I ISSUE BRIEF

SPRF.IN

Navigating India's AI Governance Framework:

Future Directions in #AIforAll

10
25

Abhiramy Sudhi Manjusha

TABLE OF CONTENTS

If you have any suggestions, or would like to contribute, please write to us at contact@sprf.in

OCTOBER 2025

Issue Brief

Navigating India's AI Governance Framework:

Future Directions in #AIforAll

II INTRODUCTION

From healthcare to education, Artificial Intelligence hype has covered every sector, with some calling it the greatest breakthrough since splitting the atom (Price, 2023) while others leveling up their research on existential threats due to AI (Future of Life Institute, 2023). Whether or not AI is worth the hype, constant and steady progress is being made globally in its development and deployment (Asia News Network, 2025; Bhuiyan, 2025). This will eventually impact all sectors of governance, and hence needs critical scrutiny. It is clear is the deployment of AI has also shown great potential for harm (Shroff, 2025), especially for women, gender minorities, and groups in other vulnerable situations like platform workers, immigrants, and refugees (AI Incident Database, 2017; Oppenheim, 2018; Hersey, 2022). Amidst this, countries are stepping up efforts to craft governance strategies and policies to balance safeguarding the interests of their citizens while allowing businesses to thrive. This pushes researchers and policy analysts to think beyond the binaries.

India currently ranks fourth globally in the Stanford Human-Centered AI (HAI) Global AI Vibrancy Tool[1], with a score of 25.54 across a balanced range of indicators including responsibility, R&D, policy and governance, infrastructure, and public engagement. This ranking underscores several core strengths: India leads globally in AI conference citations, holds the third position in AI journal publications, and ranks second worldwide in AI-related GitHub projects, highlighting the vibrancy of its developer ecosystem (Jeevanandam, 2024). Public discourse around AI is equally robust—India stands second globally in both AI-related social media voice share and total AI posts, reflecting widespread societal engagement and awareness (ibid.). However, despite these notable achievements, India still trails behind leading nations—especially the U.S. and China—in critical domains like policy and governance, responsible AI and technological infrastructure, which suppress its overall vibrancy score.

The Finance Minister of India, in his budget speech for 2018 – 2019, mandated NITI Aayog to establish the National Program on AI. In pursuance, NITI Aayog drafted a report on “National Strategy for Artificial Intelligence”. This report put forth a unique brand of #AIforAll. The report focused on how India can leverage transformative technologies to ensure social and inclusive growth in line with the development philosophy of the government (NITI Ayog, 2018).

This paper examines the conceptual and practical dimensions of building AI for All, with a focus on inclusivity, accessibility, and socio-technical equity. The paper uses a decolonial framework not simply to critique the social costs of technological advancement, but to foreground how colonial systems operated through extractive logics that depleted the economic, technological, epistemic, and social capital of colonized societies. The concern here is not a binary between technological progress and equity, but rather a deeper interrogation of how the legacies of colonial knowledge hierarchies and resource extraction shape contemporary AI systems and infrastructures and how can India prioritize technological advancement with an inclusive design. In this light, building equitable AI in India must involve a commitment to data sovereignty, epistemic inclusion, and historical redress, rather than merely localizing global models.

Drawing on critical works done on responsible and inclusive AI by scholars from the Global South, I advocate for a framework that is indigenous to India and ground it in feminist decoloniality. I propose two guiding principles that should steer inclusive AI governance frameworks in India. First, AI systems differ from existing artificial entities (of governance, e.g. banking system) in the capacity of AI to increasingly make decisions without humans in the loop. Moreover, AI systems play an

1 . https://hai.stanford.edu/ai-index/global-vibrancy-tool

increasingly important role in the operation of all the other artificial entities; for example, in financial markets, only a small sliver of all transactions are executed by humans. As a result, they leave a footprint on all aspects of the governance ecosystem. This pushes all stakeholders to rethink how capable our existing systems of governance are to welcome AI into the system (Bullock et al., 2023). Hence, the values we embed in AI systems become extremely important. Second, not thinking of AI as a governance black box (Kosinski, 2024) and enabling a consistent, open and rigorous research environment around AI. Both of these principles are inseparable from each other and without upholding both together, we risk stymying the other.

I HUMANS OUT-OF-LOOP

2025 has been marked as the year of AI agents (IBM, 2025); we have moved on from generative AI models (GPTs) to conquer new territory (Knight, 2025; Gülen, 2025). Google defines AI agents as "software systems that use AI to pursue goals and complete tasks on behalf of users. They show reasoning, planning, and memory and have a level of autonomy to make decisions, learn, and adapt." Although the concept of autonomous AI has been explored for decades in theory and research, the presence of AI agents that operate independently in our daily lives, such as managing schedules, coordinating meetings, and handling routine tasks, has only recently become a practical reality (IBM, 2024). These AI assistants are evolving from simple tools to proactive collaborators, capable of navigating complex social and organizational contexts with minimal human intervention. This shift marks a significant milestone in the integration of AI into everyday life, raising new possibilities as well as challenges related to trust, privacy, and user autonomy. This poses serious challenges for governance given the rapid progress these developments are taking and raises serious questions on where humans fit in this decision making process.

Researchers (Kapoor et al., 2024; Ng, 2024; Greyling, 2024; Lambert, 2024; Roucher et al., 2024; Mitchell et al., 2025) argue that AI agents should be understood on a sliding scale of autonomy. Figure 1 shows AI agents with varying levels of autonomy based on Mitchell et al., (2025). Their main argument is that there are no clear benefits of fully autonomous AI agents, but many foreseeable harms from ceding full human control.

Figure 1: Levels of AI Agent

Agentic Level	Description	Term	Example Code	Who's in Control?
☆☆☆☆	Model has no impact on program flow	Simple processor	`print(llm_output(llm_response))`	Human
★☆☆☆	Model determines basic program flow	Router	`if llm_decision(): path_a() else: path_b()`	Human: How functions are done; System: When
★★☆☆	Model determines how functions are executed	Tool call	`run_function(llm_chosen_tool, llm_chosen_args)`	Human: What functions are done; System: How
★★★☆	Model controls iteration and program continuation	Multi-step agent	`while should_continue(): execute_next_step()`	Human: What functions exist; System: Which to do, when, how
★★★★	Model creates & executes new code	Fully autonomous agent	`create_code(user_request); execute()`	System

Systems using machine-learned models can have different levels of agency. They can also be combined in "multiagent systems," where one agent workflow triggers another, or multiple agents work collectively toward a goal. (Mitchell et al., 2025)

Beyond AI agents, there are far more serious threats posed by autonomous decision making systems which are covered under frontier AI research. “Frontier AI models” are highly capable foundation models, which could have dangerous capabilities that are sufficient to severely threaten public safety and global security. Examples of capabilities that would meet this standard include designing chemical weapons, exploiting vulnerabilities in safety-critical software systems, synthesising persuasive disinformation at scale, or evading human control (Anderljung et al., 2023).

Hence, the first guiding principle proposes that as AI based decision making increasingly takes humans out of the loop, we need to focus on what are the values embedded in these systems and how are these prioritized in a profit-hungry system. (Bullock et al.) talk about two types of desirable goals that both play critical roles in AI development, but that require distinct governance approaches: Direct alignment: when an AI system is pursuing goals consistent with the goals of its operator, irrespective of whether it imposes externalities on other parties. Social alignment: when an AI system is pursuing goals that are consistent with the broader goals of society, taking into account the welfare of everybody who is impacted by the system. The concept of social alignment becomes even more important when AI is used for social good for example welfare targeting, healthcare etc. as in NITI Ayog’s AI for all.

The discussion of social alignment of AI must be contextualized within decolonial theories of AI, focusing on the inequality in resource allocation between cultures, languages and knowledge systems. Coloniality describes domination, including in abstract forms such as in the production of knowledge, that remains after the end of formal colonialism (Quijano, 2007). Decoloniality is the process of challenging and dismantling this coloniality (Mignolo & Escobar, 2010). (Bhambra et al. 2018) state that decolonizing has two key referents. First, it is a way of thinking about the world which takes colonialism, empire and racism as its empirical and discursive objects of study; it re-situates these phenomena as key shaping forces of the contemporary world, in a context where their role has been systematically effaced from view. Second, it purports to offer alternative ways of thinking about the world and alternative forms of political praxis. Hence, before examining how digital systems replicate and intensify existing power dynamics, it is important to recognize that “the digital” itself is not a neutral or universal construct. Rather, it is embedded in specific languages, epistemologies, and cultural frameworks. What counts as data, knowledge, or participation online is prefigured by these foundations, privileging certain worldviews while marginalizing others. Blanco et al. (2023) underscores how much of the vocabulary used in conversations around technology and data governance—particularly terms like “data protection” and “data privacy”—originates from Global North legal and cultural frameworks. These concepts often have no direct translations in many Global South languages, making it difficult to convey their full meaning or urgency in local contexts. This linguistic gap reflects a deeper issue: the dominance of colonial and Western epistemologies in shaping the digital discourse, even down to the words we use to talk about rights, ethics, and autonomy online.

Hence, I put forth a case for the Collective Benefit, Authority to Control, Responsibility, and Ethics (CARE) principles to be embedded in India’s AI governance framework, not just to align AI with societal needs, but to make it truly inclusive by foregrounding feminist and decolonial perspectives that challenge dominant power structures and prioritize historically marginalised voices.

I COLLECTIVE BENEFIT, AUTHORITY TO CONTROL, RESPONSIBILITY, AND ETHICS (CARE)

One of the major principles on data sharing and governance, the FAIR principle or, Findability, Accessibility, Interoperability, and Reuse of digital assets emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data (GO FAIR, 2017). But FAIR primarily focuses on characteristics of data that will facilitate increased data sharing among entities while ignoring power differentials and historical contexts (Global Indigenous Data Alliance, 2019). Sambasivan et al. (2020) show that conventional algorithmic fairness principles are Western in their sub-groups, values, and optimizations. In India, the distance between ML models and dis-empowered communities whom they aim to serve is large—via technical distance, social distance, ethical distance, temporal distance, and physical distance. Therefore, a myopic take on localising 'fair' model outputs alone can backfire.

To support this hypothesis and address the problem, the CARE Principles for Indigenous Data Governance are applied. The CARE Principles for Indigenous Data Governance are a set of principles intended to guide open data projects and was created in 2019 by the International Indigenous Data Sovereignty Interest Group, a group that is a part of the Research Data Alliance. Unlike FAIR, the CARE principles centre on the well-being of Indigenous Peoples and their data, and can be implemented alongside the FAIR Principles throughout data lifecycles to ensure collective benefit (Roberts & Montoya, 2023).

Figure 2: FAIR and CARE principles in complementarity (Global Indigenous Data Alliance, 2019)

Source: Global Indigenous Data Alliance, 2019

Although vastly different from each other in many significant ways, communities in the Global South share key commonalities with Indigenous communities, particularly histories of colonialism, exploitation, and decades of extractive data practices. Inclusion of CARE principles in the Global South is a research area still in its budding stages with scholars working in this interdisciplinaryfield coming together to build knowledge[2].

II COLLECTIVE BENEFIT

From a decolonial and feminist perspective, AI should not simply be optimized for efficiency but must actively confront the legacies of digital colonialism, racialized data practices, and gendered power imbalances. Care ethics helps us ask: Who benefits? Who is harmed? Of the 7,000 languages spoken, around 2,500 are considered endangered. (Kornai, 2013) argued around 12 years ago that this underestimates the threat of digital extinction as languages present evidence of a "massive die-off" caused by the digital divide. Even fewer than 5% of languages have a viable digital presence. Despite the progress of a decade, the gap has barely shrunk (Joshi et al., 2020). The colonial character of this divide is inalienable to its study. (Helm et al., 2024) show that the number of Wikipedia pages for Kiswahili, one of the major African languages with about 80 million speakers, is as high as for Breton, an endangered Celtic language in western France with about 200 thousand speakers.

The principle of Collective Benefit ensures that the advantages of AI integration and advancements are not confined to those who already benefit from historical and ongoing systems of coloniality and global inequality. Instead, it calls for AI to be developed and deployed in ways that prioritize the well-being of communities in the global south, redistribute power, and promote inclusive participation across linguistic, cultural, and geopolitical boundaries. Projects such as Papa Reo[3] (Māori language platform) illustrate how AI grounded in indigenous lifeways can generate cultural and social value, ensuring benefits flow back to the source community under kaitiakitanga (guardianship) principles. Te reo Māori, spoken by approximately 190,000 people today[4], is a language that has endured immense colonial suppression and aggressive assimilationist policies, particularly within New Zealand's education system. Papa Reo builds services around Te reo Māori, enabling smaller indigenous language communities to develop their own speech recognition and natural language processing capabilities, ensuring that the sovereignty of the data remains with them and the benefits derived from these technologies goes directly to their communities. Kaituhi is one such initiative—a web-based transcription tool that automatically transcribes te reo Māori and New Zealand English audio and video files and has capability to detect code-mixing and local cultural differences within the language. In addition to this, existing machine learning techniques require large data sets to support the development of speech to text, text to speech and speech synthesis. Papa Reo aims to make these tools and methods available for languages with smaller data sets to ensure their languages are present in the digital age.

In India, the introductions of initiatives like Bhāshini and BharatGPT language tools aim to serve citizens broadly and can be a powerful tool in ensuring collective benefit because of indigenous knowledge production. A CARE based approach insists that these systems be co-designed with marginalized communities (e.g., Adivasi, Dalit, linguistic minorities, women), so the benefits—cultural, economic, social—are shared in a non-extractive and meaningful way. For a language-heterogeneous country like India it could include initiatives like having marginalized community members in the

2 . https://inclusiveailab.org/program/
3 . https://papareo.nz/
4 . https://www.ethniccommunities.govt.nz/resources/our-languages-o-tatou-reo/languages-in-nz

design team and introducing these tools to languages that are often under-served. Hindi, despite being the fourth most spoken language in the world, accounts for less than 0.1% of web content - largely due to the absence of robust lexical and computational resources. The digital presence of other Indian languages is even more limited, highlighting a stark linguistic disparity online. As per the latest census report (Census, 2011), there are 1,369 mother tongues currently being spoken in the country. Kumar & Lahiri (2022) find that barring the 22 scheduled languages as per constitution, rest of the non-scheduled and endangered Indian languages have only negligible technological resources in comparison to the other resource-rich languages or even the 22 scheduled Indian languages. Only 10 non-scheduled languages of the country have at least some kind of language resource or technology available for it in the form of OCR, Grammatical Analysers and Text-to-speech converters.

II AUTHORITY TO CONTROL

CARE emphasizes community sovereignty over data and technologies. Feminist decolonial frameworks stress that marginalized groups must participate in design and governance - embodying "nothing about us, without us" (Ricaurte, 2025). Democratization of AI means not only that people can freely use AI, but also that people can collectively decide how AI is to be used. Monopolized in the hands of private corporations in the Global North, the development of the most capable foundation models has proceeded largely without input from the Global South. The citizens that have generated the data necessary to train models do not have input on how their data are to be used (Chan et al., 2023). India with its linguistic and cultural diversity coupled with its demographic enormity is an active knowledge creation space with centuries of history and risks losing control over it with LLMs that train on publicly available data.

Sometimes AI systems are developed externally, building in safety mechanisms from the outset isn't always possible. Instead, defensive actions and safeguards often need to be implemented after deployment especially in contexts like India, where data sovereignty is a priority and generative AI models like ChatGPT, Perplexity, and Claude fall outside direct national governance. Agnew et al. (2024) define and build "data defenses"—a novel strategy that directly empowers data owners to block LLMs from performing inference on their data. They create data defenses by developing a method to automatically generate adversarial prompt injections that, when added to input text, significantly reduce the ability of LLMs to accurately infer personally identifying information about the subject of the input text or to use copyrighted text in inference.

II RESPONSIBILITY

Responsible AI is a composite of key concepts like fairness, safety, privacy and security and transparency. But without a theoretical grounding in the histories of exclusion that produced inequities in fairness, safety, privacy, security, and transparency, Responsible AI remains a shallow promise. AI models, frontier or not, need to be embedded with responsible AI concepts that acknowledge these inequalities and take steps at each stage—from inception to deployment—to ensure that these are not reproduced in a digital form. Transparency and resulting trust and safety might be more important for communities that have been excluded systematically for decades. For instance, facial recognition systems piloted by police departments across India (Venkatesh, 2024) have raised concerns over biased surveillance, profiling and lack of consent (Sinha, 2024) especially for those who are more vulnerable to police brutality (Narula, 2000). "Facial recognition systems enable covert and remote mass authentication. This means that they work without providing notice of their existence and use, require no direct interaction with the subject, and intended deployments are usually not targeted

to specific suspects, but designed to surveil everyone." This is in line with the idea of a patriarchal authority taking advantage of AI advancements to further propagate violence in the name of protection. Narayanan & Kapoor (2024) argue how broken AI models are appealing to broken institutions. In practice, CARE demands shared accountability across stakeholders, disrupting patriarchal and colonial power structures leading to the creation of systems with embedded responsibility.

II ETHICS

Ethical harms caused by AI are emerging as a distinct research subfield, supported by growing repositories that document how AI impacts individuals and societies on multiple levels in everyday life[5][6]. Harms can range from data labor exploitation (Perrigo, 2023) to surveillance (Sen & Bennett, 2022). But these harms can be multiplied in Global South contexts. In her seminal work Unmasking AI, Buolamwini (2023) articulates the concept of "the coded gaze"- the evidence of encoded discrimination and exclusion in tech products. Applying an intersectional lens to both the tech industry and the research sector, she shows how racism, sexism, colorism, and ableism can overlap and render broad swaths of humanity "excoded" and therefore vulnerable in a world rapidly adopting AI tools. Hence, CARE framework advocates that active and consistent efforts need to be embedded in AI systems to prevent systemic biases from creeping into codes and hence amplifying ethical harms caused by AI.

This is especially important with respect to Ethical AI in employment. The Fairwork India Report 2023 that assessed 12 gigwork platforms—bigbasket, BluSmart, Swiggy, Urban Company, Zomato, Zepto, Flipkart, Amazon Flex, Dunzo, Uber, Ola, and Porter—show that many platforms have achieved worse scores than the previous year according to the Fairwork AI Principles (Fairwork, 2023b) . The Fairwork AI Principles offer a powerful foundation for reimagining worker-centric AI governance, emphasizing fair pay, safe conditions, transparent management, equitable contracts, and collective representation (Fairwork, 2023a). Therefore, ethical AI must move beyond technical fixes to interrogate what ethical values are embedded within the system and what power structures are reproduced or disrupted through its rollout.

I BREAKING OPEN THE BLACK BOX

Black box AI models refer to artificial intelligence systems where the internal decision-making process is opaque or incomprehensible, even to the people who built them (IBM, 2024b). In recent years, there has been growing concern over the deployment of deep learning systems in high-stakes environments where errors can result in serious harm (Hendrycks et al., 2023). These concerns have prompted calls to slow down or place stricter limits on the use of such technologies in critical sectors like healthcare, law enforcement, and transportation.

A notable example of a structured regulatory response is the European Union's proposed Artificial Intelligence Act, which aims to establish a comprehensive legal framework to govern AI applications across member states. The framework categorizes AI systems based on the level of risk they pose: high-risk systems, such as those used in biometric identification, critical infrastructure, or hiring decisions, are subject to stringent obligations including transparency, accountability, and human

5 . https://incidentdatabase.ai/
6 . https://airisk.mit.edu/ai-incident-tracker

oversight; in some cases, their use may be banned altogether (European Commission, 2021). In contrast, low-risk applications, such as spam filters or AI-powered recommendation engines, are subject to minimal regulatory intervention, allowing for more flexible innovation. This tiered approach reflects a precautionary principle, balancing technological advancement with the ethical imperative to prevent harm, particularly in contexts where deep learning's opacity and complexity may obscure unintended consequences (Floridi et al., 2018).

While concerns about the opacity of "black box" AI systems are shared globally, the approach to addressing these issues must be adapted to the specific political, economic, and socio-technical contexts of each region. In India, the regulatory response to opaque AI systems cannot simply mirror other regional frameworks. For India, the global discourse around regulating black-box AI must be recontextualized within a post-colonial framework that prioritizes economic development, technological sovereignty, and social equity. Unlike the European Union, which can afford to take a precautionary and heavily regulated stance on AI due to its economic stability and institutional infrastructure derived from colonial extraction, India faces a dual imperative: to mitigate technological risks while rapidly advancing innovation. Caution cannot come at the cost of progress.

As a nation with a history of colonial extraction, India cannot afford to simply replicate the regulatory conservatism of the Global North. Therefore, the first guiding principle ensures that the values we embed in our systems are derived from a CARE based approach and strives to uphold our indigenous knowledge systems. Building on the first principle, the second emphasizes the need for open, consistent, and rigorous research, while safeguarding data sovereignty and protecting indigenous knowledge systems from appropriation or erasure.

This principle is grounded in work done by Narayanan & Kapoor (2024) and Arora (2019; 2024), which argue that the way forward for AI research and deployment is not through fear-driven restrictions but through a proactive, optimistic, and evidence-based approach that distinguishes real capabilities from exaggerated claims. While Narayanan & Kapoor (2024) caution against both blanket skepticism and hype that stifles responsible innovation, Arora (2019) argues about the perils of western-technocratic narratives of "the digital south". Both of these arguments are critical to be embedded within India's AI research initiatives.

Narayanan & Kapoor (2024) point out how harmful but cheap (in both the resource and time sense) AI technology appeals to broken institutions. They identify how key actors in the system—technocrats, policymakers, academicians and the media—contribute to the AI hype. Chronically underfunded news outlets amplify hyped claims by researchers and companies. Journalists often lack the time or specialized knowledge to thoroughly fact-check these claims, leading them to publish minimally revised press releases as news stories. "And they amplify claims made by public figures, such as public intellectuals and tech CEOs, rather than engaging in grounded discourse urging instead for rigorous empirical validation and open, accountable inquiry" (ibid).

Further they argue that "Describing AI as unknowable reduces our agency by positioning AI as something that we can never understand and therefore never challenge" (Narayanan & Kapoor, 2024). This bolsters a regulatory approach that focuses on breaking open the black box. Rigorous research not just from a purely technological standpoint but from interdisciplinary fields that can critique,analyze and innovate solutions on AI such as digital humanities, policy and sociology is imperative in building systems that are focused on minimizing harms while maximizing economic and technological progress. The hype and harm narratives surrounding AI can be addressed in a more systematic and inclusive manner if India's governance framework prioritizes open, consistent, and rigorous research across disciplines that study AI.

Arora's seminal work The Next Billion Users (2019) critically dismantles Western-centric assumptions about technology adoption in the Global South and argues that India is one of the countries that has "The Next Billion Users" of digital technologies. She argues that digital users in countries like India are not passive recipients of technology but active agents who appropriate digital tools in ways shaped by local desires, aspirations, and constraints. Furthermore, in her more recent work (Arora, 2024), she expands on this by calling for a shift from pessimism to optimism in how the Global South is framed in AI discourse, advocating for frameworks that recognize creativity, resilience, and agency rather than deficit and vulnerability. These insights are crucial for shaping India's AI governance and research priorities. Embedding such perspectives ensures that AI development in India is not only technologically sound but also socially attuned, enabling a narrative that emphasizes contextual innovation over imported caution.

A practical solution to this comes in the form of explainable AI (XAI) systems, that promote transparency and accountability. It meets the increasing necessity for research that caters to the needs of general audiences, policymakers, and governance decision-makers—individuals who may not require a deep technical understanding of how AI works but nonetheless hold critical stakes in its outcomes. Explainable Artificial Intelligence (XAI) is an evolving area of research which has received a lot of attention from the research community and the broader society alike (NITI Ayog, 2018). India has been proactive by pointing out in the NITI Ayog report—National Strategy for Artificial Intelligence—that looking at the countrywise H-Index (a metric that quantifies a country's scientific productivity and scientific impact), India ranks a dismal 19th, globally. In other words, while India may be producing research pieces in numbers, their utility has been rather limited. The report also strongly argues for Incentivising Core and Applied research in AI which is visionary but given the document's focus on AI for All should integrate what that truly means. Given the disproportionately high dropout rates among historically marginalized castes in higher education, it is essential to critically examine the AI and tech research spaces we are building and ask who they include, exclude, and serve (Radhakrishnan, 2021).

AI-based learning systems can be great tools to work on exclusion but only when they are intentionally designed for equity, with local-context adaptation, bias mitigation, inclusive co-design, and robust ethics frameworks - all key to the CARE framework. Organizations like Education Above All[7], which through its Digi-Wise initiative and Ferby chatbot aim to deliver culturally relevant, accessible, and responsibly designed AI tools to learners in developing regions, can be promoted and incentivized. Education Above All emphasizes human-centered design, broad accessibility, and the ethical augmentation of traditional educational practices. Promoting and supporting local organizations that design AI tools grounded in Indian languages, resource realities, and cultural contexts will ensure relevance and reach along with being cost-effective. Despite this, AI tools should not be considered as a one stop solution to exclusion, as exclusion is a systemic issue which needs a holistic approach to be tackled.

7 . https://www.educationaboveall.org/blog/our-advocacy/ai-education-means-more-equitable-educational-future

II CONCLUSION

Revolutionary strides in AI development in India should be coupled with the principles argued in the paper, otherwise we risk building a system that is unsustainable and unjust, and will fall apart sooner or later. Embedding these principles require institutional support, resources, and political will (Ricaurte, 2025). I argue for AI rooted in Feminist and Decolonial epistemology that points out that knowledge, technology design, and technological practice are not value neutral but rather products of sociotechnical realities that reflect specific values and agendas. A feminist position suggests that simple attentiveness to inequality is not enough, and instead AI systems and explanatory efforts should be directed at active erosion of inequalities and contribute towards just and equitable sociotechnical AI ecosystems. There are three key ways to implement this in explainable AI research: Attention to power and structural inequalities, Challenging universalist and traditional rationalist modalities of explanation, and Centring marginalised perspectives (Klumbyte et al., 2023).

Further, this paper emphasizes the importance of design and governance frameworks on building ethical and AI models but it is also pertinent to address the material impact of AI on labor and social reproduction, particularly in the Global South. This is increasingly visible in India, where companies like TCS have begun phasing out mid- and senior-level roles due to automation—not just work typically considered low-skill. This reconfigures earlier assumptions about AI's threat to "routine" jobs and instead reveals how AI's deployment is reshaping workforce hierarchies, eroding roles traditionally associated with decision-making and profit-sharing. Moreover, this process aligns with the logic of capitalist creative destruction (Schumpeter 1942), where technological innovation displaces existing systems to make way for new, often more exploitative forms. In this context, future research on questions of AI governance must go beyond fairness and transparency to include redistributive justice and equitable digital access, all of which remain unevenly distributed across India's socio-economic and geographic spectrum.

As India stands at the crossroads of rapid technological advancement and deep societal transformation, its approach to AI governance must resist the temptation to simply replicate Euro-American regulatory frameworks. The European Union's risk-based AI Act provides a useful reference, but its emphasis on caution and control must be adapted to India's distinct socio-political realities and post-colonial development imperatives (European Commission, 2021). With its history of colonial extraction and its ongoing pursuit of economic and technological sovereignty, India cannot afford to be paralyzed by narratives of fear and black-box pessimism. Instead, the country requires a governance model that fosters responsible innovation, grounded in context-sensitive oversight and interdisciplinary research. Critiquing western technocratic notions Arora (2019) writes, "Co-creation with the poor appears to be an enlightened approach. There is a thin line, though, between collaboration and exploitative commodification. In the process of making legitimate consumers out of the poor, technology companies harness their data."

REFERENCES

Agnew, W., Jiang, H. H., Sum, C., Sap, M., & Das, S. (2024). Data Defenses Against Large Language Models. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2410.13138

AI Incident Database. (2017, October 26). Incident 13: High-Toxicity Assessed on Text Involving Women and Minority Groups. Incidentdatabase.ai. https://incidentdatabase.ai/cite/13/

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., ... & Wolf, K. (2023). Frontier AI regulation: Managing emerging risks to public safety. arXiv preprint. arXiv:2307.03718

Arora, P. (2019). The next billion users: Digital life beyond the West. Cambridge, Massachusetts Harvard University Press.

Arora, P. (2024). From Pessimism to Promise. MIT Press.

Asia News Network. (2025). China tops global AI development with 1,509 large models. Asianews.network. https://asianews.network/china-tops-global-ai-development-with-1509-large-models/

Bhuiyan, J. (2025, July 30). Zuckerberg claims “superintelligence is now in sight” as Meta lavishes billions on AI. The Guardian. https://www.theguardian.com/technology/2025/jul/30/zuckerberg-superintelligence-meta-ai

Bullock, J. B., Hudson, V. M., Korinek, A., Young, M. M., & Dean, M. (2023). The Oxford Handbook of AI Governance. Oxford University Press.

Buolamwini, J. (2023). Unmasking AI. Random House.

Chan, A., Bradley, H., & Rajkumar, N. (2023, May 21). Reclaiming the Digital Commons: A Public Data Trust for Training Data. ArXiv.org. https://doi.org/10.48550/arXiv.2303.09001

Chowdhary, R. (2024). Shifting Paradigms of Multilingual Publishing and Scholarship in India. The Journal of Electronic Publishing, 27(1). https://doi.org/10.3998/jep.5592

Fairwork. (2023a, July 27). Fairwork AI Principles. Fair.work. https://fair.work/en/fw/principles/ai-principles/

Fairwork. (2023b). Fairwork India Ratings 2023: Labour Standards in the Platform Economy. In Fairwork . https://fair.work/en/fw/publications/fairwork-india-ratings-2023-labour-standards-in-the-platform-economy/

Future of Life Institute. (2023, March 22). Pause Giant AI Experiments: an Open Letter. Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Greyling, C. (2024, October 11). 5 Levels Of AI Agents (Updated) - Cobus Greyling - Medium. Medium. https://cobusgreyling.medium.com/5-levels-of-ai-agents-updated-0ddf8931a1c6

Gülen, K. (2025, July 28). Meet Opal: Google’s vibe coding AI agent. Dataconomy. https://dataconomy.com/2025/07/28/meet-opal-googles-vibe-coding-ai-agent/

Helm, P., Bella, G., Koch, G., & Giunchiglia, F. (2024). Diversity and language technology: How language modeling bias causes epistemic injustice. Ethics and Information Technology, 26(1). https://doi.org/10.1007/s10676-023-09742-6

Hendrycks, D., Mazeika, M., & Woodside, T. (2023, June 26). An Overview of Catastrophic AI Risks. ArXiv.org. https://doi.org/10.48550/arXiv.2306.12001

Hersey, F. (2022, September 15). Prominent facial recognition researcher scraped videos of trans people, left dataset exposed. Biometric Update. https://www.biometricupdate.com/202209/prominent-facial-recognition-researcher-scraped-videos-of-trans-people-left-dataset-exposed

IBM. (2024a, October 11). Agentic AI. https://www.ibm.com/think/insights/agentic-ai

IBM. (2024b, October 29). What is black box artificial intelligence (AI)? https://www.ibm.com/think/topics/black-box-ai

IBM. (2025, March 4). AI Agents in 2025: Expectations vs. Reality. https://www.ibm.com/think/insights/ai-agents-2025-expectations-vs-reality

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity

and Inclusion in the NLP World. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Kissinger, H., Schmidt, E., & Huttenlocher, D. P. (2021). The age of A. I.: And our human future. Little Brown & Company.

Klumbyte, G., Piehl, H., & Draude, C. (2023). Towards Feminist Intersectional XAI: From Explainability to Response-Ability. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.03375

Knight, W. (2025, July 16). Former Top Google Researchers Have Made a New Kind of AI Agent. WIRED. https://www.wired.com/story/former-top-google-researchers-have-made-a-new-kind-of-ai-agent/

Kornai, A. (2013). Digital Language Death. PLoS ONE, 8(10), e77056. https://doi.org/10.1371/journal.pone.0077056

Kosinski, M. (2024, October 29). What is black box artificial intelligence (AI)? IBM. https://www.ibm.com/think/topics/black-box-ai

Kumar, R., & Lahiri, B. (2022). Language Resources and Technologies for Non-Scheduled and Endangered Indian Languages. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2204.02822

Lambert, N. (2024, December 18). The AI Agent Spectrum. Interconnects.ai. https://www.interconnects.ai/p/the-ai-agent-spectrum

Magallanes Blanco, C., Mhlambi, S., Nyabola, N., Couldry, N., Nothias, T., & Siminyu, K. (2023, March). Decolonizing Data, One Language At A Time [Interview]. In Digital Civil Society Lab at Stanford University. https://pacscenter.stanford.edu/publication/decolonizing-data-one-language-at-a-time/

Mignolo, W. D., & Escobar, A. (2010). Globalization and the Decolonial Option. In Routledge eBooks. Informa. https://doi.org/10.4324/9781315868448

Mitchell, M., Ghosh, A., Luccioni, A. S., & Pistilli, G. (2025). Fully Autonomous AI Agents Should Not be Developed. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2502.02649

Narayanan, A., & Kapoor, S. (2024). AI Snake Oil. Princeton University Press.

Narula, S. (2000). Broken people : caste violence against India's untouchables. Human Rights Watch.

Ng, A. (2025). X (Formerly Twitter). https://x.com/AndrewYNg/status/1801295202788983136

NITI Ayog. (2018). National Strategy for Artificial Intelligence. In NITI Ayog. https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf

Oppenheim, M. (2018, October 11). Amazon scraps "sexist AI" recruitment tool. The Independent. https://www.independent.co.uk/tech/amazon-ai-sexist-recruitment-tool-algorithm-a8579161.html

Perrigo, B. (2023, January 18). Exclusive: OpenAI Used Kenyan Workers on Less than $2 per Hour to Make ChatGPT Less Toxic. Time. https://time.com/6247678/openai-chatgpt-kenya-workers/

Price, C. (2023, November). AI is greatest breakthrough since splitting the atom, says King Charles. The Telegraph. https://www.telegraph.co.uk/business/2023/11/01/king-charles-ai-greatest-breakthrough-safety-summit/

Quijano, A. (2007). Coloniality and modernity/rationality. Cultural Studies, 21(2-3), 168–178.

Radhakrishnan, V. (2021, August 14). Data | Drop out rate of SC/STs relatively higher in IITs and IISERs. The Hindu. https://www.thehindu.com/data/data-drop-out-rate-of-scsts-relatively-higher-in-iits-and-iisers/article61430188.ece

RAND. (2024, March 21). Artificial Intelligence and Biotechnology: Risks and Opportunities. Www.rand.org. https://www.rand.org/pubs/articles/2024/artificial-intelligence-and-biotechnology-risks-and.html

Ricaurte, P. (2025). How can feminism inform AI governance in practice? Unesco.org. https://www.unesco.org/en/articles/how-can-feminism-inform-ai-governance-practice?

Sambasivan, N., Arnesen, E., Hutchinson, B., & Prabhakaran, V. (2020). Non-portability of Algorithmic Fairness in India. ArXiv.org. https://arxiv.org/abs/2012.03659

Sen, A., & Bennett, D. K. (2022). AI flagged "irrelevant" posts as unsafe at colleges, investigation shows.

Dallasnews.com. https://interactives.dallasnews.com/2022/the-black-box-social-sentinel/

Shroff, L. (2025, July 24). The Atlantic. The Atlantic; theatlantic. https://www.theatlantic.com/technology/archive/2025/07/chatgpt-ai-self-mutilation-satanism/683649/

Sinha, A. (2024, March 13). The Landscape of Facial Recognition Technologies in India | TechPolicy.Press. Tech Policy Press. https://www.techpolicy.press/the-landscape-of-facial-recognition-technologies-in-india/

Venkatesh, S. (2024, December 17). The legal gaps in India's unregulated AI surveillance. The Hindu. https://www.thehindu.com/opinion/lead/the-legal-gaps-in-indias-unregulated-ai-surveillance/article68996389.ece

WWW.SPRF.IN

If you have any suggestions, or would like to contribute, please write to us at **contact@sprf.in**