KAUTILYA

WORKING PAPER #3

Governing Artificial Intelligence in India: Data Sourcing, Synthetic Content, and Technological Sovereignty

Anoushka Sharan, Abhijeet Raut, Adarsh Roy, Annwesha Ghosh, Sidhant Singh & Sannidhya Seth

MAY 2026

Governing Artificial Intelligence in India: Data Sourcing, Synthetic Content, and Technological Sovereignty

Anoushka Sharan[1], Abhijeet Raut[1], Adarsh Philip Roy[2], Annwesha Ghosh[2], Sidhant Singh[2], Sannidhya Seth[2]

[1]Kautilya School of Public Policy, GITAM

[2]West Bengal National University of Juridical Sciences

Governing Artificial Intelligence in India: Data Sourcing, Synthetic Content, and Technological Sovereignty

Abstract

Artificial intelligence is possibly the most revolutionary technological development of the 21st century. While advances in artificial intelligence have been happening for a long time, it is only recently that it has started interacting with the broader public, triggering a plethora of complex and problematic instances. In this paper, we highlight three concerns, primarily related to data sourcing, synthetic content, and digital sovereignty. The data-related risks are explored in the context of the creative sector, and the concerns of digital sovereignty cut across diverse sectors. The concerns affect diverse policy actors, including AI service providers, users, the creative industry, students, and other communities. To alleviate the risks to these diverse impacted policy actors, we propose a detailed framework for sourcing data for training large language models by focusing on the type of data input and creating a mix of policies by drawing on provisions from the existing Digital Personal Data Protection Act. Finally, when it comes to the indigenisation of technology, we build on the existing IndiaAI mission by highlighting that, in the short run, a public-private partnership model can be sustained, with a focus on collectively bargaining with these technology firms. We believe that these policies are well-suited to the Indian context, as they build on the existing AI technical and legal policy ecosystem, acknowledging the very requirements and how to navigate them.

Keywords: Artificial intelligence; governance; data sourcing; digital sovereignty; synthetic content

Governing Artificial Intelligence in India: Data Sourcing, Synthetic Content, and Technological Sovereignty

1. Introduction

Knowledge, creative, and commercial operations are undergoing active transformation, with AI as one of the biggest drivers. These very systems, which promise economic productivity and inclusion, also inherently expose the ecosystem to vulnerabilities in data governance, content authenticity and technological sovereignty (Occhipinti et al., 2025).

In this paper, we address three challenges that directly face India's current ecosystem, which we found to be interlinked and structural. First, data sourcing for model training; second, synthetic content; and third, non-indigenous tech. It can also be clubbed into a supply-and-demand problem, with data clearly on the supply side, indigenous tech on the demand side, and synthetic content, a product of data sourcing and non-indigenous supply chains that cannot be monitored or subject to checks and balances.

This working paper draws on tools from design thinking, logical framework analysis, and the theory of change (ToC) to derive insights on risks and policy options. The aim is to integrate policy analysis, technological assessment, and human-centric design to ensure that final recommendations are empirically grounded and operationally feasible.

To derive Public Policy insights to address the risks and concerns of AI, it is imperative that we first identify the risks, understand the different stakeholders involved, and prioritise regulations. Focusing on AI technologies, trying to decipher and define them, will inevitably be an exercise that contributes nothing significant to policy studies due to the pacing problem: the technology evolves faster than regulation can keep pace. In this working paper, we consider diverse AI applications as AI+Systems, where the systems are the education and creative sectors. This will also enable us to explore the diverse policy actors involved in the system and how they are significantly affected by many of its applications.

1.1 Policy Actors

Among the diverse policy actors in the policy universe, a policy subsystem will be affected by different AI applications. We categorised them as impact groups derived from the works of (Pillai & Matus, 2020), analogous to Perrow's victim group classification (Perrow, 2011). This would include domestic and global AI developers, cloud service providers,

compute manufacturers, and even data marketplaces. Within the creative and cultural sector, artists, writers, musicians, folk creators, photographers, and content creators are the primary right-holders affected by AI training on copyrighted content. They frame the part of fair remuneration and consent frameworks. Within academia, universities and R&D institutes advance research, provide datasets, and invest further in developing technical standards. Civil society and consumers remain the primary stakeholders in every public policy problem, given the public nature of these problems. International partners determine the possibility and extent of interoperability.

Table 1 :Policy Actors and their Characteristics

Policy Actors	**Category**	**Characteristics**	**References to the existing frameworks and past research**
Domestic and global AI developers, cloud service providers, compute manufacturers, and data marketplaces	Service providers	Considerably higher knowledge about the AI system. Receives the benefits and lower risk exposure. Awareness on the technological advancement.	Victim Group 1 (Operators) of (Perrow, 2011).
Artists, writers, musicians, folk creators, photographers, and content creators. Students and teachers in universities and R&D institutes	Users who have used these applications.	Lower knowledge than the service providers, however, gains the benefits from the application of the technology. Have considerably higher exposure to risks than the service providers.	Victim Group 2 (Users) of (Perrow, 2011)
Artists, writers, musicians, folk creators,	Innocent bystanders- if they haven't	Has less knowledge than service providers and users and is not gaining any	Victim Group 3 (Innocent

photographers, and content creators. Students and teachers in universities and R&D institutes	used these systems, however, got impacted as the data (and the work), if being used to train these AI models.	benefits from the technology. Exposed to risks without benefitting from the technology. They are not aware that their data is being used.	bystanders) of (Perrow, 2011)
Other communities than those in the academic and creative industries – the civil society			Victim Group 3 of (Perrow, 2011) and Community as mentioned by (Pillai & Matus, 2020)

The regulators and other policy actors in the policy implementation subsystem focus on the direct, or sensory proximity of the person with the problem at hand. Upon integration, certain key stakeholders within the regulatory bodies were noted to be the Ministry of Electronics and Information Technology, the Ministry of Law & Justice, and CERT-In for policy design, enforcement, coordination across governance- data or non-data, cybersecurity, AI innovation, and the Ministry of Culture to represent and negotiate on behalf of the creators. Suppliers of infrastructure, models, and datasets who must also comply with provenance requirements and engage in full disclosure, while also having a stake in tech indigenisation, will be the industry and technology providers.

1.2 Identified Issues

In this working paper, we emphasise 3 key risks: data sourcing, consent generation, and sovereignty concerns when it comes to creative sector.

1. Data Sourcing: Ownership, Consent, and Fair Value

The reliability of AI models relies heavily on the datasets used to power them (Kapar et al., 2025). Data collection, as a process in itself, has become a point of contention, raising questions about consent, fairness, and ownership (Yao et al., 2025). The data is scraped from

the open web, with or without the original creator's knowledge. This is an implied challenge that sits squarely at the input layer of the AI value chain and determines both the ethical foundation and quality of AI outputs.

Unlicensed scraping and use of copyrighted or personal material erode both creator trust and data integrity. The dominance of Western datasets leads to systemic bias that marginalises local voices and limits the applicability of technology, especially for languages other than English (Paik et al., 2025). The lack of provenance and transparency in dataset construction makes it difficult to trace how AI systems make decisions or to hold developers accountable for harm (Bashayreh et al., 2023). Further, while large technology firms profit from these datasets, the creators and citizens whose data fuels them remain invisible and uncompensated.

2. Synthetic Content: Restoring Trust and Accountability

Generative AI creates what are called the “hyper-realistic” images, texts, audio and video contents without taking much time, very highly productive in that sense; however, pose the risks that this becomes indistinguishable from human- generated contents **(Gayathri et al., 2025)**. While the previous risks that we discussed are based on inputs, these are the characteristics of the outputs generated by LLMs. These are the ones that are nicknamed as deepfakes, which blur the line between reality and being fabricated. We have seen the implications of the same in the sphere of politics, where deepfakes created reputational damage, and it spread so fast, impeding public trust in our institutions. This risk is not restricted to political communications but has also seen significant ramifications on other sectors, such as academics, challenges to intellectual property rights and so on.

3. Non-Indigenous Technology: Building Digital Sovereignty

While the first issues is related to the inputs to the AI, the second one was related to the outputs. This one is more technological, the hardware and software technologies that power the AI systems. India is highly dependent on foreign technological infrastructure hampering it’s ability to develop indigenous innovation on the AI domain. This issue is very deep-rooted and need to be discussed on a systemic level addressing the diverse requirements of an AI infrastructure.

To address the AI race, relying on the foreign AI infrastructure will reap benefits short-term but definitely mask the long-term vulnerabilities. The lack of domestic compute capacity creates strategic and economic risks, as India’s AI development becomes tethered to external supply chains which can be seen as, unreliable in today’s techno-societal complex policy

regimes that remain contingent upon the rising multipolar world and thus also prone to change, and pricing structures that may or may not be determined by a superpower's engagement with the country. Such reliance also exposes sensitive data and model operations to the stakeholders in jurisdictions beyond India's legal reach, threatening India's cybersecurity and raising sovereignty concerns.

1.3 Interplay of the Three Issues

While the challenges pose individual barriers at different layers of the value chain, they cannot be isolated from one another. The nature of data sourcing determines the quality of AI models. The models generate content that, if the data is opaque, makes authentication difficult. Ultimately, this entire process is carried out within an infrastructure that is not Indian, thus offering little scope for true change.

As a result, a feedback loop opens: poor data governance feeds the AI models; the models then become unreliable and produce unverified content; and the lack of indigenous infrastructure weakens the country's ability to regulate or monitor the process.

2. Policy Proposals

The policy proposals that can address the diverse risks mentioned here fall into three categories: amendments to existing authority-based policy tools, awareness programs, and organisational or infrastructural development using the State's operational capacity. To address the risks related to data sourcing and synthetic content, we have amendment recommendations under diverse acts, such as the Copyright Act, 1957, and the Digital Personal Data Protection (DPDP) Act 2021, drawing on insights from regulations such as the GDPR. These are the regulatory amendments and strengthening regulatory oversight; however, other soft policy tools are also proposed in the policy mix to augment these regulatory tools, especially to build capacity to ensure digital sovereignty.

2.1 Recommendations for AI Scraping, Copyright, and Indigenous Art

1. Amend the Copyright Act, 1957, to Address AI Training Data:

The regulation of AI training on copyrighted and culturally significant material can proceed along two primary models.

The first is the text-and-data-mining (TDM) exception approach, widely used under the EU Digital Single Market (DSM) regime. In this proposed model, creators can opt out, preventing their works from being used for AI training. This opt-out can be implemented through technical

signals such as robots.txt, though its effectiveness is limited unless reinforced by statutory obligations. A more robust version would embed the opt-out requirement in an AI-specific legislative framework, thereby imposing binding duties on developers to honour individual and community-level exclusions from AI datasets.

The second option is a mandatory licensing system, which means that companies must get permission from copyright holders (an "opt-in" mechanism) before they can use their works to train models. Changes to the Copyright Act or the creation of a separate AI law could make this happen. Licensing can happen in two ways: directly with the creators or through a state-run system that sets standard terms and lowers the cost of transactions. Adding clear TDM rules that set apart non-commercial research (with the option for creators to opt out) from commercial AI training, which may need a license or payment depending on the situation. The TCE training should also require a licence to be valid. It is essential to embed safeguards for community-held cultural expressions and to incorporate collective moral rights into the legal framework.

2. Licensing & Royalties Framework – Establish Dedicated AI Societies:

Establishing AI societies that are similar to the copyright societies that will be vested with the power of licensing art and cultural materials needs to be established to facilitate transparency and accountability for the same. These societies will work in a fair manner through transparency in distributing licenses and will ensure that the indigenous communities are adequately recommended. These societies will definitely prevent the need for negotiations between the artists and the big AI developers, thereby preventing them from being exploited and facilitating fair distribution of money used for training purposes to be shared with the artists. This can be implemented through the amendments or provisions of the Copyright Act.

3. Strengthen Attribution, Consent, and Moral Rights:

It is difficult to attribute the outputs of the AI systems to the contribution of a particular artist, since the AI systems learn from patterns of a wide range of an enormous corpus of data that they use for training. However, there should be technological solutions, by some means, which should be made compulsory before deploying these AI systems that can enable attribution labels for the AI- generated content especially through embedding “machine-readable” metadata marking on the outputs that will be produced by the respectful algorithms.

To protect cultural integrity, all AI-generated images that imitate indigenous styles should include clear, community-respecting labels, for example:

Tag:

"This image is AI-generated in the style of Madhubani art. It does not represent authentic work created by Madhubani artists or the Mithila community and should not be treated as original or culturally authoritative."

We posit that this labelling exercise can prevent misrepresentation, confusion among consumers and the dilution of cultural significance, if operationalised through an opt-in mechanism for the artists that they can decide whether they want to enrol in the process of their contents being used for AI training and content generation.

4. Redressal and Enforcement Mechanisms:

India should establish regional tribunals specialising in AI, copyright, and cultural-rights disputes, and have them handle those disputes as first instances, subject to the supervisory jurisdiction of the High Courts under Article 226 of the Constitution. These tribunals should be allocated based on regional litigation loads, dividing India into five zones—North, East, South, Central, and West.

Appeals from these tribunals should lie directly with the respective High Courts and, further, with the Supreme Court of India, ensuring judicial oversight while preserving specialised, expertise-driven adjudication at the first level. The government should also formally promote mediation and conciliation-based Alternative Dispute Resolution (ADR) for AI-art and cultural-heritage disputes, including those affecting indigenous communities. Expedited takedown procedures (ideally within 48 hours) must be mandated to protect creators from ongoing harm.

5. Align with Global AI, Copyright, and Data Protection Standards:

We propose the following measures to align with the global regulations and standards:

1. Treat large-scale scraping of identifiable images as personal-data processing under the DPDPA. Clarify in the DPDPA on the issue of widespread scraping of publicly available data.

2. Require Data Protection Impact Assessments (DPIAs) before training models using scraped images, especially where content includes identifiable individuals or culturally sensitive data. This can be done by adding AI companies as SDFs.
3. Amend Section 39 of the DPDPA to allow civil compensation for unauthorised scraping of personal or cultural data, aligning India with GDPR-style remedies.
4. Adopt transparency obligations inspired by the EU AI Act, requiring disclosure of dataset categories and risk-mitigation measures.

6. Architectural, Infrastructural, and Technical Safeguards via AI Act or sui generis law:

The following safeguards through the AI Act will ensure the architectural, infrastructural and technical safeguards:

1. Develop machine-readable "noAI" metadata that web crawlers must legally honour.
2. Maintain high-level provenance documentation for all training datasets, filed with MeitY as part of annual mandatory disclosures.
3. Mandate dataset filtering and periodic retraining whenever communities or creators opt out.
4. Promote watermarking, provenance tracking, and authenticity standards to help communities identify imitative AI outputs.

7. Mandatory Transparency and Scraping Oversight via the AI Act:

To mandate the transparency of the AI processes and the outputs so generated, the AI developers must publish annual transparency reports disclosing dataset sources, scraping practices, and risk controls. These reports must be filed with MeitY as a statutory obligation, with penalties for non-compliance.

We also propose establishing a Meity Scraping Registry in which companies must register the datasets used to train high-impact generative models, including summaries of sources (e.g., LAION, stock libraries, cultural archives).

8. Public Awareness, Community Support, and Capacity Building

Finally, to build awareness, we propose launching nationwide programmes to educate artists and indigenous groups about the risks of scraping, data rights, attribution issues, and legal remedies. It is imperative to promote tools such as "Have I Been Trained" to help creators verify whether their work appears in datasets.

2.2. Addressing the foreign dependency of the technology

When it comes to the foreign dependency of technology, particularly key technologies linked to artificial intelligence, there is a systemic understanding that there must be a push for indigenization of these technologies and the India AI mission, particularly, has the following pillars that can ensure the data set availability, infrastructure buildup and all the relevant cycles of the AI value chain are worked upon in order to achieve tech independence and indigenization. The relevant policies under the IndiaAI mission and the objectives are the same as follows:

> Under the IndiaAI Compute Pillar, the Mission is developing a scalable AI computing ecosystem to support India's growing AI startup and research community. This initiative includes the establishment of a state-of-the-art AI compute infrastructure featuring 18,000+ GPUs, built through public-private partnerships. The Union Minister of Electronics & IT, Railways, and I&B has announced that eligible users can access AI compute at up to 40% reduced cost under the IndiaAI Mission, which has a budgetary outlay of ₹10,372 Cr.

Beyond the aforementioned long-term policy to achieve technology indigenisation, addressing how startups can deal with high subscription costs, limited computing power and chip availability, etc. While the answer lies essentially in the longer run with the success of India AI mission to achieve technology indigenization and heightened affordability, for starters, there can still be a shorter run, small time policy whose aim is to quickly enable the government to negotiate with larger multinational firms that are dealing with artificial intelligence in order to provide subsidized rate of these crucial technologies to startups within the country.

The government of India can make provision for the formation of an independent board with 2 secretaries, one representing MeitY and the other the Ministry of Corporate Affairs, and representation from India Inc. and relevant think tanks, including NITI Aayog. The board would encourage private players to launch a special category for business models. The model will include the following provision: organisations will reach out to individual content creators, such as artists and researchers, to obtain consent for AI companies to use their creative works for training their LLMs. The core principle here is that companies can increase the net bargaining power of these creators through union-like efforts. The increase in bargaining power means that it not only increases net revenue for creators, but also allows a stronger bargaining

power to Indian companies and given that these companies will act on the behest of the board, it will allow them to negotiate lower local prices for subscriptions, etc., which can then be provided to startups in the critical sector. The board will also take separate initiatives to negotiate with AI companies to offer these subscriptions at much lower prices to startups in return for easing regulations and improving Ease of Doing Business.

3. Conclusion

AI's rapid diffusion has outpaced existing legal and institutional frameworks, exposing students, artists, and indigenous communities to significant integrity, privacy, and cultural harms. In the creative ecosystem, large-scale scraping of copyrighted and culturally significant material, the absence of clear TDM rules, weak enforcement capacity, and lack of protection for Traditional Cultural Expressions (TCE) enable uncompensated extraction and cultural misappropriation.

A single regulatory strategy that anchors on some definitions of AI won't address these concerns, as systemic risks remain unaddressed: the rapid advancement and in-depth proliferation of technology. Here, we argue that incremental, tool-centric fixes are insufficient. India requires a coherent regulatory architecture: national guidelines on AI use in the creative sector, systematic faculty training, integrity-by-design assessment reforms, and a risk-based AI Act that classifies and regulates high- risk educational systems, coupled with due-process guarantees and shared liability for developers and institutions. The suggested policy interventions include amendments to existing regulations, awareness programs, and infrastructure developments to uphold AI sovereignty. In parallel, amendments to the Copyright Act, a dedicated licensing and royalty framework for AI training data, a sui generis regime for TCEs, and MeitY-led transparency and oversight mechanisms are needed to realign incentives, recognise community rights, and provide practical avenues for redress.

Though these are abstract-level analyses, they lay the foundation for further in-depth research on each of these risks, test the diverse frameworks and derive regulatory insights. The insights are not empirically driven; however, they are exploratory in their contribution to ongoing policy discussions on the procedural and systemic challenges of AI governance.

References

Bashayreh, M. H., Tabbara, A., & Sibai, F. N. (2023). The need for a legal standard of care in the AI environment. *Sriwijaya Law Review*, 73–86.

Gayathri, N., Kumar, S. R., Chandran, R., Chelliah, P. R., & Pelusi, D. (2025). *Generative AI*. wiley.

Kapar, J., Koenen, N., & Jullum, M. (2025). What's wrong with your synthetic tabular data? using explainable ai to evaluate generative models. *World Conference on Explainable Artificial Intelligence*, 19–43.

Occhipinti, J.-A., Prodan, A., Hynes, W., Buchanan, J., Green, R., Burrow, S., Eyre, H. A., Skinner, A., Hickie, I. B., & Heffernan, M. (2025). Artificial intelligence, recessionary pressures and population health. *Bulletin of the World Health Organization*, *103*(2), 155.

Paik, S., Novozhilova, E., Mays, K. K., & Katz, J. E. (2025). Who benefits from AI? Examining different demographics' fairness perceptions across personal, work, and public life. *Discover Artificial Intelligence*, *5*(1), 39.

Perrow, C. (2011). *Normal accidents: Living with high risk technologies-Updated edition*. Princeton university press.

Pillai, V. S., & Matus, K. J. M. (2020). Towards a responsible integration of artificial intelligence technology in the construction sector. *Science and Public Policy*, *47*(5), 689–704. https://doi.org/10.1093/scipol/scaa073

Yao, J., Liu, F., & Han, B. (2025). Trustworthy AI under Imperfect Web Data. *Companion Proceedings of the ACM on Web Conference 2025*, 65–68.