

[← back to home](#)

Strategic Priorities for India

Beyond definitions: How India's AISI can espouse a sociotechnical approach to AI safety.

 Jhalak Kakkar & Joanne D'Cunha

Abstract

Colonial legacies that have shaped the progress of other digital technologies continue to influence the global AI ecosystem, with Global Majority countries facing risks from the import of AI systems developed without sufficient attention to their lived experiences, priorities, and safety needs. This essay examines the emergence of AI Safety Institutes (AISIs) within the evolving global AI governance landscape and argues for an evidence-based, sociotechnical conception of AI safety. It contends that countries in the Global Majority must be collaborators and designers in shaping how AI is safely and ethically integrated into their societies. Against this backdrop, the essay offers insights into the structure, functions, and role of India's AISI, arguing that it can develop AI safety approaches grounded in the social, economic, and cultural realities of India and other Global Majority contexts.

Introduction

Over the last several years, we have seen a rapid evolution of the global AI ecosystem with significant breakthroughs including the explosion of the LLM domain, multimodal

integration and the advent of agentic AI. AI can be leveraged to unlock opportunities for inclusion and access as well as economic development, and the push for AI adoption is rapidly intensifying in the Global Majority.

However, if not governed effectively and developed and deployed safely, it can lead to significant harm to individuals, communities, and society. Countries in the Global Majority and their user populations are most vulnerable to such harms. Well known is the dearth of indigenous datasets, perpetuation of existing social bias through algorithmic bias, and inadequate digital infrastructure, that exacerbates these emerging AI risks. For these reasons amongst many other emerging ones, there is an urgent need for the development of robust domestic institutions such as AI Safety Institutes (AISIs) to develop effective safety practices, contextualised to local realities, and oversee the implementation of mechanisms for risk mitigation.

The rapid adoption of AI in the Global Majority is happening in the backdrop of a significant geopolitical race between the United States and China as they seek to dominate the AI ecosystem and capture markets around the world. Within this context, the entrenched effects of industrialisation and colonialism are well-known; it has built historical power structures resulting in deep economic dependencies through dominance, extraction and captive markets. The new technological ecosystem imitates these effects in various ways through economic and political power asymmetries, control over critical infrastructure, and technological dependencies.

In the domain of AI, scholars continue to emphasise the colonial legacy that currently shapes AI production, its supply chains, knowledge epistemologies, and market dominance. The global “AI arms race” has clearly demonstrated competitive interests in capturing markets through technological hegemony, geopolitical influence, and military prowess. The United States, for example, has established this interest transparently in many ways. A key showcasing of this desire for technological dominance has been by championing the export of a “full-stack American AI” to the world. In fact, the United States has titled their AI Action Plan as “Winning the AI Race”. In the past, we have seen the harmful effects of market capture and colonisation in the Global Majority – individual level harms, collective harms to marginalised groups, and erosion of governance institutions. Most recently with the proliferation of internet/social media platforms, Global Majority countries saw significant harms at individual, community, and societal levels. It is crucial to reflect on learnings from this recent wave of tech colonisation and rethink approaches as we now deal with the impact of AI.

To ensure resilience and equitable development of the AI ecosystem, it is imperative that Global Majority countries are intentional in designing and operationalising institutional architecture, policies, legislation, and systems. These structures and mechanisms need to ensure that Global Majority countries not only occupy a significant place in the global value chain of AI but also safeguard populations from harm – individual harm, collective level group harm as well as negative impacts to their democratic institutions and societies.

One approach to achieving the above is moving up the AI value chain. This can be in the form of industrial policy that encourages the development of local AI ecosystems. As part of this, investments by government and other stakeholders are critical to ensure inputs into AI – such as semiconductors, datasets and support to startups and other companies – to undertake the development of AI applications. Strategic policies can help identify the role the country can play in the global AI value chains and ensure there are points of involvement for Global Majority countries at the “high” points of the value chain such as design and development of semiconductors rather than only the “low points” of the value chain such as data labellers, extraction of minerals and resources, and the setting up of data centres that have detrimental environmental and labour impacts.

On the other hand, is harm prevention. Safeguarding people and institutions from harm can be addressed through a combination of the work of regulators, AI Safety Institutes, the judicial system, amongst others. The critical role these institutions occupy can ensure the safe and trusted development and deployment of AI systems, and development and effective implementation of new and existing laws and policies.

AI Safety Institutes are gaining momentum as a pathway to guide the safe and trusted deployment of AI. To examine the value of an AISI in the Global Majority, the essay first explores the shifts in the global AI governance landscape that has informed the contours of AI safety and the emergence of AISIs. It then argues for the adoption of an evidence-based sociotechnical conception of AI safety by Global Majority AISIs including India. With this lens, the essay offers insights on the structure, functions, and role that India's AISI can play in charting AI safety approaches for the Indian context, that are grounded in the broader social, economic, and cultural realities of countries in the Global Majority.

How should the Global Majority and its AISIs think of AI safety?

The objectives of this essay are situated in the second avenue through which Global Majority countries can build a resilient and equitable AI ecosystem - by contextualising

risk and safeguarding against harm. There are numerous avenues through which to address the risks and harms posed by this technology. Key amongst these has been regulation; however, with AI, there has been a widespread reluctance by governments across the world to regulate. Some justification of this response can be attributed to the complexities of AI systems and challenges with regulation meeting the rapid speeds of development and adequately covering the multifaceted nature of the technology. However, framing of principle-based regulation with the creation of independent institutions and regulators empowered to frame rules, regulations, and guidance can go some way to address these challenges. Parallely though is a significant push towards deregulation that stems from an overwhelmingly false dichotomy between regulation and innovation, often advanced by industry. Over the last few years, this dichotomy has grown in emphasis due to the sudden advancement of foundational AI models such as OpenAI's GPT series, Google's Gemini, Anthropic's Claude and Meta's Llama and consequently, their novel and emerging risks. Notably in this context, the current case against regulation has stemmed partly as an overcorrection. One, the initial over-emphasis on future existential AI risks within global AI safety discourse became at odds with "AI opportunity" and innovation and led to a few countries drastically shifting away from safety, and two, a perception that regulation is at odds with an "AI arms race" that demands an unencumbered pace of domestic AI development and its deployment across global markets.

In trying to address broader future AI risks such as existential risks, governments across the United States and the European Union responded by ramping up efforts to address safety risks of advancing AI models. Some have been through voluntary commitments by industry, updating some regulations to cover foundational models, building global consensus through the ongoing series of AI Summits and their country-level declarations and the United Nations AI Scientific Panel and Global Dialogue on AI. There was a great deal of initial momentum on AI safety efforts, particularly at the international level, which also led to the establishing of institutional architecture for AI safety through AI Safety Institutes across a number of countries. However, those keeping track have seen discourse and shifts away from safety, relegating both safety agendas and regulation to the background.

These shifts are increasingly critical to acknowledge and examine; they demonstrate impact on the way safety is discussed at high-level dialogues, how institutions such as AISIs interpret and determine their safety priorities, and importantly, how countries in the Global Majority countries engage with conceptions of AI safety in their contexts.

The latter two are particularly relevant for Global Majority countries shaping their AI ecosystems amidst unique contextual risks. In examining the relevance of these safety related developments and the forming of AI safety institutions, two core aspects on AI safety are worth noting for the Global Majority:

1. Recognising the source of contemporary AI safety conceptions

Simply put, a theoretical examination of the field of AI safety is that it seeks to prevent harms or risks from AI. By this boundary, all research towards this effort would be categorised within the domain of AI safety. However, in practice, contemporary approaches to AI safety have been informed (and funded) by communities importing values, perspectives, and objectives from other philosophical or ethical priorities. Note for instance that AI safety has historically been associated with a focus on existential risk – concerns that heavily borrow from movements of transhumanism, effective altruism, and longtermism. Additionally, a great deal of AI safety research and discourse naturally emanates from technical disciplines, with approaches tied to narrow forms of methodology and specific objectives such as alignment, robustness, or reliability. Expectedly, tech billionaires and Silicon Valley, situated in the Global North, that are developing these technologies play a dominant role in globally furthering such narrowly conceived AI safety objectives and technical approaches.

2. Recognising the context in which existing conceptions are defined

In examining the progression of AI safety discourse - both through the technical discipline and global governance agendas, we observe that existing approaches to AI safety have been wholly articulated by stakeholders in the Global North. These conceptions naturally reflect priorities, resources, and reliance on well-established Global North institutions. However, the current global prioritisation of existential risks and technical disciplines within AI safety does not account for contexts at vastly different stages of economic and societal development – which in the Global Majority, necessitates urgent focus on immediate existing harms and future needs in this regard before prioritising existential AI frontier harms.

Known centralisation of power and decision making within the Global North emphasises the need for Global Majority's proactiveness in determining its own contours of AI safety. These observations do not indicate a lack of parallel focus on non-existential AI risk and harm in the Global North. It simply highlights that the Global North is in a position and capacity to be able to achieve both objectives.

Moving beyond purely catastrophic and technical interpretations by adopting a sociotechnical approach to AI safety

The limitations of current conceptions of AI safety necessitate rescoping. In examining this, scholars such as Harding and Kirk-Giannini contrast the articulated objectives of the field of AI safety and the manner in which the field operates, in reality. Their work highlights that while the field of AI safety focuses on harm prevention or reduction, it often does not include or directly consider within the domain, any research examining algorithmic bias, misinformation, breaches of privacy, albeit research with the same goal of mitigating harm. These demarcations, they explain, occur due to the tendency to do the following: first, “explicitly or implicitly” designate the domain’s focus to catastrophic risks from future AI systems; and second, limit the domain based on methodological constraints, restricting it to technical AI safety research.

Limiting the focus of AI safety to catastrophic harms or methodologies of technical disciplines, fails to acknowledge the underlying reality that AI systems operate within societal structures. A “sociotechnical approach” to AI safety acknowledges that harm is not “located” only in AI. Instead, that AI is not a neutral system and any associated harms result from its interactions as a “sociotechnical assemblage” embedded within a social, political, and economical environment. Such an understanding of safety broadens its focus to account for a wider spectrum of harms and to address them, expanded interdisciplinary approaches are needed beyond just the technical.

Importantly, scholars exploring an African perspective of AI safety also adopt a sociotechnical lens and encompass both short-term and long-term risks within its fold. As a result, it may be useful to then consider characterising AI safety based on its objective – to prevent or reduce harms from AI systems, and not limited to specific types of harm, disciplines, or timeframes. This holistic reframing of the AI safety paradigm, through a sociotechnical lens, will enable a better encompassing of real-world systemic risks and harms, and additionally, a conception of AI safety that is more suited to the current contexts and needs of the Global Majority.

Higher stakes with incompatible definitions demand a sociotechnical approach to AI safety in the Global Majority

In investigating the need for re-examination of AI safety through a sociotechnical approach, it's crucial to reiterate that AI systems that are not safe, equitable, and resilient will have disproportionate impact on the Global Majority.

While risks and harms posed by AI systems are globally occurring, there is a distinct layer to the harms experienced by countries in the Global Majority. This is broadly articulated in four ways. *First* and as indicated above, the Global Majority operates within historically continuing hierarchies, structural inequalities and relational power asymmetries, and AI ecosystems reproduce related harms like extraction, inequity, Eurocentric knowledge systems. For instance, an extractive practice underpinning key AI technologies is low-wage labour situated in the Global Majority. *Second*, the import of AI designed from and for the Global North. These systems are often not designed or developed for the diverse contextual and lived experiences of the Global Majority; consider inherently biased datasets, ill-equipped models, under-evaluated deployment contexts, amongst others. Consequently, all of this fails to adequately account for intersectional and marginalised identities, varied linguistics, or socioeconomic and political conditions of the context of deployment in the Global Majority. *Third*, the underdeveloped or in-progress regulatory landscape and research ecosystems in some Global Majority countries are limited by capacity to identify harms and present peer reviewed research and evidence of such harms, and hence, overall, hinders strong mitigating measures. *Fourth*, most global technology companies operate out of multiple distinct contexts and are required to prioritise critical efforts. For instance, learning from practices of social media technology companies demonstrates a significant investment towards safety concerns within the United States or Europe, as compared to countries in the Global Majority. A consequence of this is financial constraints or limited incentives to invest in local contexts of non-priority markets, and as a result, an underrepresentation of their needs.

An additional point of note, Global Majority efforts in developing AI are often with Global North technology companies, in a bid to keep pace with global developments and the rapidly advancing geo-political AI race. This means that the Global Majority is operating in the global AI ecosystem with significant dependencies on the Global North, skewed incentives, and are consequently positioned to accept detrimental trade-offs impacting the safety and well-being of users.

Emphasising an evidence-based sociotechnical approach to AI safety for India's AISI

A recent report by IBM purports that AI could “unlock” more than USD 500 billion in economic value by 2030 in India, if supported by the necessary infrastructure, capacity, and national level alignment. During India's hosting of the AI Impact Summit in February 2026, Reliance Industries committed USD 110 billion as part of the Indian government and other Indian firms aim to invest over USD 200 billion in AI infrastructure. In late 2025, an SAP study on AI adoption in the Indian market projects that organisational investments

in AI could rise at about 33 per cent over the next two years. To better understand a portion of realities on the ground, the India AI Impact Summit released compendiums of real-world AI development and deployment. A compendium on health indicates use cases focused on AI enabled brain CT decision-support for radiologists and deep-learning image analysis to determine adequacy of sputum. Similarly, another compendium on education highlights use cases such as AI NESTHAM – an AI literacy platform that functions as a gamified curriculum, and Amira Learning – an AI powered oral reading tutor. These are a few instances that showcase the varied push for AI development and implementation underway and, consequently, the imminent expectations from AI within the Indian context.

Albeit a single-dimensional coverage of the Indian AI ecosystem, the above push demonstrates the existing vision for AI and simultaneously, the dire need for ensuring safeguards of this emerging technology to protect against a range of societal harms. However, the fragmented and on-going experimentation, along with an overall lack of information with regard to on ground realities of AI deployment, poses a challenge in decisively determining the approach to these safeguards. As a result, this impasse calls for a greater evidence base that can support AI safety through regulation, policy, standard-making, and technical safeguards. The International Safety Report highlights how key decisions on AI policy are made without adequate evidence. It emphasises that acting without a base of evidence that informs policymaking can result in ineffective approaches while inaction can cause societal vulnerabilities. Scholars also showcase the role that policy can play in fostering an evidence base for policymaking.

Hence, it is critical that safety standards and benchmarks are framed, relevant to the contextual realities, to enable the qualitative and quantitative measurement, assessment, and evaluation of AI systems that are being developed and deployed and an ecosystem for such evaluation and assessments is fostered. This will generate this much needed evidence base. As discussed above, it is critical that these assessments and evaluations are not only from a technical lens but a sociotechnical lens. It is within these objectives of an evidence-based sociotechnical approach, that the role of AISIs finds critical value and is explored further in the essay.

Zooming into India's vision of an AI Safety Institute

The previous section provides rationale for the reframing of AI safety, with a specific focus on its benefit to the Global Majority. A sociotechnical approach not only accounts for cultural contexts, socioeconomic realities, and intersectional concerns but also enables

addressing developmental and deployment harms in a manner that situates the rights and needs of individuals. The rest of this essay leverages the sociotechnical approach to expand the role and responsibilities of India's AISI, beyond a primary focus on the technical.

Risks and harms arising from the development and deployment of AI inevitably require to be addressed by companies and governments. However, this is often in direct tension with their push for innovation and market expansion. There is therefore a need for robust domestic governance institutions such as AISIs that can ensure contextualised safety is integrated into domestic AI design, development, deployment. This further enables the seeding and creation of an ecosystem for monitoring and evaluating AI systems in a manner that is adapted to local contexts. Additionally, the pursuit of these functions by such institutions can facilitate the creation of a robust community of experts across the social, legal, policy, and technical realm that can come together to combine expertise to shape and execute these objectives.

Relevance of the AISI in a pro innovation-oriented policy environment: India is at a unique moment in time with it being one of the first Global Majority countries in the process of establishing its own AI Safety Institute. It announced the Institute in January 2025 as it prepared to co-chair the Paris Action Summit in February 2025. The AISI is set up within the IndiaAI Mission's broader objectives of *"designing and developing adequate guardrails to advance the responsible development, deployment, and adoption of AI"* under its Safe and Trusted AI Pillar. India's AISI aims to play a critical role in fostering AI safety within the Indian context.

However, it is noteworthy that during the India AI Impact Summit 2026, India signed the Pax Silica Declaration with the United States, in a bid to secure India's access to AI supply chains. Alongside this Declaration, India signed a bilateral addendum focused on pro innovation approaches including pledging the adoption of a "pro-growth regulatory regime" that advances "innovation and investment" with the aim to empower both developers and deployers such as platforms to "rapidly scale" secure and trusted AI ecosystems". Even prior to signing this agreement, the Indian government has iterated its commitment to enabling a pro-innovation environment for the development and deployment of AI and has articulated an approach of *"responsible innovation over cautionary restraint"*. At minor odds with this evidently heavy position is also a muted acknowledgement that the potential of AI has to be harnessed *"while addressing the risks it may pose to individuals and society"*.

Given the strong pro-innovation approach to AI, an interesting question is how can the Indian AISI operationalise AI safety mechanisms to address AI risks, that in the short term will likely impede the speed of innovation? Perhaps the clue is in Secretary of MeitY, S. Krishnan's Foreword to the AI Governance Guidelines. Krishnan's remarks acknowledge that without robust trust and safety measures such as the Indian AISI, the IndiaAI Mission's efforts in infrastructure development, capacity building, and AI application development and adoption might be impeded due to societal and systemic risks. This highlights a critical point that underlines AI innovation. The adoption of AI and growth of the user base and market in India is deeply dependent on public trust in the AI that is being developed and deployed. The Indian AISI is optimally placed to engender this trust by shaping AI safety, evaluating and addressing the risks that arise in AI systems.

Structure and model: Coming at the heels of the 2025 Paris AI Action Summit, the IndiaAI Mission housed within the Ministry of Electronics and Information Technology began the process of operationalising elements of the India AI Safety Institute. It published a call for partnerships inviting applications from institutions such as academic institutions, R&D organisations, startups, private sector entities and government departments, to build and expand the Institute's capacity. Compared to AISIs across other jurisdictions, India's is envisioned through a unique "hub-and-spoke" model decentralising the capacity of the AISI – a central hub to coordinate the network of partner institutions and a model that allows the AISI to draw on the expertise of the broader Indian ecosystem. Each Institute Partner is expected to host a dedicated AISI cell along with a technical resource to be co-located with the AISI at IndiaAI, MeitY. Interestingly, a point that emphasises the AISI's heavy technical orientation is that the dedicated AISI cell's staff of an Institute Partner is expected to primarily have technical AI experience.

Functions: India's AISI is envisaged to drive critical AI research and development and to mitigate AI risks. In the articulation of its functions, it is mandated to contribute to technological advancement through technical tools, benchmarks, measurement systems, and also further governance mechanisms such as frameworks, protocols and best practices. Consistent with India's overall AI developmental strategy including through the IndiaAI Mission, the AISI is expected to foster these responsibilities through indigenous research and India-specific contextualised approaches. India's AI Governance Guidelines released in November 2025 by a drafting committee constituted by the Ministry of Electronics and Information Technology, with the guidelines focused on fostering AI innovation that is safe and inclusive, recommends that the AISI: (i) drive research to develop India specific risk frameworks and common standards, (ii) drive adoption of

voluntary commitments, (iii) carry out risk assessments and testing and evaluating of AI systems, operationalise AI incidents systems and pilot regulatory sandboxes.

Additionally, the AISI is also expected to focus on capacity building including training and awareness programs for key stakeholders. Relatedly, the stakeholder engagement by AISI is envisioned to be multistakeholder, engaging and collaborating with all relevant stakeholders including academia, civil society, startups, industry, and government.

Indian AISI's role in operationalising a sociotechnical approach to AI Safety: Recommendations on composition, structure and functions

As India charts out its AISI, it can pave the way for the reshaping of a more inclusive understanding of AI safety with an emphasis on a sociotechnical approach. The AISI has to enable public trust, transparency, and accountability and safe AI in practice. As discussed earlier, AI safety cannot be reduced to technical evaluations and testing of AI systems, and needs to encompass a broader sweep of sociotechnical assessments. Additionally, there is a need for translating safe AI into practice, bridging the gap between principles and practice, through standards and frameworks that are not purely technical but account for the complexity of deploying these systems in varied contexts and environments. Enhancing the capacity of regulators, government institutions, companies, researchers and other stakeholders is necessary to collaboratively develop and implement principles that achieve the above objectives and that can also be shared internationally towards global cooperation. In this section, we make recommendations around the composition, structure, and functions of the AISI with a focus on short-term recommendations to be operationalised in the next six months and longer-term recommendations to be operationalised over the course of a two-year period.

1. Composition:

A critical function of the AISI is to enable safe and publicly focused innovation that responds to community needs. To do so, necessitates that the user and common person is kept in focus during the development and deployment of AI systems and especially in addressing any resulting harms and real-world impact. The involvement of stakeholders that are equipped and can dispense an accurate representation of on-ground realities becomes critical as is their role in generating an evidence base for policy makers.

Short-term Recommendation: Draw on multistakeholder and multidisciplinary experts

It is imperative that the experts being engaged with are not only multistakeholder but multidisciplinary. Currently, as pointed to earlier, the expertise and orientation of the Institute is anchored in the technical domain. Hence, there is a danger that the AISI will default to a focus on the technical aspects of AI safety, at a moment when it is pivotal that the work of the AISI engages with the social, economic and cultural aspects of AI and its impact on society. The AISI will need to involve not only technical researchers, industry and startups but also academia from beyond engineering, AI computer science, and other technical programs to broader programs such as law, the social sciences and ethics, as well as civil society and think tanks that undertake on ground research and studies on user impacts. Involvement of a cross section of experts will enable us to bridge some of the democratic deficit currently arising in AI governance. To this end, the Spokes/AI Safety Cells of the AISI should have multistakeholder and multidisciplinary expertise.

There needs to be space for experimentation with mechanisms that meaningfully integrate non-technical approaches and perspectives. This approach is required across the spectrum of the AISI's functions – evaluations of AI systems, development of frameworks, and operationalising of safety mechanisms such as issue based working groups, incident reporting and response, sandboxes, etc. Any initiative undertaken by the AISI must have a balance of representation from disciplines and stakeholder groups.

Long-term Recommendation: Engage with grassroots perspectives and stakeholders

India's AISI also has an opportunity to pioneer innovative participatory approaches relevant to the Global Majority. These will be vital in determining ground realities that reflect the impacts of AI deployment and use, on users from vulnerable communities, rural areas, labour groups and representative groups from other domains. The AISI will have to develop pathways and modalities to effectively engage with and draw in the perspectives of these stakeholder groups. Mechanisms could include their membership of Working Groups, Spokes/AI Safety Cells of the AISI, development of mechanisms such as AI Impact Assessments for the deployment of new systems drawing on the mechanism of Environmental Impact Assessments and the involvement of these groups in these assessments, and providing community input during evaluations being conducted as part of AI Sandboxes.

2. Structure:

Short-term Recommendation: Constitute multistakeholder steering committee and thematic based working groups to provide research evidence and guidance to

Multistakeholder Steering Committee: There needs to be a coordinated approach by the AISI to identify the most pressing areas of research and themes to be taken up to prevent the duplication of work across Spokes and Working Groups of the AISI; oversee and coordinate the setting up and composition of the Working Groups; and set up a mechanism of public knowledge management of various outputs such as research, guidance, assessments and transparency disclosures by companies. Towards this end, the composition of the Steering Committee must imitate the multistakeholder and multidisciplinary nature of the AISI.

Thematic Working Groups (WGs): The proposed hub and spoke model of the AISI and also the colocation of experts from the Spokes within the AISI goes some way towards fostering an AI community. This model creates an opportunity for collaboration, cross learning and exchange of perspectives and ideas of the AISI with a cross section of experts.

The proposed hub and spoke approach allows the AISI to draw on the expertise of the Indian AI ecosystem. To operationalise a coordinated approach to developing research and evidence, there is a need to constitute thematic based WGs that can look at issues from an ecosystem wide lens. Structuring these efforts ensures that work is not disconnected across different stakeholder groups. For instance, these WGs can focus on particular use cases in a sector, an emerging challenge in the domain or examine a harm that has occurred. WGs will also facilitate the development of expert consensus and evidence on the state of play in a particular domain of AI in India, evaluate the risks and opportunities, and can provide context specific guidance and mitigations required. This will be valuable to inform policy makers and enable policy level alignment. Thematic issues to explore could include AI and data governance, child safety, online safety, diagnostic health applications, edtech, labour and competition as well as other sectoral focused working groups. The WGs can develop codes, standards, guidelines, evaluation metrics, testing frameworks and SOPs in the relevant domain.

To leverage a cross section of expertise and thinking, the WGs should be composed of 15-20 experts. The WGs need to maintain an equitable balance between technical experts and experts from the domain of social science, law, policy, etc. from across stakeholder groups such as industry, government, CSOs, and academia. Adopting this approach acknowledges that no single group of experts across the domains of technology, social science, law, or policy should unilaterally decide what risks count, who is harmed and what

harms matter, and what values should align with safe AI. Particular attention has to be paid to ensure that engagement with company representatives is structured in a way to maintain independence of recommendations and prevent capture by industry interests.

Long-term Recommendation: Set up an Evaluation Cell to foster a tiered independent evaluation mechanism and build evaluation infrastructure

Sociotechnical safety assessments and evaluations of AI systems will be a critical function for the AISI to undertake. For this, it is important for the AISI to foster the evaluation ecosystem and develop the infrastructure needed to undertake them. Towards this end, it would be important to set up an Evaluation Cell within the AISI that can operationalise this tiered mechanism over a period of time.

Enabling Self Assessment: The Evaluation Cell in conjunction with Spokes of the AISI and relevant WGs should develop guidance and SoPs for companies to undertake rigorous self assessment. There should be a transparency obligation on companies to submit reports to the AISI at pre-determined intervals.

Fostering an ecosystem of independent evaluators: Standards should be developed by the Evaluation Cell in conjunction with Spokes of the AISI and relevant WGs to enable auditors to undertake independent evaluations and assessments of AI systems. Criteria will have to be evolved by the AISI to designate institutions that can qualify to undertake these sociotechnical evaluations. Relatedly, it is key to consider funding models and mechanisms to enable independence of the evaluators and prevent evaluator shopping by companies.

AISI and expert-led assessments: As the Evaluation Cell at the AISI enhances capacity and has adequate staffing, the evaluation function could eventually complement the independent evaluator ecosystem, particularly in critical instances. As the AISI is being operationalised in the medium term, there is an opportunity to leverage the expertise of the Spokes of the AISI to pull together ad hoc taskforces for evaluations such as technical and social audits of specific AI systems. In the longer term, the Evaluation Cell at AISI can be staffed with experts to undertake continuous monitoring of emergent issues in the ecosystem and prioritise and implement mitigations. It is critical that staffing of the Evaluation Cell from the beginning is balanced in composition between technical experts and experts from the social sciences, law, policy and related fields as well as CSOs and other relevant organisations.

3. Functions

The institutional focus for the AISI can be grouped into the following three interconnected functions:

1. Research, Development and Expert Engagement

Short-term Recommendation: Developing research, evidence and guidance

Developing research and guidance: WGs and Spokes of the AISI can develop India specific risk assessment frameworks, evaluations benchmarks and quantitative and qualitative metrics. This could include process and management standards and product testing, and performance and measurement standards. Within these efforts, it is critical to bring social and policy dimensions into what have often been overly technical evaluation and benchmarks. Besides this, WGs and Spokes can build research that investigates and analyses the societal, economic and legal impacts of AI systems. All of this would be critical in building evidence and understanding of the AI ecosystem, harms arising and pathways ahead to prevent, mitigate and address these harms.

Long-term recommendations: Development of Shared Resources and Transparency Trackers

Development of AI Safety Commons: An AI Safety Commons will support access to datasets, benchmarks, tools and frameworks for the ecosystem. Some work has already been initiated around this including projects that have been funded by the IndiaAI Mission. It can be further populated with the work of the Spokes and Working Groups to be used not only domestically but internationally as well.

Transparency and Accountability Tracker: The AISI and its Spokes can also anchor transparency and accountability trackers around transparency disclosures on risk assessments, evaluations and voluntary commitments by industry and various government departments. The trackers could be critical for measuring progress on addressing identified risks, contributing to an evidence base of harms, bridging knowledge gaps, and the status of the implementation of any commitments.

2. Capacity Building: Transparency, Accountability and Training

Short-term Recommendation: Transparency mechanisms, Incident Reporting and Procurement Standards

Regulatory Capacity Building: In the context of social media regulation, one of the key challenges was empowering regulators with the requisite information around the working of these platforms and enabling them to have nuanced access to the kinds of harms arising. There are some learnings we can draw from social media regulation in this context and some actions that can be taken that are more particular to AI:

- *Transparency mechanisms:* AI platforms should give governments, evaluators and other stakeholders a record of what their system can do, how risks are being evaluated and addressed, model cards on the launch of new versions and risk reports.
- *Incident Reporting Database:* Companies should report significant safety incidents to a national database of AI incidents, to enable policymakers to have real-world evidence and insight into the risks and pattern of harms emerging from AI systems and in turn enable the development of relevant risk assessment and classification frameworks for India. This is the first step towards building accountability for developers and deployers in India and an evidence-based approach to future pathways for policy making.
- *Common Reporting Standards:* Currently, across the globe, even when reported, incidents involving AI systems are reported inconsistently. There is a need for common taxonomies for documenting failures, vulnerabilities, or misuse. This will allow us to draw insights, better identify harms and build safer systems.

Procurement Standards and Training: Given the government will be a significant procurer of technology across domains such as health, education and agriculture, developing government procurement standards is a way to set the floor for AI safety for the AI ecosystem in India. Often when government institutions are negotiating a contract with a vendor, the vendor is more technically knowledgeable and government functionaries may not have the skills, capacity or budgets to assess safety and monitor performance. The speed of AI development makes this capacity gap even more acute. The AISI can help bridge this gap by working with the Spokes and Working Groups to periodically put out procurement standards and guidance and support the monitoring of the safety performance of these systems as well as undertake training of government officials on these aspects.

Long-term Recommendations: AI Sandbox, Procurement Guidance and Capacity Building of Stakeholders

Regulatory Capacity Enhancement:

- *AI Sandbox:* AI sandboxes seek to foster innovation while enabling safety by testing within a controlled context in the pre-market stages. This can enable a sociotechnical evaluation of a system and assess its potential post deployment impact not only from a technical lens but a societal one. In conjunction with sectoral regulators, it is an opportunity for the AISI to set up AI Sandboxes to evaluate AI safety and sociotechnical implications for systems in critical sectors.
- *Collaboration with Sectoral Regulators:* Another key aspect of regulatory capacity building would be the AISI collaborating with sectoral regulators to enhance their capacity to deal with the implications of AI integration in their sectors such as finance, health, or education. While sectoral regulators would be well versed with the nuances of their domain, it would be imperative for the AISI to co-work with the sectoral regulators to bring its broader AI safety expertise to the table. In fact, it has been envisaged that AISI will *“form and oversee sectoral Sub-Committees that will complement efforts of and work in conjunction with sectoral regulators”*. This is an opportunity to involve multidisciplinary experts from beyond the domains of industry and government to build sociotechnical sector specific approaches to AI Safety. Such committees/ working groups can in collaboration with AISI and the sectoral regulators develop draft guidelines, standards, evaluation benchmarks and testing frameworks. In particular, it would be important for the AISI to work closely with sectoral regulators to respond to any significant incidents that may arise in their domains and develop research evidence on future pathways to policymaking and addressing harms in the sector.

Training: Fostering a multidisciplinary community may need capacity enhancement of experts from various disciplines to engage with the sociotechnical aspects of AI and contribute perspectives. In particular there is a need for capacity building of grassroots organisations working in domains that are seeing significant adoption of AI such as health and education, to empower them to engage with and identify the implications and challenges of adoption of AI in their domains.

3. International Cooperation and Coordination

Short-term Recommendation: Knowledge sharing and pooling of resources

While the Indian AISI can work towards contextualising and translating conceptions of AI safety to Indian realities, it can also take a lead in driving international coordination and cooperation amongst Global Majority countries through sharing learnings, pooling of

safety resources and serving as an example for other Global Majority countries, on how to navigate a process of translating AI safety to local socioeconomic contexts.

Long-term Recommendation: Coordinated network and building interoperable infrastructure

Coordinated Global Network: India should either orchestrate a Global Majority AI Safety Network of Institutes, or propose the creation of a subgroup within the current AI Safety Institutes Network to enable more structured collaboration, cross learning and coordination to ensure that sociotechnical approaches inform AI safety and hence adequately identify and address the challenges emerging from AI development and deployment in the Global Majority. This is an opportunity to share benchmarks, datasets, evaluation mechanisms and safety frameworks.

Interoperable standards and infrastructure: Given the reluctance of governments globally to regulate AI, standards will play a key role in governance and harmonisation globally. Consequently, it will be important to embed sociolegal notions, ethics, democratic values and user-centric approaches within the development of these technical standards and the AISI can play a critical role with respect to this not only domestically but internationally. This is in line with the Indian government's perspective which is that AISI should coordinate with government bodies and standard-setting organizations to ensure harmonization of efforts towards AI governance.

[← Previous Essay](#)

**Building an Intelligence Flywheel
for India's AI Economy**

[Share ↗](#)

[Next Essay →](#)

**Future of Digital Welfare in India:
Critical Reflections on Agentic AI
and Welfare Delivery Governance**