80% of the rate for the privileged group).

11.3 LONG-TERM ACTIONS (18+ MONTHS): THE FUTURE-PROOFING

A. For Government & Academia

Action 1: Sovereign Safety Models

- Develop indigenous constitutional AI models trained specifically on Indian legal and ethical texts (The Constitution, IPC, Supreme Court Judgments).

- Utility: These models will serve as the Supreme Court for other agents, adjudicating ethical conflicts (e.g., ***Should I reveal this private data to law enforcement?***) based on Indian law, not Western training data.

B. For Regulators: The Regulatory API Shift

Action 1: Transition to Code is Law

Concept: Regulators (RBI, SEBI) must transition from publishing PDF circulars to hosting live Compliance APIs.

Workflow:

1. RBI updates lending norms (e.g., LTV ratio changed to 75%).
2. Update is pushed to api.rbi.org.in/lending-norms.
3. All regulated Fintech Agents query this API before disbursing a loan.
4. Result: Instant, 100% compliance without manual code updates.

11.4 SECTOR-SPECIFIC CHECKLISTS

Healthcare Leaders

- **Audit:** Do your triage agents have a hard-coded escalation path to a human doctor if confidence < 90%?
- **Data:** Is patient PII (Personally Identifiable Information) masked before it is sent to the LLM inference layer?
- **Consent:** Does the agent explicitly state I am an AI at the start of every interaction (NMC requirement)?

BFSI Leaders

- **Circuit Breaker:** Is there a system-level kill switch if the agent's transaction volume spikes by >50% in 5 minutes?
- **Fairness:** Have you tested your credit scoring agent against a counterfactual dataset (e.g., changing only the gender/caste of the applicant) to prove non-discrimination?

Public Sector (e-Gov) Leaders

- **Inclusion:** Does the grievance bot support voice input for at least the top 3 local languages of the district?
- **Resilience:** Can the agent function in a store-and-forward mode during internet outages?

12. Conclusion and Strategic Outlook

This study provides a comprehensive, multidimensional assessment of India's readiness for agentic AI governance, integrating quantitative evidence across six governance dimensions with qualitative insights from industry, policy, and technical experts. Together, these findings reveal a governance landscape that is fragmented, uneven, and characterised by structural gaps that cut across sectors, while also highlighting emerging strengths and clear strategic opportunities.`

12.1 SYNTHESIS OF FINDINGS: THE ORCHESTRATION GAP

This study, anchored in the perspectives of 110 experts from India's education, healthcare, BFSI, and public sectors, uncovers a critical orchestration gap. While Indian organizations are aggressively

adopting Agentic AI, with 90% expected to deploy agents by 2026, the governance mechanisms to manage them remain dangerously static.

Our findings reveal a bi-modal safety landscape:

- The Fortified Islands: The BFSI sector, driven by the RBI's FREE-AI framework, has built robust, human-supervised agentic workflows. BFSI stands out as the most consistently advanced, supported by clearer mandates, risk cultures, and compliance expectations, findings strongly echoed in interviews, where participants noted that financial-sector governance culture forces you to think in layers and instills recursive accountability almost by default.
- The Open Plains: Education and Public Services, despite high adoption, lacks the recursive accountability structures needed to safely manage autonomous agents. A concerning 40% of education respondents lack mechanisms to detect emergent goal drift, posing a long-term risk to the cognitive development of students interacting with AI tutors. These sectors remain transitional, heavily reliant on manual oversight, limited documentation, and ad hoc practices. Interviewees from these sectors frequently described their environments as resource-constrained, procedurally fragmented, or not yet ready for autonomy.

Across sectors, both quantitative and qualitative findings converge on three deeper structural issues. First, ethical autonomy remains the least developed governance layer, with weak distinctions between ethical and operational decisions and limited transparent reasoning mechanisms. Interviewees repeatedly highlighted this as a conceptual blind spot: as one expert explained, we have technical autonomy but no moral architecture. Second, recursive accountability structures, critical for distributed agents, are substantially absent. The quantitative results show minimal subsystem-level traceability and inconsistent audit trails, and qualitative perspectives reinforced this gap, emphasising that organizations rarely map responsibility beyond the primary system boundary. Third, regionally adaptive governance, essential in the Indian context, remains inconsistent, with significant disparities between sectors. While some experts stressed the urgency of building regionally sensitive safety protocols, several interviewees noted that adaptation only happens when there is a crisis or regulatory pressure.

The core insight of this research is that infrastructure is a safety parameter. In the Global South, a safe agent is not just one that is aligned with human values, but one that is resilient to network volatility, linguistically diverse, and compliant with evolving data sovereignty laws like the DPDP Act.

12.2 STRATEGIC OUTLOOK: INDIA AND THE GLOBAL SOUTH (2026-2030)

As we approach the India AI Impact Summit 2026 (the first global AI summit hosted in the Global South), India is positioned to define a new paradigm of developmental AI safety.

1. The Shift to Sovereign Agentic AI

Dependency on foreign foundation models creates digital colonialism risks, where an agent managing Indian land records might enforce Western data norms. The strategic response, reflected in the IndiaAI Mission, is the rise of sovereign AI, indigenous small language models, and agents built on Indian data and compute. By 2026, we project that sovereign agents integrated with DPI will become the standard for government-to-citizen services, reducing reliance on global tech giants.

2. Trust as the New Currency

As agentic automation moves from pilot to production, trust will replace capability as the primary differentiator. Organizations that adopt frameworks like the ASF, implementing digital audit twins and guardian agents, will secure a trust premium, enabling them to deploy agents in high-stakes zones (like clinical triage) where competitors dare not tread.

3. The Global South as a Governance Lab

India's DPI + AI model offers a blueprint for the Global South (Africa, LATAM, SE Asia). Unlike the West's litigation-heavy approach or China's state-surveillance model, India offers a techno-legal model: embedding regulation directly into the API layer (e.g., DEPA for data sharing). This allows for scalable, low-cost safety enforcement, making it an attractive export to other developing nations facing similar state-capacity constraints.

12.3 ROADMAP TOWARD 2026 AND BEYOND

To realize this vision, the ecosystem must execute a synchronized transformation:

- Immediate term (2025): The safety reset. Organizations must conduct zone of autonomy audits to identify unmonitored agents. Regulators must publish compliance APIs to replace static guidelines.
- Medium term (2026): The agentic economy. As agent-to-agent commerce begins (e.g., a procurement agent negotiating with a supplier agent), we will need antitrust watcher agents to prevent algorithmic collusion. The focus will shift from individual agent safety to systemic market stability.
- Long Term (2030): The constitutional era. We envision a future where every critical agent carries a digital constitution, a hard-coded set of ethical values (derived from the Indian Constitution) that cannot be overridden by prompt engineering or reward hacking.

12.4 CLOSING PERSPECTIVE

The era of move fast and break things is over. In the age of agentic AI, where code executes real-world actions, transferring money, diagnosing disease, grading exams, the cost of breaking things is too high.

This report argues that safety is not a constraint on innovation; it is the enabler of scale. Just as brakes allow a car to drive faster, robust safety architectures like the ASF will allow India to deploy agentic AI at a velocity and scale that the world has never seen. By solving for the constraints of the Global South, infrastructure, language, and trust, India can build not just safer AI, but better AI for all humanity.

Agentic AI stands to transform critical services in India and the Global South, but only if its deployment is guided by robust safety and governance measures. This report has mapped the complex landscape of agentic AI risks, particularly as they interact with infrastructural inequalities and cultural contexts. We find that compliance and safety maturity currently trail adoption ambitions, leaving significant gaps in education and public sectors.

By synthesizing global frameworks (EU's AI Act, OECD Principles, National Institute of Standards and Technology Risk Management Framework (NIST RMF)) and local case studies, we outline an ASF that integrates continuous risk management with socio-technical awareness. Central to our vision is treating network reliability and ethical constraints as first-class safety parameters, agents must know when to pause or seek human help, not just how to act.

We emphasize actionable change: regulators should embed AI norms in live systems (APIs), industry should operationalize safety like cybersecurity, and governments should proactively fund the necessary AI ecosystem (data, evaluation, skill-building). The alternative – neglect – risks not only technical failures, but deepening inequities under a new form of digital colonialism.

In conclusion, steering agentic AI toward inclusive benefit requires collaborative governance and engineering. With the right mix of policy innovation and technical rigor, countries like India can leapfrog from being passive data suppliers to active shapers of the AI future. The recommendations herein offer a roadmap: if enacted, they will help ensure that agentic AI enhances societal well-being rather than undermining it.

REFERENCES

Karthikeyan, C., 2025. Smart Governance With AI for Transformative Urban Development in India: Harnessing Data-Driven Insights to Shape the Future of Indian Cities. In Nexus of AI, Climatology, and Urbanism for Smart Cities (pp. 217-246). IGI Global Scientific Publishing

.

"Building AI for India!", AI4Bharat, 2024. https://ai4bharat.iitm.ac.in/

Tallam, K., 2025. From autonomous agents to integrated systems, a new paradigm: Orchestrated distributed intelligence. arXiv preprint arXiv:2503.13754

Trombino, D., Pecorella, V., De Giulii, A. and Tresoldi, D., 2025. Knowledge Base-Aware Orchestration: A Dynamic, Privacy-Preserving Method for Multi-Agent Systems. arXiv preprint arXiv:2509.19599.

"AI agent orchestration patterns", Microsoft Ignite, 2025. https://learn.microsoft.com/en-us/azure/architecture/ai-ml/guide/ai-agent-design-patterns

"What Is AI Agent Orchestration? Examples & Benefits", DOMO, 2025. https://www.domo.com/glossary/ai-agent-orchestration

Nisa, U., Shirazi, M., Saip, M.A. and Pozi, M.S.M., 2025. Agentic AI: The age of reasoning—A review. Journal of Automation and Intelligence.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. and Chi, E.H., 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Sapkota, R., Roumeliotis, K.I. and Karkee, M., 2025. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges. arXiv preprint arXiv:2505.10468.

Bahangulu, J.K. and Owusu-Berko, L., 2025. Algorithmic bias, data ethics, and governance: Ensuring fairness, transparency and compliance in AI-powered business analytics applications. World J Adv Res Rev, 25(2), pp.1746-63.

Acharya, D.B., Kuppan, K. and Divya, B., 2025. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. IEEE Access.

Ranjan, R., Gupta, S. and Singh, S.N., 2025. Fairness in Agentic AI: A Unified Framework for Ethical and Equitable Multi-Agent System. arXiv preprint arXiv:2502.07254.

"Amazon scrapped 'sexist AI' tool", BBC, 2018. https://www.bbc.com/news/technology-45809919

"What Is Algorithmic Bias?, IBM, 2025. https://www.ibm.com/think/topics/algorithmic-bias#:~:text=Algorithmic%20bias%20occurs%20when%20systematic,socioeconomic%2C%20racial%20and%20gender%20biases

"Addressing Transparency & Explainability When Using AI Under Global Standards," Mayer Brown, 2025.https://www.mayerbrown.com/-/media/files/perspectives-events/publications/2024/01/addressing-transparency-and-explainability-when-using-ai-under-global-standards.pdf%3Frev=8f001eca513240968f1aea81b4516757

Pawar, A., 2025. Ethical and Governance Challenges of Agentic AI. International Journal of Humanities and Information Technology, 7(03), pp.76-82.

"Agentic AI - Threats and Mitigations. OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative", AI Governance Library, 2025. https://www.aigl.blog/content/files/2025/04/Agentic-AI---Threats-and-Mitigations.pdf

Tran, K.T., Dao, D., Nguyen, M.D., Pham, Q.V., O'Sullivan, B. and Nguyen, H.D., 2025. Multi-agent collaboration mechanisms: A survey of llms. arXiv preprint arXiv:2501.06322.

Mitra, C., 2025. Synchronization Dynamics of Heterogeneous, Collaborative Multi-Agent AI Systems. arXiv preprint arXiv:2508.12314.

"Agent Orchestration 101: Making Multiple AI Agents Work as One", lyzr, 2025. https://www.lyzr.ai/blog/agent-orchestration/

Evani, P.K., Agentic AI Security: A Control Framework for Autonomous Decision-Making Systems. Available at SSRN 5332681.

Joshi, S., 2025. Advancing US Competitiveness Through Governance Tools and Trustworthy Frameworks for Autonomous GenAI Agentic Systems.

Joshi, D., 2024. AI governance in India – law, policy and political economy. Communication Research and Practice, 10(3), 328–339. https://doi.org/10.1080/22041451.2024.2346428

NITI Aayog, National strategy for Artificial Intelligence. Government of India

Salvi, P., 2025. Analytical Study of Artificial Intelligence Integration in Indian Banks and Its Implications for the Future of Banking and Financial Services in India. Available at SSRN 5705943.

"FREE-AI Committee Report - Framework for Responsible and Ethical Enablement of Artificial Intelligence", Reserve Bank of India, 2025. https://rbidocs.rbi.org.in/rdocs/PublicationReport/Pdfs/FREEAIR130820250A24FF2D4578453F824C72ED9F5D5851.PDF

Sahoo, S. and Behera, K., 2025, Artificial Intelligence in Education: Opportunities and Challenges for enhancing teacher competence in Indian classrooms.

Hajam, K.B. and Purohit, R., 2024. The Ethics of Artificial Intelligence: A Critical Examination of Moral Responsibility and Autonomy. Indian Journal of Educational Technology, 6(II), pp.398-407.

"India Takes the Lead Establishing the IndiaAI Safety Institute for Responsible AI Innovation", IndiaAI Mission, 2025.https://indiaai.gov.in/ article/ india-takes-the- lead-establishing-the-indiaai-safety-institute-for-responsible-ai-innovation

"AI IN SCHOOL EDUCATION: TOWARDS A PREPAREDNESS FRAMEWORK", ICRIER Prosus Centre for Internet and Digital Economy, 2025. https:// icrier.org/pdf/ AI_in_School_Education.pdf

"National AI Portal (INDIAai)", National e-Governance Division (NeGD), 2025, https://negd.gov.in/national-ai-portal-indiaai/

ABBREVIATIONS

ASF – Agentic Safety Framework
AI – Artificial Intelligence
AICTE – All India Council for Technical Education
API – Application Programming Interface
ASMI – Agentic Safety Maturity Index
AutoGPT – Autonomous Generative Pre-trained Transformer
BFSI – Banking, Financial Services, and Insurance
BIS – Bureau of Indian Standards
BoM – Bill of Materials
CAG – Comptroller and Auditor General
CBSE – Central Board of Secondary Education
CoT – Chain-of-Thought
DIKSHA – Digital Infrastructure for Knowledge Sharing
DEPA – Data Empowerment and Protection Architecture
DPI – Digital Public Infrastructure
DPDP – Digital Personal Data Protection
EHR – Electronic Health Record
EU – European Union
FREE-AI – Framework for Responsible and Ethical Enablement of Artificial Intelligence
GDPR – General Data Protection Regulation
HELM – Holistic Evaluation of Language Models
HIPAA – Health Insurance Portability and Accountability Act
ICRIER – Indian Council for Research on International Economic Relations
ICU – Intensive Care Unit
IEC – International Electrotechnical Commission
ISO – International Organization for Standardization
IT – Information Technology
KYC – Know Your Customer
LLMs – Large Language Models
MeitY – Ministry of Electronics and Information technology
ML – Machine Learning
MMLU – Massive Multitask Language Understanding
NABH – National Accreditation Board for Hospitals and Healthcare providers
NASSCOM – National Association of Software and Service Companies
NBFC – Non-Banking Financial Company
NEP – National Education Policy
NITI Aayog – National Institution for Transforming India Aayog
NMC – National Medical Commission
NSAI – National Strategy on Artificial Intelligence

OECD – Organization for Economic Co-operation and Development
ONDC – Open Network for Digital Commerce
OWASP – Open Worldwide Application Security Project
PDF – Portable Document Format
PDPB – Personal Data Protection Bill
PMLA – Prevention of Money Laundering Act
RBI – Reserve Bank of India
RLHF – Reinforcement Learning from Human Feedback
RMP – Registered Medical Practitioner
SEBI – Securities and Exchange Board of India
SLMs – Small Language Models
STOS – Socio-Technical Orchestration Safety
SWAYAM – Study Webs of Active-Learning for Young Aspiring Minds
UGC – University Grants Commission
UPI – Unified Payments Interface
US – United States

SAIGS 2025

Safe AI for the Global South

Empowering the Global South with safe and responsible AI ecosystems

aisafety@am.amrita.edu

www.amrita.edu/events/ai-safety-conclave

Amritapuri Campus, Kollam, Kerala

PARTNERS

RAISE LABS
Research in AI, Information Security, Sustainability and Education

Amrita Center for Cybersecurity Systems & Networks

AMRITA VISHWA VIDYAPEETHAM | CENTER FOR POLICY RESEARCH

nasscom ai
Research Questionnaire Partner