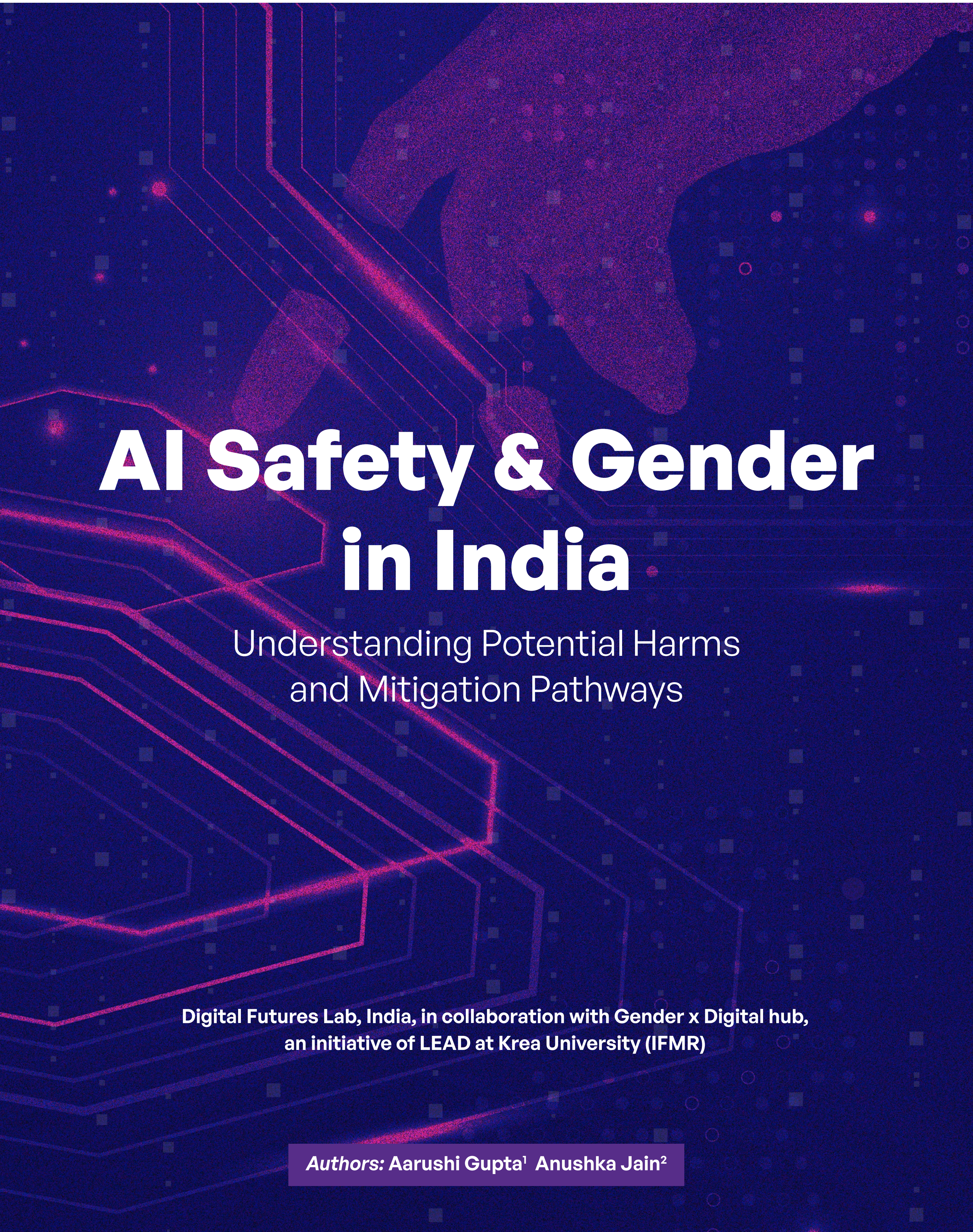

AI Safety & Gender in India

Understanding Potential Harms and Mitigation Pathways

Digital Futures Lab, India, in collaboration with Gender x Digital hub,
an initiative of LEAD at Krea University (IFMR)

Authors: Aarushi Gupta[1] Anushka Jain[2]

About

digital futures lab

www.digitalfuturesslab.in

Digital Futures Lab is an India-based, interdisciplinary research network that examines the complex relationship between technology and society in the Majority World. Through evidence-based research, participatory foresight, and public engagement, we identify pathways toward equitable, safe, and caring futures.

GxD hub
Digital Equity. Connectivity. Insights.

www.gxdhub.org

Gender x Digital hub (GxD hub) is a cross-disciplinary platform dedicated to advancing meaningful digital connectivity among women and girls in India. It is an initiative of LEAD at Krea University (IFMR). LEAD at Krea University is an action-oriented research centre housed at the Institute for Financial Management and Research (IFMR), a not-for-profit society which is also the Sponsoring Body of Krea University.

Acknowledgements

This brief emerges from Key Informant Interviews (KII) and secondary research conducted by the Digital Futures Lab from November 2025 to February 2026, as part of a collaborative study undertaken with the GxD hub (an initiative of LEAD at Krea University).

We extend our sincere gratitude to Sharon Buteau, Executive Director, LEAD at Krea University, for her inputs in shaping the direction of this research, and Diksha Singh, Associate Director of Communications and Outreach, for editorial inputs on the draft.

We extend our deepest gratitude to all the experts who generously shared their time and insights with us - Nihar Sahoo (BharatGen), Saurabh Karn (Sarvam AI), Shweta Gupta (Microsoft), Alpan Rawal (Wadhwani AI), Bishakha Datta (Point of View), Shohini Bannerjee (Point of View), Gokul S. Krishnan (Centre for Responsible AI, IIT Madras), Geetha Raju (Centre for Responsible AI, IIT Madras), Sunayana Sitaram (Microsoft Research India), Zara Schroeder (Research ICT Africa), & Shweta Chaudhury (Haqdarshak).

We would also like to thank Anita Gurumurthy (IT For Change) & Shweta Gupta (Microsoft) for being generous with their time and providing us with valuable feedback during our peer review process.

The views and opinions expressed herein are those of the contributing authors, and do not necessarily represent the official views of any organisation or entity associated with this research.

Authors: Aarushi Gupta, Anushka Jain (Digital Futures Lab)

Technical Leads: Urvashi Aneja (Digital Futures Lab),[3] Yashita Jhurani (GxD hub – LEAD at Krea University)[3]

Design: Sakthivel Arumugum, Keerthana Ramaswamy (LEAD at Krea University)

Image Credit: Pexels/EqualStock IN

Table of Contents

1. Background & Context

As AI systems are increasingly integrated into a wide range of everyday services and technologies in India, they play an important role in shaping how people access information, public services, and economic opportunities. Informational chatbots, AI-enabled speech and language tools for Indic languages, generative media platforms, and predictive systems used in service delivery are now increasingly getting adopted across both public and private sectors.[4] In parallel, there are growing efforts to integrate AI into elements of India's digital public infrastructure (DPI), further expanding the reach and influence of these systems.[5]

The increasing deployment of AI across sectors is accompanied by a growing body of evidence on the risks associated with these systems. AI models have been shown to produce biased or discriminatory outcomes, amplify misinformation and harmful content, and operate with limited transparency, making it difficult for affected users to understand, contest, or seek redress for harmful outputs.[6]

The Need for a Contextualised Understanding of AI Harms in India

While these risks are widely acknowledged, they do not manifest uniformly across users or contexts. In practice, the harmful impacts of AI systems are shaped by the socio-cultural contexts, institutional settings, and user profiles among which these systems are deployed.[7]

There is thus an urgent need to contextualise global AI risk frameworks and understand the unique forms these risks take within the diverse contexts presented by the Indian landscape. The recently released AI Governance Guidelines emphasise a similar approach, recommending the development of a suitable AI risk assessment and classification framework for India that accounts for its unique social, economic, and cultural context, based on which appropriate risk mitigation measures can be deployed.[8]

However, while contextualisation at a country level is necessary, it is insufficient on its own. Within a country as diverse as India, AI-related harms are likely to be experienced unevenly across different groups of users, based on their gender, income level, geographical location, caste, education level, and a variety of intersecting factors therein.[9] As a result, any overarching or aggregate risk framework can obscure how the harmful impacts of AI are distributed and experienced in practice.

Therefore, in response, this brief proposes a disaggregated framework for understanding AI harms, using gender as a starting lens to examine how many of the generic AI risks are experienced differently or disproportionately by different groups of users.

Gender is a particularly important lens in the Indian context, given persistent gender gaps in digital access, differential exposure to online harms, and the ways in which AI systems can amplify existing inequalities in public service delivery, information access, and economic participation.[10]

Inclusivity as an Integral Dimension of AI Safety

In providing a contextualised taxonomy of gendered harms of AI, this brief also seeks to redefine what safety in AI means in practice within the Indian context, where AI systems are increasingly being deployed at scale, across diverse linguistic, cultural, and socio-economic settings.

In India, AI is not confined to experimental or high-end applications, but is being integrated into public service delivery, digital platforms, and population-scale technologies, often in environments marked by uneven digital literacy, variable institutional capacity, and deep social inequalities.[11] In such contexts, narrow or purely technical notions of AI safety risk overlooking more immediate and consequential harms, including exclusion, unequal system performance, automation bias, and barriers to redress.[12]

This brief, therefore, adopts a broader and more grounded understanding of AI safety, one that extends beyond model robustness or misuse to consider how AI systems interact with real users, institutional workflows, and existing service delivery or grievance redressal infrastructures. When AI is integrated into population-scale systems such as DPI, safety cannot be an afterthought. It must shape how these systems are designed and used, particularly in terms of who can access them, who is protected, and how harms are addressed.

Therefore, while this brief adopts a gender lens, the framework and recommendations are designed to enable more inclusive AI safety practices that are responsive to the needs of a wider set of marginalised and underrepresented communities, including rural and remote populations, linguistic minorities, persons with disabilities, informal sector workers, and communities facing various other intersecting socio-economic disadvantages.

Overall, the brief challenges the tendency to treat inclusivity and safety as separate pillars, arguing instead that AI safety is inherently an inclusion-exclusion question: safe for whom, in which contexts, and at whose cost. In this, we view inclusivity as not an adjacent objective, but a defining condition of what it means for an AI system to be safe.

Structure of the Brief

We provide our key findings and arguments in three parts. Section 1 outlines a contextual taxonomy of gender-related AI safety risks as they manifest across different use cases and deployment settings in India. This initial taxonomy was built through extensive secondary research on AI-enabled gendered harms and risks, along with virtual interviews with technology developers, AI researchers, and feminist scholars (see the Annexure for a detailed note on our methodology).

Section 2 then examines the gaps in existing safety tools and evaluation practices, highlighting where current approaches in India fall short in addressing these risks. Finally, in Section 3, building on our analysis of harms and gaps in safety tooling, we set out a targeted set of recommendations for policymakers, regulators, philanthropies, and AI practitioners to ensure more inclusive and safe AI innovation in India.

Taken together, this brief offers a set of starting points for many of these stakeholders working in a fast-changing AI landscape. We also want to acknowledge that the challenges discussed in this brief are complex and cut across policy, technology, and institutions, and cannot be addressed through any single measure. Therefore, rather than seeking to be exhaustive or overly prescriptive, the brief outlines broad practical directions to help stakeholders navigate this space and take more informed action over time.

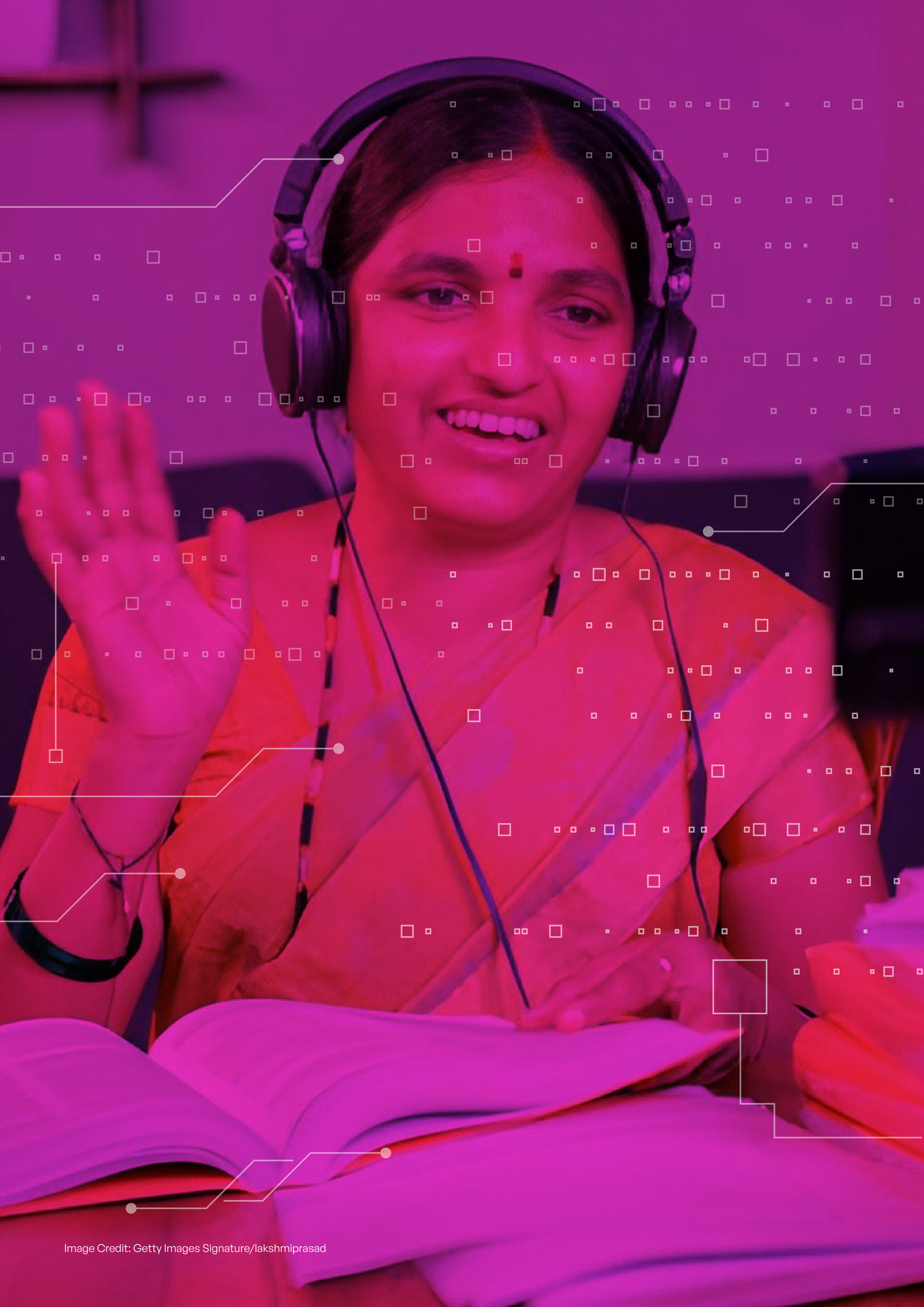

Image Credit: Getty Images Signature/lakshmiprasad

2. Understanding Gendered Harms of AI in India

Across India, AI systems are increasingly embedded in digital platforms, public service delivery, financial services, workplaces, and online spaces. Early evidence from real-world deployments points to a range of gendered harms, including algorithmic exclusion in financial services, AI-enabled harassment, surveillance, and non-consensual imagery, socio-cultural stereotyping and misclassification, as well as harms that undermine women's participation in public life and access to rights-based entitlements.[13]

While these harms are increasingly visible through ground reports by media or civil society organisations, they are rarely examined together as part of a coherent AI risk assessment framework.[14] Instead, they tend to be discussed in isolation, as sector-specific or issue-specific problems, limiting the ability of policymakers and developers to identify recurring risk patterns or design targeted safety interventions.

To address this gap, this section proposes an initial taxonomy of gendered AI harms in the Indian context (see Table 1). The taxonomy organises these harms across four broad domains — economic, bodily and autonomy-related, socio-cultural, and political — and provides illustrative evidence from the Indian context, where feasible.15

The taxonomy builds on existing AI harms frameworks, but adapts them to focus specifically on gender and on the social and institutional realities of India.[16] Please note that this taxonomy is intended to support more systematic safety assessments, contextual evaluations, and policy action, rather than to serve as an exhaustive catalogue of all possible harms.

Table 1: A Taxonomy of Gendered Harms of AI in India

Harm Category	Harm	Description	Emerging Evidence
Economic Harms AI can negatively impact women's livelihood and financial opportunities	Opportunity loss due to unsafe online spaces	Unsafe online spaces, riddled with harassment, non-consensual intimate imagery (NCII), and fraud, may curb women's active participation on digital platforms. In response to the prevalence of such risks, women may tend to self-censor or avoid online platforms altogether, losing out on professional, entrepreneurial, and civic engagement opportunities that are increasingly mediated through digital platforms.	Recent evidence on online harms for women in India shows that the fear of being targeted using AI-based nudification and deepfake tools has led several women to avoid/limit their presence on the internet.[17]
	Opportunity loss due to biased platform algorithms and policies	Opportunity loss due to biased platform algorithms and policies occurs when opaque practices such as downranking, shadow-banning, or demonetisation limit users' visibility and income without clear justification or recourse. These unaccountable decisions can disproportionately restrict women and marginalised groups' access to economic opportunities and participation in digital spaces.	Studies indicate that some AI-based image moderation and classification tools disproportionately label images of women as sexually suggestive relative to comparable images of men. This disparity is especially evident in cases involving visible aspects of female anatomy, pregnant bodies, or athletic activity. Consequently, when social media companies rely on these or similar algorithmic systems, they risk systematically reducing the reach of content featuring women's bodies. This can adversely affect women-led businesses that depend on visual engagement, particularly in sectors such as fitness, maternal health, reproductive education, and apparel.[18]
	Opportunity loss due to biased algorithms in labour and financial markets	Discriminatory or gender-blind algorithms in hiring, pay, appraisal, and credit scoring may fail to account for structural gender inequalities, such as women's career breaks for caregiving, safety-related constraints on work, and non-linear employment patterns.[19] Trained on historical data reflecting male-dominated labour markets and continuous full-time work norms, such hiring or performance appraisal algorithms may penalise women for gaps in experience, while credit models may reproduce biases linked to gender gaps in income, employment continuity, and asset ownership. As a result, such discriminatory AI systems may reinforce existing disparities in employment, pay, promotion, and access to credit rather than mitigating them.	In India, machine-learning(ML)-based credit systems have expanded access to finance overall, but evidence shows they often allocate larger loan amounts to men than to women, reinforcing existing gender gaps in financial inclusion. This pattern is driven in part by credit-scoring models that rely on gendered proxies such as continuous formal employment and income, which can disadvantage women with career breaks or those working in the informal sector, resulting in lower loan approvals or smaller ticket sizes, despite women's stronger repayment outcomes.[20]
	Job displacement due to AI-led automation in sectors with high participation from women	While there is currently limited evidence of large-scale AI-driven job displacement in India, particularly given the dominance of informal and low-resource labour markets, existing research suggests a plausible risk that AI-based automation could disproportionately affect women as adoption expands. This risk is most evident in sectors where women are over-represented, such as clerical, administrative, customer support, and data-entry roles, which are among the earliest targets of automation globally.[21]	....
	Normalisation of online gender-based harassment/violence	Content recommendation systems on platforms that are optimised for virality through engagement metrics, amplify and normalise gender-based harassment by prioritising misogynistic, toxic content. Algorithms slowly expose users to increasingly radical material, starting with "soft" humour, escalating to overt violence glorification, framing it as entertainment to boost shares and views. This creates feedback loops where harassment goes mainstream, desensitising audiences and fueling real-world harms against women.	Studies indicate that social media recommendation systems can amplify and normalise harmful ideologies by increasing users' exposure to radical material, including misogyny and gender-based violence, often framing such content as entertainment that attracts high engagement.[31] Evidence further suggests that the core business models of digital platforms, which prioritise monetisation and user engagement, create structural incentives for the amplification of emotionally charged and misogynistic content, reinforcing cycles of toxicity, visibility, and profit.[32] While India currently lacks systematic, publicly available evidence that quantifies these dynamics in the Indian context, these findings offer important insights and analytical frameworks that can inform assessments of similar risks within India's digital ecosystem.

Table 1: A Taxonomy of Gendered Harms of AI in India

Harm Category	Harm	Description	Emerging Evidence
Harms to Physical Safety and Personal Control AI can threaten physical safety or restrict an individual's control over their body, movements, or personal decisions	Gendered stereotyping	AI systems may perpetuate gendered stereotyping by embedding societal biases from training data into outputs, for example, portraying women as emotional or domestic, while men may appear as public leaders or be portrayed as rational actors.[25]	Large language models customised for Indic languages may reproduce gender stereotypes in their outputs, including the systematic association of certain occupations (e.g. man = doctor, woman = nurse) or activities (e.g. woman = cook, man = go to work) with specific genders. This is evident when gender-neutral English pronouns are translated into gendered forms in low-resource languages such as Bengali.[26]
	Misclassification	AI systems may misclassify gender minorities, including non-binary and transgender individuals as many of them are trained to default to male/female labels, leading to errors in facial recognition or identity verification.[27]	Evidence from testing on datasets of Indian faces shows that commercially available facial processing tools for face detection, gender classification, and age estimation are unable to accurately classify individuals of the third gender.[28]
	Cultural misrepresentation	Cultural misrepresentation may occur when AI systems oversimplify, stereotype, or inaccurately represent communities due to biased training data and dominant cultural narratives.[29] In India, this harm intersects with gendered stereotypes to misrepresent women by reinforcing reductive norms around appearance, behaviour, roles, and morality.	Evidence indicates that text-to-image models overwhelmingly generate outputs depicting outfits of female-presenting Indians as wearing traditional Indian garments such as sarees.[30]
	Normalisation of online gender-based harassment/violence	Content recommendation systems on platforms that are optimised for virality through engagement metrics, amplify and normalise gender-based harassment by prioritising misogynistic, toxic content. Algorithms slowly expose users to increasingly radical material, starting with "soft" humour, escalating to overt violence glorification, framing it as entertainment to boost shares and views. This creates feedback loops where harassment goes mainstream, desensitising audiences and fueling real-world harms against women.	Studies indicate that social media recommendation systems can amplify and normalise harmful ideologies by increasing users' exposure to radical material, including misogyny and gender-based violence, often framing such content as entertainment that attracts high engagement.[31] Evidence further suggests that the core business models of digital platforms, which prioritise monetisation and user engagement, create structural incentives for the amplification of emotionally charged and misogynistic content, reinforcing cycles of toxicity, visibility, and profit.[32] While India currently lacks systematic, publicly available evidence that quantifies these dynamics in the Indian context, these findings offer important insights and analytical frameworks that can inform assessments of similar risks within India's digital ecosystem.
Political Harms AI can impinge on women's ability to participate in political life and access their constitutional rights	Targeted disinformation and harassment of politically vocal women	AI-based targeted disinformation and harassment involve the use of automated and generative systems to amplify false narratives, abuse, and threats against politically vocal and/or active women. These systems are often deployed to produce coordinated smear campaigns, deepfakes, sexualised misinformation, and mass harassment at scale, exploiting gendered stereotypes to undermine women's credibility, silence dissent, and deter political participation. In India, these harms are frequently intersectional, disproportionately targeting women from marginalised communities and reinforcing existing power imbalances in public and political life.	A UN Women-commissioned survey of women in the public sphere, focused on women human rights defenders, activists, and journalists, found that 24% of surveyed participants had experienced AI-enabled online violence.[33]
	Exclusion from rights-based entitlements	AI-enabled exclusion from rights-based entitlements occurs when automated decision-making systems used for identification, eligibility determination, or service delivery deny or restrict women's access to welfare, healthcare, employment benefits, or financial services. These systems often rely on incomplete, inaccurate, or gender-blind data that fail to reflect women's lived realities, such as name changes after marriage, informal or unpaid work, limited documentation, or irregular access to digital systems. As a result, AI systems can systematically exclude women, particularly those from marginalised communities, from essential public services and legal entitlements, reinforcing existing structural inequalities, with limited avenues for seeking grievance redressal.	Entity resolution is an ML-based process to link records referring to the same individual across different databases. When such algorithmic systems are poorly implemented in welfare service delivery, errors in data can lead to arbitrary and unaccountable denial of benefits. Reports on Telangana's Samagra Vedika, an integrated data platform designed to create a comprehensive profile of citizens by merging various government databases, highlight these risks.[34] Women may be disproportionately affected due to factors such as name or address changes after marriage.

Image Credit: Getty Images Signature/Deepak Sethi

3. Mapping AI Safety Tools for Gender

The harms outlined in the preceding index underscore an important distinction in how gendered risks emerge in AI systems in practice. While some harms are closely tied to data gaps, model design choices, or evaluation blind spots, others are more deeply embedded in institutional workflows, platform incentives, and governance arrangements within which AI systems are deployed. This distinction has direct implications for the kinds of interventions that can meaningfully mitigate risk.

A growing ecosystem of AI safety tools has emerged globally to support risk identification, evaluation, and mitigation across different stages of the AI lifecycle. These include data-related tools such as dataset augmentation methods; model- and evaluation-focused tools such as bias benchmarks, red-teaming, and reinforcement learning through human feedback; procedural tools such as model cards, audits, and impact assessments; and platform-level tools such as algorithmic content moderation. Collectively, these tools are often positioned as mechanisms to surface harmful behaviours, improve system performance across demographic groups, and reduce the likelihood of biased or unsafe outputs.

Table 2 provides an overview of commonly referenced AI safety tools and evaluates their relevance for addressing gendered harms, with a particular focus on the Indian context. Rather than assessing these tools in the abstract, the table examines who primarily uses them, whether they have seen application in India, and the key gaps that limit their effectiveness and applicability for mitigating gendered AI risks in the Indian context.

Table 2: An Overview of Existing AI Safety Tools

AI Safety Tool	Purpose & Relevance for Gender	Primary User of the Tool	Application within the Indian Context (If any)	Gaps Identified
Data-related Tools				
Data documentation cards	These tools provide structured documentation describing dataset provenance, biases, demographic coverage, annotation pipeline, and known limitations (including gender skew).[35]	Upstream model developers, Downstream application developers, Auditors, Policy	....	**Tool Adoption Gap:** Data cards are widely referenced as best practice, but are inconsistently produced and rarely required in Indian AI development, procurement, or public dataset hosting. **Gender and Intersectionality Gap:** Even when documentation exists, it often omits gender-relevant fields (such as gender representation and intersectional coverage, gendered harm risks, annotation guidance on gendered abuse), limiting its usefulness for diagnosing downstream bias and exclusion.
Counterfactual Data Augmentation (CDA)	This tool is used to generate minimally altered text examples (gender swaps) to mitigate stereotypes or spurious correlations.[36]	Upstream model developers, Downstream application developers	Upstream model developers, Downstream application developers	**Socio-cultural Contextualisation Gap:** Tools are adopted but with limited adaptation to India's diverse socio-cultural contexts
Curated Pretraining / Finetuning Data	This tool provides high-quality, representative multilingual corpora.[37]	Upstream model developers, Downstream application developers	1. IndicLLM Suite[38] 2. AI4Bharat-IndicNLP Corpus[39]	**Gender Specificity Gap:** Tools are adapted to Indian contexts, but do not explicitly encode gender or intersectional gender risks
Model-related Tools				
Benchmarking Datasets	This tool provides high-quality, representative multilingual corpora.[37]	These datasets are used to evaluate and compare the performance of algorithms or models.	1. IndiBias[40] 2. BharatBBQ[41] 3. IndiCASA[42] 4. OTSC-Hindi and Wino-MT Hindi[43] 5. Uli[44]	**Coverage Gap:** Existing safety tools only address a narrow subset of gendered harms
Red-teaming	This is a structured method to surface, characterise, and document harmful behaviours, such as stereotypes or bias against women, that emerge when AI systems are used in real-world, adversarial, or edge-case contexts, especially those that standard benchmarks and metrics fail to capture.[45]	Upstream model developers, Downstream application developers, Auditors, Platforms, Policy	1. Deccan AI's red-teaming dataset for evaluating cultural biases in LLMs.[46]	**Epistemic Gap:** Weak feedback loops between practice and research on gender-focused safety tooling
Reinforcement Learning through Human Feedback (RLHF)	This technique uses adversarial prompts or human feedback to reduce stereotypes or harmful outputs.[47]	Upstream model developers	....	**Socio-cultural Contextualisation Gap:** Tools are adopted but with limited adaptation to India's diverse socio-cultural contexts **Gender Specificity Gap:** Tools are adapted to Indian contexts, but do not explicitly encode gender or intersectional gender risks

AI Safety Tool	Purpose & Relevance for Gender	Primary User of the Tool	Application within the Indian Context (If any)	Gaps Identified
Procedural Tools				
Model cards	These are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type) and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains.[48]	Upstream model developers, Downstream application developers, Auditors, Policy	....	**Tool Adoption Gap:** Tools exist, but are sparsely applied to AI systems in India
AI Audits	These are systematic evaluations of an AI system in operation to determine whether it performs as claimed, produces biased or harmful outcomes, and complies with legal or ethical standards. It may be used to help anticipate gendered harms.[49]	Auditors, Upstream model developers, Downstream application developers, Policy	....	**Tool Adoption Gap:** Tools exist, but are sparsely applied to AI systems in India
Impact Assessments	These are structured processes conducted before (and sometimes during) deployment to identify, assess, and mitigate risks associated with an AI system. It may be used to reveal gender disparities in system performance.[50]	Downstream application developers, Policy, Auditors, Upstream model developers	....	**Tool Adoption Gap:** Tools exist, but are sparsely applied to AI systems in India
Platform-related Tools				
Algorithmic content moderation	This method uses machine learning or rule-based systems to detect, flag, demote, or remove content (text, images, audio, video) that violates platform rules, often at scale.[51]	Platform providers	....	**Socio-cultural Contextualisation Gap:** Tools are adopted but with limited adaptation to India’s diverse socio-cultural contexts

3.1 Prominent Gaps in Current AI Safety Tooling for Gender

Taken together, the mapping of AI safety tools highlights a set of recurring limitations in how safety practices are currently designed, adopted, and operationalised in Indian contexts. While a wide range of tools exist across the AI lifecycle, including documentation practices, audits, benchmarks, red-teaming methods, and impact assessments, their effectiveness in addressing gendered harms in India is constrained by several structural gaps.

A first and widely visible limitation is uneven adoption. Many of the tools listed in the table are well established in global AI safety discourse and are routinely referenced in policy discussions and technical guidance. However, their use in India remains inconsistent and highly uneven.[52] Given the capacity, resources, and time needed to develop, customise, and deploy existing safety tools, their adoption is largely concentrated among large technology firms or specialised technical institutes. Smaller developers and social impact organisations face several challenges in meaningfully integrating or building safety suites for the AI tools they are developing.

A second limitation concerns the lack of gender specificity. Even where safety tools or evaluation techniques are adapted for Indian languages or datasets by domestic actors, they often do not explicitly encode gender or intersectional gender risks.

This gap is visible, for example, in widely used evaluation benchmarks for Indic language models such as the IndicLLM evaluation suite. While these benchmarks play an important role in assessing linguistic coverage, task performance, and general bias across Indian languages, they do not explicitly test for gendered or intersectional harms.[53] Evaluation tasks typically focus on accuracy, fluency, or high-level bias detection, without probing how models reproduce gender stereotypes, make gendered assumptions in advice-giving, or respond to prompts involving women's safety, work, or caregiving roles.[54] This gap is visible, for example, in widely used evaluation benchmarks for Indic language models such as the IndicLLM evaluation suite. While these benchmarks play an important role in assessing linguistic coverage, task performance, and general bias across Indian languages, they do not explicitly test for gendered or intersectional harms.[53] Evaluation tasks typically focus on accuracy, fluency, or high-level bias detection, without probing how models reproduce gender stereotypes, make gendered assumptions in advice-giving, or respond to prompts involving women's safety, work, or caregiving roles.[54] As a result, a model may perform well on Indic language benchmarks while still exhibiting gendered failures in real-world use, particularly in contexts that reflect women's lived experiences. This illustrates how tools that are adapted for Indian languages can nonetheless lack gender specificity, limiting their ability to surface harms that disproportionately affect women and gender minorities.

Beyond these core limitations, reading the harms taxonomy and the safety tools table together also surfaces additional gaps that cut across tool categories and stages of the AI lifecycle. For example, feedback loops between real-world incidents and safety evaluations remain underdeveloped, meaning that documented harms do not consistently inform how safety tools are designed, adapted, or updated over time.[55] Additionally, it is also important to acknowledge the inherent limitations of 'safety tooling' in the first place, particularly when it comes to complex sociotechnical harms such as those discussed in our harms taxonomy. Harms arising from misclassification, biased outputs, or evaluation blind spots may be partially addressed through improved datasets, benchmarks, and testing practices. By contrast, harms linked to platform business models, institutional decision-making, enforcement failures, or barriers to redress often lie beyond the reach of tooling alone. In such cases, safety tools can help make risks visible and legible, but cannot resolve the structural conditions that produce harm.[56]

These gaps point to the need for a combination of interventions: improving the design and contextual relevance of safety tools; embedding safety and inclusion into institutional and regulatory processes; and strengthening capacity to monitor, respond to, and anticipate harms over time. Taken together, mapping such gaps also helps clarify where existing safety tools can meaningfully address gendered risks, and where harms extend beyond the reach of tooling alone, requiring governance measures, accountability mechanisms, and institutional interventions. The recommendations that follow translate these gaps into a set of priority actions across the AI ecosystem, reflecting the range of actors and interventions needed to address gendered harms of AI in India.

4. Recommendations

Inclusivity and fairness already feature prominently in India's emerging approach to AI governance. The AI Governance Guidelines, for instance, identify equity as a guiding principle for the development and deployment of AI systems.[57] At the same time, evidence from real-world deployments suggests that translating these commitments into practice remains uneven, particularly in contexts shaped by existing social, economic, and institutional inequalities.

The analysis in this brief shows that gendered AI harms in India arise across sectors and deployment contexts, and take varied forms. As discussed in the previous sections, some risks are linked to identifiable gaps in data, model design, or evaluation practices, while others are embedded in institutional workflows, platform incentives, and decision-making processes that extend beyond the scope of technical fixes alone. Together, these findings underscore the need to distinguish between harms that can be partially mitigated through improved tooling and those that require broader institutional and governance responses.

Against this backdrop, this section sets out priority intervention points across the AI ecosystem, using gender as a starting analytical lens to identify where action is most urgently needed. Gender is not treated as a standalone category, but as a way to surface how well-documented AI safety risks manifest differently or disproportionately for certain users, and how these risks intersect with factors such as language, geography, caste, disability, and socio-economic status in the Indian context.

Many of the harms identified in this brief disproportionately affect users with limited visibility, weaker bargaining power, or constrained access to grievance mechanisms. This suggests that market incentives alone may be insufficient to drive timely and consistent investment in gender-responsive AI safety, particularly where harms do not immediately translate into commercial or reputational consequences for system developers and deployers. In such contexts, coordinated action by governments, regulators, philanthropic funders, and other ecosystem actors becomes critical.

The recommendations that follow are therefore directed at a range of stakeholders, including policymakers and regulators, AI developers and deployers, philanthropic funders, and civil society organisations. Some focus on governance and regulatory levers, while others are framed as practice-oriented or ecosystem-level measures that can be taken forward alongside formal government action.

Please note that some of the recommendations build on existing policy advisories and governance pathways that are already widely recognised. However, we adapt and contextualise them through a gender lens, focusing on how they can be applied more meaningfully within India's specific social and institutional realities.

The brief also puts forward a set of more novel recommendations. These are presented as initial ideas that go beyond existing guidance and would require further research, stakeholder consultations, and deliberation to assess their feasibility and impact.

Table 3: Recommendations to Address Gendered Harms of AI Systems in India

Key Intervention Levers	Recommendations	Relevant Stakeholders
Data Layer *Building training datasets and evaluation datasets that reflect women's lived realities in India, so AI systems are both trained and tested in ways that surface gendered risks and enable context-aware safety assessments*	**Build Gender-focused Training Datasets** Support the collection and curation of training datasets that meaningfully represent women and other underrepresented communities, particularly for high-stakes AI use cases such as healthcare, agriculture advisory, and financial inclusion, where data gaps can directly translate into denied services, unsafe recommendations, or exclusion from essential public services, leading to a variety of economic or bodily harms for women. **Relevant Action Points:** • **Who the Data Represents:** Prioritise participation from women and affected communities whose needs are often underrepresented in existing datasets, which are typically drawn from historically biased administrative records or online sources and therefore fail to capture the lived experiences of digitally excluded or marginalised user groups. Support use-case-specific data collection approaches that are designed in partnership with domain experts and trusted local organisations, enabling participation from women and affected communities who are unlikely to be reached through conventional, top-down data collection pipelines. For example, women's self-help groups (SHGs) can play a critical role in supporting dataset development for AI systems used in financial inclusion and fintech.[58] • **What Counts as Data:** Ensure datasets go beyond numerical indicators to include qualitative inputs that capture lived experiences and real-world constraints, particularly in sectors where gender norms shape access to services and outcomes. • **How Data is Collected**: Use multi-modal and use-friendly collection methods suited to low-literacy and low-connectivity contexts, including voice notes, assisted data entry, offline tools, and facilitated focus group discussions. These approaches are particularly important for engaging women in rural and informal settings. • **Who Collects and Maintains the Data:** Data collection should be conducted with trusted, on-ground organisations, particularly women-led groups, as prevalent gender norms in many parts of India often limit who women feel comfortable speaking to and what they are willing to share. • **How Often is Data Collected:** Adopt a phased and iterative approach to data development. Representation gaps should be identified and documented upfront, with datasets expanded, corrected, or refined across subsequent phases as new evidence emerges. • **Who Owns and Derives Value from Community Data:** Women and local communities contributing to AI datasets should be enabled to retain control over the collected data, including how it may be shared and used further. Benefits accrued from the commercialisation of these datasets should be shared back with the data contributors. Organisations such as Myna Mahila Foundation and Digital Green illustrate how use-case-specific, gender-responsive data collection can fill critical gaps in existing datasets.[59] By grounding data collection in women's lived experiences within specific domains such as health, livelihoods, and agriculture, these data collection mechanisms help surface patterns of needs, constraints, and decision-making that are typically absent from administrative or online data. AI systems trained on such contextualised data are, in turn, more locally relevant, context-aware, and responsive to women's realities.	**IndiaAI Mission:** Fund, commission, and coordinate dataset creation for priority sectors and sovereign AI efforts. **Proposed Indian AI Safety Institute (AISI):** Provide methodological guidance on how gender-responsive datasets should be constructed. **Philanthropic Actors:** Finance data collection where market incentives are weak, especially for high-risk, low-visibility user groups - particularly women from marginalised castes, rural communities.
	Build Context-aware Evaluation Datasets to test AI Systems As discussed in Section 3, most existing AI benchmarks or evaluation datasets for detecting biases or toxicity in AI-generated content are not designed to reflect the lived realities of women in India and how they actually experience such harms. Many such datasets are also predominantly in the English language, with limited ability to detect gendered bias, abuse, or toxic content generated in Indian languages and mixed-language settings. **Relevant Action Points:** • There is therefore a need to develop context-aware evaluation datasets that capture how harms such as gender bias, exclusion, and abuse are experienced by women and other marginalised groups in practice. These datasets would enable developers and researchers to more meaningfully test AI systems for harmful behaviour in the real-world contexts in which they are deployed. • Building such datasets requires translating lived experiences into structured test cases, creating reusable evaluation artefacts that can be applied across benchmarking, sandbox testing, and other safety assessments. • Such datasets should adequately reflect both India's linguistic diversity as well as the gendered and intersectional risk patterns that shape how such harms manifest on the ground. The Uli dataset is one such example of a context-aware evaluation dataset. Built in collaboration with experts who identify as women or a member of the LGBTQIA+ community in South Asia, the dataset focuses on gendered abuse in three languages, Hindi, Tamil and Indian English, enabling detection of harms often missed by generic or English-centric benchmarks.[60]	**IndiaAI Mission:** As the central public funder and coordinator for AI infrastructure, IndiaAI is best placed to commission and support the development of India-specific evaluation datasets, particularly for high-stakes use cases and Indic language contexts. **Applied Research Institutions and Community-based Organisations:** Research centres and community-based organisations play a critical role in translating lived experiences into structured, reusable evaluation artefacts.

Key Intervention Levers	Recommendations	Relevant Stakeholders
	Bias and Representation Disclosures as a Prerequisite for Training Data Releases Disclosures as a Prerequisite for Training Data Releases Many harms experienced by women and other marginalised groups are rooted in gaps in how training data is sourced, documented, and disclosed. In practice, developers and deployers often lack visibility into whether datasets adequately represent women's lived realities, linguistic diversity, or intersectional identities, making it difficult to anticipate downstream risks or assess the suitability of datasets for specific deployment contexts.[61] **Relevant Action Points:** • There is a need to encourage the adoption of standardised data documentation, such as data cards or dataset disclosures, that explicitly surface representation gaps and known limitations in training data. • These disclosures should go beyond high-level provenance details to include information on demographic coverage, language distribution, annotation practices, and known sources of bias, with specific attention to gender and intersectional risks. • Such disclosures are particularly important for training datasets hosted on government-supported or publicly funded platforms, such as AI Kosh, where datasets are likely to be reused across multiple applications and sectors. Standardised representation disclosures would enable downstream developers, auditors, and public sector procurers to make more informed decisions about dataset reuse, adaptation, or the need for additional safeguards.	**MeitY/IndiaAI Mission:** As the steward of government-backed AI infrastructure and platforms such as AI Kosh, MeitY and the IndiaAI Mission are best positioned to require standardised data cards as a condition for hosting or releasing training datasets.
Model Layer *Ensuring that general-purpose and foundation models developed for Indian contexts are designed, tested, and documented with explicit gender- and context-specific safety safeguards, rather than relying on generic assumptions about acceptable model behaviour and safety guardrails.*	**Gender-specific Safety Screenings as a Prerequisite for Indic AI Models** Require general-purpose AI models developed for or deployed in Indian contexts to undergo baseline gender-safety screening through adversarial testing and red teaming in sandboxed environments, with a specific focus on Indic languages and mixed-language use. **This screening should test whether the model:** • Generates gendered slurs, sexualised abuse, or derogatory stereotypes in Indian languages. • Produces biased or exclusionary assumptions in advice-giving or informational responses (for example, defaulting to male perspectives or reinforcing gender roles). • Performs differently across different demographic or socio-economic groups, in order to identify disproportionate error rates, exclusion, or harm that may not be visible in aggregate metrics. • Fails to recognise, respect, or appropriately handle gender identity and pronouns in neutral or everyday contexts. **Mandate Gender-Specific Safety Specifications and Disclosure** Require developers to document and publish model-level gender safety specifications (similar in intent to OpenAI's Model Spec[62]) but adapted to Indian linguistic and social contexts. To support such disclosures, provide standardised templates for (a) model cards, (b) data cards, and (c) evaluation disclosure reports (what was tested, guardrails applied, observed risks, known limitations), with mandatory fields for gender and social biases. **These disclosures should clearly specify:** • What constitutes *undesirable outputs* for the model, in the context of gender or other social biases in the Indian context. • Which gender-related behaviours the model is explicitly trained to avoid • Known limitations or unresolved risks related to gender harms or harms affecting other marginalised or protected groups • Coverage gaps in their training datasets Determining what constitutes appropriate or unsafe model behaviour is not a neutral or purely technical exercise. Developers should therefore define gender-related safety boundaries through consultation with feminist scholars, gender experts, and relevant domain specialists, grounded in Indian social and linguistic contexts. As recent critiques by model developers themselves have highlighted, safety specifications that rely on abstract or imported norms risk mischaracterising harm in non-Western settings.[63] Indian general-purpose model developers should address this from the outset by grounding safety fine-tuning in context-aware evaluation datasets that reflect how gendered harms are articulated and experienced locally. This ensures that decisions about acceptable behaviour, failure modes, and known limitations are shaped by real-world use and community realities, rather than generic assumptions. **Post-Deployment Obligations** • Make continued government support contingent on post-deployment monitoring/grievance redressal offered. • Where technically feasible and contextually appropriate, mandate watermarks or provenance mechanisms for content generated by these models.	**General-purpose Model Developers (Publicly Funded or Commissioned):** Developers of foundation and general-purpose models intended for Indian contexts are responsible for ensuring that baseline safety behaviours are robust across languages and social settings, prior to downstream adaptation or deployment. **IndiaAI/AISI (Standard-setting role):** IndiaAI and the proposed AISI can support consistency by issuing reference templates and minimum expectations for gender-safety testing and disclosure, reducing ambiguity for developers and evaluators.

Key Intervention Levers	Recommendations	Relevant Stakeholders
AI Evaluations Layer *Embedding evidence-first approaches into AI development and deployment, especially in high-stakes use cases, so risks are identified and addressed based on how systems actually behave in practice.*	**Provide Safety & Inclusivity Sandboxes for High-stakes Models & Applications** Regulatory sandboxes offer a unique avenue to foreground evidence-first approaches in AI development and deployment. By offering a controlled, assistive environment, they help spur AI innovation and experimentation, while ensuring alignment with ethical, safety, and policy frameworks.[64] This is particularly valuable for emerging technologies such as foundation models, where system behaviour and downstream impacts remain insufficiently understood and require empirical testing prior to wider deployment. Set up an "AI Safety & Inclusivity Sandbox" that will allow AI developers to test and refine their models and/or applications in a controlled environment, with a specific focus on risks related to misuse, bias, and discrimination. These sandboxes can also offer developers the necessary compute infrastructure needed for robust testing and iterations. A prominent example of such inclusivity-focused sandboxes is AI+ Alliance's *'Feminist AI Innovation Sandbox' that aims to provide a secure, human rights-focused digital environment with open-source AI models and development tools to organisations looking to develop and refine gender-responsive AI applications.*[65] In the Indian context, such sandboxes can be piloted as a standalone initiative or can be an integral part of emerging sectoral sandboxes (such as the AI Innovation Sandbox proposed by the Reserve Bank of India for the fintech sector), with technical and compute support provided through MeitY, the India AI Mission, and the upcoming Indian AISI.[66] As emphasised in the IndiaAI Governance Guidelines, such AI sandboxes should produce evidence with published details of what was tested, guardrails applied, and the risks observed.[67]	**MeitY/IndiaAI Mission, Sectoral Regulators:** These actors can help set up sandboxes by offering infrastructure, coordination, and support for testing AI systems before they are deployed.
	Supporting Community-based Evaluations of AI Applications (particularly for high-stakes sectors) Sandbox-based evaluations of AI models and applications are typically conducted in controlled, time-bound, and limited-scope environments. While some sandboxes may involve user groups or sectoral stakeholders, they often do not entail sustained collaboration with local, on-ground communities, particularly those most likely to experience downstream harms. Furthermore, if adequate care is not taken, even the datasets used in such testing environments may suffer from a lack of diverse representation of different gender identities or intersectional demographics. Additionally, evaluating complex social biases related to gender, class, or caste requires deep and sustained engagement with real-world contexts. Such engagement is often limited within sandboxes, which are inherently small-scale, controlled, and time-bound, making it challenging to draw reliable insights into how AI systems will behave across diverse user groups and social settings. Therefore, it is important to support and promote community-based evaluations in addition to sandbox testing. Community-based evaluations involve testing AI models and applications in close coordination with affected communities, civil society organisations, and domain practitioners. For example, teams conducting red-teaming exercises should ensure adequate representation of women when aiming to surface, characterise, and document real-world harmful behaviours, such as stereotypes or bias against women. Such evaluations help surface forms of harm, exclusion, or usability barriers that are often missed in narrowly technical testing, including issues related to accessibility, cultural relevance, and trust. [68] Such community-focused evaluations can be made mandatory for applications that directly affect bodily integrity, access to rights, or essential services, such as AI tools intended for frontline healthcare workers or sexual and reproductive health use cases. *Evidence generated through sandbox and community-based evaluations can feed into the development of gender-sensitive benchmarks and context-aware evaluation datasets.*	**Philanthropic Funders:** Philanthropic actors play a critical role in supporting community-based evaluations by funding sustained engagement, compensating participants, and enabling work that is unlikely to be commercially viable or state-funded in early stages. **Public Funders and Regulators:** Public institutions can strengthen the impact of such evaluations by recognising their findings as valid evidence for procurement decisions, risk assessments, or further regulatory action.
	AI Incident Database Support the development and maintenance of a centralised, crowdsourced database of emerging instances of AI misuse, bias, discrimination, and exclusion. This can be built or pooled using various sources, including, but not limited to: • Self-reporting by affected individuals through grievance redress platforms. • Self-reporting by AI-developing and deploying organisations with evidence from sandbox or community-based evaluations, pilots, user logs, or impact assessments. • Reporting by civil society organisations or philanthropic actors operating helplines such as 'Meri Trustline' (Rati). • Web scraping of emerging reports in news outlets and on social media platforms.	**MeitY/IndiaAI Mission (Central steward):** A central public authority is best placed to host and govern a national AI incident reporting database. This includes setting reporting standards, ensuring interoperability across sectors, safeguarding sensitive data, and enabling aggregated analysis of gendered and other harms.

Key Intervention Levers	Recommendations	Relevant Stakeholders
AI Safety Tooling Layer *Ensuring AI safety tools used in India are able to detect and assess risks as they actually arise in local languages, social contexts, and use cases, rather than relying on tools designed for English-language or Global North settings.*	**Contextual Scaling of Existing Safety Tools** Evaluate the performance and efficacy of existing safety tools when used for diagnosing and mitigating risks in AI systems developed/customised and deployed in local contexts within India. For example, more research needs to be undertaken to assess how existing red-teaming tools/prompts (predominantly English-language-centred) for stress-testing generative AI systems perform in Indic language and mixed language contexts and what steps need to be taken to ensure their effectiveness for such languages/local contexts. Contextualised prompts for red teaming can also be developed using evidence available from incident reporting databases.	**IndiaAI Mission/proposed AI Safety Institute:** Public institutions can play a coordinating and standard-setting role by supporting research on the effectiveness of existing safety tools in Indian contexts, issuing guidance on contextual adaptation, and enabling shared repositories of Indic-language and mixed-language evaluation artefacts. **Civil Society or Academic Research Organisations:** Research institutions and civil society actors are well placed to develop and validate contextualised red-teaming prompts and test cases, drawing on evidence from incident reporting databases and documented user experiences. These actors can help translate real-world harms into evaluation inputs that improve the relevance and robustness of safety testing.
	Develop New AI Safety Tools for Indian Contexts Support the development of new, India-specific AI safety tools to surface and mitigate forms of harm that are poorly captured by existing safety approaches, particularly harms that are linguistic, contextual, or relational rather than purely semantic. These harms can include indirect bias in advice systems or context-specific abuse expressed through local languages and social norms. Supporting innovation in this area would allow safety assessments to move beyond surface-level toxicity detection toward identifying how AI systems reinforce or exacerbate gender inequalities in real-world use. Such work can also include the co-development of LLM guardrails with a broader range of stakeholders, including end-user communities as well as policymakers.	**Applied research institutions and technical universities:** Research institutions are well placed to lead the design and development of new safety tools that address gendered harms specific to linguistic, cultural, and social contexts in India. **IndiaAI Mission:** The Mission can play a catalytic role by funding and commissioning the development of India-specific safety tools, supporting pilot deployments, and integrating successful tools into national benchmarking, sandboxing, or evaluation initiatives.
	Build an AI Safety Leaderboard (with Bias & Discrimination as Indexing Parameters) Create an India-specific AI safety leaderboard that aggregates and publicly displays results from bias, discrimination, and safety evaluations of frontier and foundation models developed for or deployed in Indian contexts. The leaderboard would improve transparency and comparability across models by making safety performance visible over time, and by surfacing findings from benchmarks, sandboxes, and independent evaluations, rather than relying on isolated or ad-hoc disclosures.[69]	**IndiaAI Mission/proposed AI Safety Institute:** To steward the leaderboard, define indexing parameters, and ensure methodological consistency and neutrality.

Key Intervention Levers	Recommendations	Relevant Stakeholders
Regulatory Layer *Ensuring that existing laws, rules, and regulatory guidance meaningfully address AI-related harms as they arise in practice, including clarifying responsibilities across the AI value chain, strengthening response timelines for high-risk harms, and enabling effective redress for affected users, particularly in cases of gendered abuse and misuse.*	**Regulatory Gap Mapping for AI-Enabled Discrimination Across Sectors** While much of the regulatory focus on AI and gender has centred on harms such as deepfakes and online abuse, several of the harms identified in this brief — including economic exclusion, discriminatory credit scoring, opaque welfare eligibility systems, and algorithmic hiring biases — have not received much traction within the regulatory governance discourse in India. It also remains unclear to what extent existing legal and regulatory frameworks, especially within specific critical sectors such as finance and healthcare, adequately address AI-enabled forms of discrimination. There is thus a need to explore whether current sectoral regulations are equipped to anticipate, detect, and remedy AI-driven harms, particularly those that disproportionately affect women and marginalised communities**.**	MeitY/IndiaAI Mission, Sectoral Regulators
	Relevant Action Points: • Commission a structured regulatory review, led by an appropriate inter-ministerial or expert body, to assess how existing sectoral laws and regulatory frameworks apply to AI-driven decision-making systems. • Direct sectoral regulators to map current and emerging uses of AI within their domains, particularly where automated systems affect access to credit, employment, public benefits, healthcare, or other essential services. ◦ Require sectoral regulators, as part of this process, to undertake gender-sensitive risk assessments examining potential discrimination, exclusion, opacity, and due-process concerns arising from such systems. Based on these findings, identify regulatory gaps, if any, and recommend proportionate responses - which may include clarifications to existing rules, supervisory guidance, enhanced transparency requirements, stronger enforcement mechanisms, or, where demonstrably necessary, the consideration of new legal instruments.	
	Moving Towards Proactive Regulation of AI-enabled Gender Abuse Public visibility into AI-enabled gender abuse is often triggered only after high-profile incidents, by which point harm has already occurred and spread at scale. Incident-driven scrutiny limits the ability of regulators to assess whether safeguards are functioning in advance, identify emerging misuse patterns, or intervene before risks escalate. Additionally, a significant share of gendered deepfake abuse originates upstream, through readily available image manipulation and nudification tools that make it easy to generate harmful content before it ever reaches platforms. Existing governance frameworks focus largely on (downstream) intermediaries such as social media platforms, leaving these tools insufficiently addressed despite their central role in enabling abuse at scale. **Relevant Action Points:** • Prohibit developers from making available certain high-risk features where evidence of harm already exists, such as nudification and clothes-swapping. • Require developers of high-risk synthetic media tools to conduct contextualised jail-breaking or adversarial testing exercises before these tools are made available to the public. Reports containing the outcomes of these exercises and actions taken subsequently should also be made available publicly alongside the tool. ◦ Ensure that these exercises evaluate for resilience in low-resource and non-English contexts, where jailbreaks and circumvention of safety controls are often easier due to data sparsity, linguistic variation, and weaker moderation coverage. • Require online platforms where synthetic media is hosted or circulated, to publish regular and standardised transparency reports on AI-generated abuse. These reports should include the volume of harmful content reported, response times, and actions taken, broken down by type of harm where feasible. Regular reporting would allow regulators to track trends, assess whether safeguards are working, and intervene early, rather than responding only after harm has already occurred.	
	Clarify Responsibilities Across Key Actors in the AI Value Chain Current AI governance arrangements do not clearly specify how responsibilities are distributed across key actors in the AI value chain. For example, while the intermediary guidelines regulate platform response to AI-enabled gender abuse, they do not address the responsibilities of actors such as model providers, API providers, application developers, platforms, and users in relation to AI-enabled gender abuse. This lack of role clarity creates enforcement gaps and allows responsibility for prevention and response to be diffused across actors. Issue role-based guidance/standard operating protocols that clearly delineate safety, disclosure, monitoring, and cooperation obligations for different actors across the AI value chain, particularly for high-risk synthetic media capabilities.	

Key Intervention Levers	Recommendations	Relevant Stakeholders
Regulatory Layer *Ensuring that existing laws, rules, and regulatory guidance meaningfully address AI-related harms as they arise in practice, including clarifying responsibilities across the AI value chain, strengthening response timelines for high-risk harms, and enabling effective redress for affected users, particularly in cases of gendered abuse and misuse.*	**Introduce Risk Tiers for Tackling Ai-Generated Deepfakes** Current legal frameworks tend to treat all deepfakes through similar notice-and-action protocols, despite wide variation in severity and impact. For example, sexualised deepfakes/NCII can cause acute and long-lasting harm to women, including reputational damage, psychological distress, and withdrawal from public and professional life. Treating such content on par with benign/trivial synthetic media delays meaningful/effective intervention for cases where it's most urgently required. Introduce a risk-tiered framework for synthetic media that explicitly identifies high-risk categories such as sexualised deepfakes, non-consensual intimate imagery, and child sexual abuse material and attaches stricter response obligations (from relevant platforms) and faster timelines, and higher escalation thresholds to these categories.	MeitY/IndiaAI Mission, Sectoral Regulators
	Establish a Single-Window Reporting and Escalation Mechanism for AI-Enabled Image Abuse While legal remedies and cybercrime reporting channels exist, survivors of gendered online abuse are often required to report the same harm separately to platforms and law enforcement, resulting in delays, repeated exposure/continued circulation.[71] We therefore recommend establishing a single-window pathway through which one report of sexualised deepfakes or non-consensual intimate imagery can trigger coordinated platform action and accelerated takedowns, without requiring survivors to navigate multiple systems. This single-window mechanism can be implemented as a dedicated, fast-track workflow within the existing National Cyber Crime Reporting Portal, rather than as a separate helpline or platform.[72]	
Capacity Layer *Strengthening the ability of regulators, enforcement agencies, developers, and civil society organisations to identify, assess, and respond to AI-related risks in real-world contexts, including building long-term anticipatory and forecasting capacities to detect emerging and structural harms that may not surface through short-term evaluations.*	**Strengthen Institutional Capacity for AI Safety Enforcement and Redress** Build the capacity of regulators, enforcement agencies, judicial actors, and civil society organisations to understand and respond to AI safety risks in practice. Strengthen institutional capacity within enforcement and oversight bodies to handle AI-enabled harms, including evidence handling, coordination with platforms, and timely response mechanisms.[73] • Conventional evidence-collection models often prioritise technical completeness and prosecutorial utility over the safety, agency, and well-being of the person experiencing harm. This approach can inadvertently retraumatise survivors, escalate risk, or undermine autonomy. Develop survivor-centric methods of digital evidence collection and handling which prioritise the safety, agency, and consent of the person experiencing violence, rather than just the technical or legal aspects of a case.[74] These methods should include trauma-informed documentation practices which focus on minimising re-traumatisation and integrate documentation with providing necessary emotional support to survivors.	MeitY/IndiaAI Mission, Relevant Training and Judicial Academies
	Build Capacity for Anticipatory Governance and Long-term Risk Assessment Many gendered harms linked to AI do not emerge as immediate safety failures, but as structural shifts over time. For example, as detailed in Table 1, reduced female labour force participation as a result of AI-led automation is likely to unfold as a risk in the coming years, as AI gets integrated within industries and workflows. To better anticipate and respond to such risks, there is a need to strengthen the capacity of policymakers and civil society organisations to undertake mixed-methods forecasting and horizon-scanning on the longer-term and cumulative impacts of AI deployment. These forms of harm are difficult to detect through model-level testing or short-term evaluations alone, yet can become deeply entrenched once they materialise.	MeitY/IndiaAI Mission, NITI Aayog, Sectoral Ministries

5. Annexure: Methodology

This brief is based on a mixed-method qualitative approach combining secondary research and Key Informant Interviews (KIIs) to develop a contextualised understanding of gendered AI safety risks in India. We provide a detailed breakdown of our approach below:

- **Desk research for conceptual mapping:**
    - We conducted a structured review of scholarship at the intersection of AI safety, gender studies, and feminist technology research. Through this review, our aim was:
        - To examine established AI harms frameworks and understand how these frameworks have theorised AI safety risks in global literature.
        - To analyse how such frameworks may adapt differently or disproportionately for women and marginalised gender groups, particularly in the Indian context.
- **Desk research for tool analysis:**
    - Concurrently, we analysed existing AI safety tools and evaluation mechanisms to assess and identify gaps related to:
        - Their adoption in India;
        - Coverage of harms specific to India's diverse socio-cultural contexts; &
        - The extent to which they incorporate gender-related considerations.
- **Semi-structured Interviews with practitioners:**
    - To understand prevalent concerns related to gender and AI safety in India, we conducted 7 semi-structured KIIs with AI developers, technical researchers, and feminist scholars. Through these interviews, we assessed:
        - The practical challenges involved in operationalising gender-sensitive AI safety across AI development and deployment processes in India;
        - Institutional, technical, and capacity constraints within the Indian ecosystem; &
        - Opportunities for more contextually grounded evaluation and governance mechanisms.

Insights from these interviews were used to refine the harms taxonomy, validate identified gaps in safety tool adoption and implementation in India, and inform the recommendations presented in this brief.

Endnotes

[1] Senior Research Manager, Digital Futures Lab

[2] Research Associate, Digital Futures Lab

[3] Founder & Director, Digital Futures Lab; Associate Director, Strategy & Partnerships, GxD hub.

[4] '22 Languages, Digitally Reimagined', https://www.pib.gov.in/www.pib.gov.in/Pressreleaseshare.asbased onpx?PRID=2182427.

[5] 'AI Has Potential to Make India's DPI Significantly More Efficient: Abhishek Singh', IndiaAI, https://indiaai.gov.in/article/ai-has-potential-to-make-india-s-dpi-significantly-more-efficient-abhishek-singh.

[6] Peter Slattery et al., 'The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence', arXiv:2408.12622, preprint, arXiv, 10 April 2025, https://doi.org/10.48550/arXiv.2408.12622.

[7] P. R. Biju and O. Gayathri, 'Structural Oppression and AI: A Systematic Review of Data Policy Frameworks in India', Technological Forecasting and Social Change 223 (February 2026): 124415, https://doi.org/10.1016/j.techfore.2025.124415; Urvashi Aneja et al., From Code to Consequence: Interrogating Gender Biases in LLMs within the Indian Context (2024), https://digitalfutureslab.notion.site/From-Code-to-Consequence-Interrogating-Gender-Biases-in-LLMs-within-the-Indian-Context-1069c92254ab80e4bdfce1f2b004a42f.

[8] MeitY, India AI Governance Guidelines (2025), https://static.pib.gov.in/WriteReadData/specificdocs/documents/2025/nov/doc2025115685601.pdf.

[9] Dharish David et al., 'Algorithmic Bias and Discrimination in India: A Looming Crisis', Journal of Development Policy and Practice 11, no. 1 (2026): 81–104, https://doi.org/10.1177/24551333251343358.

[10] Anita Gurumurthy and Nandini Chami, Digital Technologies and Gender Justice in India -An Analysis of Key Policy and Programming Concerns Input to the High Level Committee on the Status of Women in India (IT for Change, 2014), https://www.researchgate.net/publication/344158173_Digital_Technologies_and_Gender_Justice_in_India_-An_analysis_of_key_policy_and_programming_concerns_Input_to_the_High_Level_Committee_on_the_Status_of_Women_in_India.

[11] Biju and Gayathri, 'Structural Oppression and AI'.

[12] Jameela Sahiba, 'India's AI Safety Institute: Key Considerations for a Critical Initiative', Tech Policy Press, 23 October 2024, https://www.techpolicy.press/indias-ai-safety-institute-key-considerations-for-a-critical-initiative/.

[13] Genevieve Smith and Ishita Rustagi, 'When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity (SSIR)', Stanford Social Innovation Review, 31 March 2021, https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity; Darrell M. West, 'AI Poses Disproportionate Risks to Women', Brookings, 20 November 2023, https://www.brookings.edu/articles/ai-poses-disproportionate-risks-to-women/; Rumman Chowdhury et al., Tackling Gender Bias and Harms in Artificial Intelligence (AI) (UNESCO, 2025), https://www.unesco.org/ethics-ai/en/articles/tackling-gender-bias-and-harms-artificial-intelligence-ai; Luhang Sun et al., 'Smiling Women Pitching Down: Auditing Representational and Presentational Gender Biases in Image Generative AI', Journal of Computer-Mediated Communication 29, no. 1 (2023): zmad045, https://doi.org/10.1093/jcmc/zmad045.

[14] Aisha Down, '"The Chilling Effect": How Fear of "Nudify" Apps and AI Deepfakes Is Keeping Indian Women off the Internet', Global Development, The Guardian, 5 November 2025, https://www.theguardian.com/global-development/2025/nov/05/india-women-ai-deepfakes-internet-social-media-artificial-intelligence-nudify-extortion-abuse; Madhavi Ravikumar, 'AI Deepfakes: Women, Misogyny, and the Digital Threat in India', Frontline, 28 November 2025, https://frontline.thehindu.com/social-issues/gender/ai-deepfake-abuse-misogyny-crisis-india/article70335218.ece.

[15] This initial harms taxonomy has been developed through secondary research, drawing on existing harms frameworks and adapting them to the Indian context. It provides a structured starting point for analysis and can be further refined, where needed, through additional research and stakeholder consultations to enhance its applicability in real-world settings in India.

[16] Mia Hoffman and Heather Frase, Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework (Center for Security and Emerging Technology, 2023), https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/; Renee Shelby et al., 'Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction', arXiv:2210.05791, preprint, arXiv, 19 July 2023, https://doi.org/10.48550/arXiv.2210.05791.

[17] Aisha Down, '"The Chilling Effect": How Fear of "Nudify" Apps and AI Deepfakes Is Keeping Indian Women off the Internet', Global Development, The Guardian, 5 November 2025, https://www.theguardian.com/global-development/2025/nov/05/india-women-ai-deepfakes-internet-social-media-artificial-intelligence-nudify-extortion-abuse.

[18] Gianluca Mauro and Hilke Schellmann, '"There Is No Standard": Investigation Finds AI Algorithms Objectify Women's Bodies', Technology, The Guardian, 8 February 2023, https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies.

[19] Abhiramy S M, Algorithmic Blindness and Gender Sensitivities in Gig Work (SPRF, 2025), https://sprf.in/algorithmic-blindness-and-gender-sensitivities-in-gig-work/.

[20] Genevieve Smith, 'Mindsets and Management: AI and Gender (In)Equitable Access to Finance', arXiv:2504.07312, preprint, arXiv, 17 July 2025, https://doi.org/10.48550/arXiv.2504.07312.

[21] Anna Desmarais, 'Women Three Times More Vulnerable to Having Job Taken by AI than Men, New Report Warns', Euro News, 24 May 2025, https://www.euronews.com/next/2025/05/24/womens-jobs-three-times-more-vulnerable-to-being-taken-by-ai-than-mens-new-report-warns.

[22] Pihu Yadav, 'Babydoll Archi Is 2025's Revenge Porn and She Never Even Existed', CNBCTV18, 23 July 2025, https://www.cnbctv18.com/technology/babydoll-archi-is-2025s-revenge-porn-and-she-never-even-existed-19642079.htm.

[23] Siddharth Pillai et al., 'Make It Real - Mapping AI-Facilitated Gendered Harm', 2025, https://tattle.co.in/blog/make-it-real/.

[24] Pillai et al., 'Make It Real - Mapping AI-Facilitated Gendered Harm'.

[25] Daniel van Niekerk et al., Challenging Systematic Prejudices - An Investigation into Bias Against Women and Girls in Large Language Models (UNESCO, 2024), https://ocivvlhqiwtxvnycqnia.supabase.co/storage/v1/object/public/new-kerko/2129771/S9ZX7IXX/RCKFTSRA/file.pdf.

[26] Sourojit Ghosh and Aylin Caliskan, 'ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five Other Low-Resource Languages', Proceedings of the 2023 ACM Conference on International Computing Education Research V.1, 7 August 2023, 397–415, https://doi.org/10.1145/3568813.3600120.

[27] Morgan Klaus Scheuerman et al., 'How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services', Proc. ACM Hum.-Comput. Interact. 3, no. CSCW (2019): 144:1-144:33, https://doi.org/10.1145/3359246.

[28] Gaurav Jain and Smriti Parsheera, 'Cinderella's Shoe Won't Fit Soundarya: An Audit of Facial Processing Tools on Indian Faces', arXiv:2112.09326, preprint, arXiv, 17 December 2021, https://doi.org/10.48550/arXiv.2112.09326.

[29] Yan Tao et al., 'Cultural Bias and Cultural Alignment of Large Language Models', PNAS Nexus 3, no. 9 (2024): pgae346, https://doi.org/10.1093/pnasnexus/pgae346.

[30] 'Do Generative AI Models Output Harm While Representing Non-Western Cultures: Evidence from A Community-Centered Approach', accessed 24 January 2026, https://arxiv.org/html/2407.14779v1#bib.bib7.

[31] Kaitlyn Regehr et al., 'Normalizing Toxicity: The Role of Recommender Algorithms for Young People's Mental Health and Social Wellbeing', Frontiers in Psychology 16 (n.d.): 1523649, https://doi.org/10.3389/fpsyg.2025.1523649.

[32] Paul Morris and Roxanne Khan, Digital and Algorithmic Normalisation of Violence Against Women and Girls (OnEvidence and HARM network, 2025); Esli Chan, 'Toxic Monetization and Amplification in Digital Platform Design: The Case of Online Gendered Hate', in Hate Crime Perpetrators: New Perspectives from Theory, Research and Practice, Volume II: Developing Responses to Hate in Online and Offline Locations, ed. Jon Garland et al. (Springer Nature Switzerland, 2025), https://doi.org/10.1007/978-3-031-92670-9_7.

[33] UN Women, 'Tipping Point: The Chilling Escalation of Violence against Women in the Public Sphere', 2025, https://www.unwomen.org/en/digital-library/publications/2025/12/tipping-point-the-chilling-escalation-of-violence-against-women-in-the-public-sphere-in-the-age-of-ai.

[34] Tapasya et al., 'How an Algorithm Denied Food to Thousands of Poor in India's Telangana', Al Jazeera, 24 January 2024, https://www.aljazeera.com/economy/2024/1/24/how-an-algorithm-denied-food-to-thousands-of-poor-in-indias-telangana.

[35] Timnit Gebru et al., 'Datasheets for Datasets', Commun. ACM 64, no. 12 (2021): 86–92, https://doi.org/10.1145/3458723.

[36] Ran Zmigrod et al., 'Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology', arXiv:1906.04571, preprint, arXiv, 27 May 2020, https://doi.org/10.48550/arXiv.1906.04571.

[37] Samuel Gehman et al., 'RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models', arXiv:2009.11462, preprint, arXiv, 25 September 2020, https://doi.org/10.48550/arXiv.2009.11462; 'OpenWebText2', accessed 12 December 2025, https://openwebtext2.readthedocs.io/en/latest/.

[38] Mohammed Safi Ur Rahman Khan et al., 'IndicLLMSuite: A Blueprint for Creating Pre-Training and Fine-Tuning Datasets for Indian Languages', Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, 15831–79, https://doi.org/10.18653/v1/2024.acl-long.843.

[39] Anoop Kunchukuttan et al., 'AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages', arXiv:2005.00085, preprint, arXiv, 30 April 2020, https://doi.org/10.48550/arXiv.2005.00085.

[40] IndiBias: Benchmark measuring multiple social bias axes (gender, caste, religion, region) in English and Hindi; Nihar Ranjan Sahoo et al., 'IndiBias: A Benchmark Dataset to Measure Social Biases in Language Models for Indian Context', arXiv:2403.20147, preprint, arXiv, 3 April 2024, https://doi.org/10.48550/arXiv.2403.20147.

[41] BharatBBQ: Multilingual QA bias benchmark across 8 Indian languages & 13 social categories including gender; Aditya Tomar et al., 'BharatBBQ: A Multilingual Bias Benchmark for Question Answering in the Indian Context', arXiv:2508.07090, preprint, arXiv, 9 August 2025, https://doi.org/10.48550/arXiv.2508.07090.

[42] IndiCASA: A bias benchmarking dataset and evaluation framework comprising 2,575 human-validated sentences spanning five demographic axes, including gender; Santhosh G. S et al., 'IndiCASA: A Dataset and Bias Evaluation Framework in LLMs Using Contrastive Embedding Similarity in the Indian Context', arXiv:2510.02742, preprint, arXiv, 3 October 2025, https://doi.org/10.48550/arXiv.2510.02742.

[43] OTSC-Hindi and Wino-MT Hindi: Two evaluation sets for gender bias evaluation of NMT models; Pushpdeep Singh, 'Gender Inflected or Bias Inflicted: On Using Grammatical Gender Cues for Bias Evaluation in Machine Translation', Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop, 2023, 17–23, https://doi.org/10.18653/v1/2023.ijcnlp-srw.3.

[44] Uli: Task-specific benchmark dataset on gendered abuse in Hindi, Tamil and Indian English, comprising of tweets annotated along three questions pertaining to the experience of gender abuse, by experts who identify as women or a member of the LGBTQIA community in South Asia; Arnav Arora et al., 'The Uli Dataset: An Exercise in Experience Led Annotation of oGBV', arXiv:2311.09086, preprint, arXiv, 24 June 2024, https://doi.org/10.48550/arXiv.2311.09086.

[45] Rumman Chowdhury et al., Tackling Gender Bias and Harms in Artificial Intelligence (AI) (UNESCO, 2025),https://www.unesco.org/ethics-ai/en/articles/tackling-gender-bias-and-harms-artificial-intelligence-ai.

[46] Guneet Singh, 'Red Teaming LLMs with India's Cultural Complexity', Deccan AI, 16 October 2024, https://www.deccan.ai/blogs/red-teaming.

[47] Long Ouyang et al., 'Training Language Models to Follow Instructions with Human Feedback', arXiv:2203.02155, preprint, arXiv, 4 March 2022, https://doi.org/10.48550/arXiv.2203.02155; Nisan Stiennon et al., 'Learning to Summarize from Human Feedback', arXiv:2009.01325, preprint, arXiv, 15 February 2022, https://doi.org/10.48550/arXiv.2009.01325; Deep Ganguli et al., 'Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned', arXiv:2209.07858, preprint, arXiv, 22 November 2022, https://doi.org/10.48550/arXiv.2209.07858.

[48] Margaret Mitchell et al., 'Model Cards for Model Reporting', Proceedings of the Conference on Fairness, Accountability, and Transparency, 29 January 2019, 220–29, https://doi.org/10.1145/3287560.3287596.

[49] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, 21 January 2018, 77–91, https://proceedings.mlr.press/v81/buolamwini18a.html.

[50] Treasury Board of Canada Secretariat, 'Algorithmic Impact Assessment Tool', 30 May 2024, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

[51] Robert Gorwa et al., 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', Big Data & Society 7, no. 1 (2020): 2053951719897945, https://doi.org/10.1177/2053951719897945.

[52] Nithya Sambasivan et al., '"Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI', Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (New York, NY, USA), CHI '21, 7 May 2021, 1–15, https://doi.org/10.1145/3411764.3445518.

[53] Mohammed Safi Ur Rahman Khan et al., 'IndicLLMSuite: A Blueprint for Creating Pre-Training and Fine-Tuning Datasets for Indian Languages', Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, 15831–79, https://doi.org/10.18653/v1/2024.acl-long.843.

[54] Hadas Orgad and Yonatan Belinkov, 'Choose Your Lenses: Flaws in Gender Bias Evaluation', arXiv:2210.11471, preprint, arXiv, 20 October 2022, https://doi.org/10.48550/arXiv.2210.11471.

[55] Lujain Ibrahim et al., 'Towards Interactive Evaluations for Interaction Harms in Human-AI Systems', arXiv:2405.10632, preprint, arXiv, 30 July 2025, https://doi.org/10.48550/arXiv.2405.10632.

[56] Thilo Hagendorff, 'Blind Spots in AI Ethics', AI and Ethics 2, no. 4 (2022): 851–67, https://doi.org/10.1007/s43681-021-00122-8.

[57] MeitY, India AI Governance Guidelines.

[58] Astha Kapoor and Bapu Vaitla, ‘Data Co-Ops: How Cooperative Structures Can Support Women’s Empowerment’, Brookings, 2 November 2022, https://www.brookings.edu/articles/data-co-ops-how-cooperative-structures-can-support-womens-empowerment/.

[59] Roshini Deva et al., ‘“Kya Family Planning after Marriage Hoti Hai?”: Integrating Cultural Sensitivity in an LLM Chatbot for Reproductive Health’, Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (New York, NY, USA), CHI ’25, 25 April 2025, 1–23, https://doi.org/10.1145/3706598.3713362; Digital Green, Engendering Agriculture: Improving Women Farmer’s Access to Agricultural Information (Digital Green, 2023).

[60] Uli: Task-specific benchmark dataset on gendered abuse in Hindi, Tamil and Indian English, comprising of tweets annotated along three questions pertaining to the experience of gender abuse, by experts who identify as women or a member of the LGBTQIA community in South Asia; Arnav Arora et al., ‘The Uli Dataset: An Exercise in Experience Led Annotation of oGBV’, arXiv:2311.09086, preprint, arXiv, 24 June 2024, https://doi.org/10.48550/arXiv.2311.09086.

[61] Ahmed Umar Otokiti et al., ‘Gender and Racial Bias Unveiled: Clinical Artificial Intelligence (AI) and Machine Learning (ML) Algorithms Are Fanning the Flames of Inequity’, Oxford Open Digital Health 3 (2025): oqaf027members, https://doi.org/10.1093/oodh/oqaf027.

[62] OpenAI, ‘OpenAI Safety Practices’, 21 May 2024, https://openai.com/index/openai-safety-update/.

[63] OpenAI, ‘Introducing IndQA’, OpenAI, 3 November 2025, https://openai.com/index/introducing-indqa/.

[64] ‘AI Sandboxes for the Intelligent Age’, World Economic Forum, August 2025, https://www.weforum.org/publications/shaping-the-ai-sandbox-ecosystem-for-the-intelligent-age/.

[65] ‘Gender & AI Innovation Collective’, A+ Alliance, n.d., accessed 11 February 2026, https://aplusalliance.org/gender-ai-innovation-collective/.

[66] FREE-AI Committee Report (Reserve Bank of India, 2025), https://rbidocs.rbi.org.in/rdocs/PublicationReport/Pdfs/FREEAIR130820250A24FF2D4578453F824C72ED9F5D5851.PDF.

[67] MeitY, India AI Governance Guidelines.

[68] For example, community-based evals of MynaBolo, a women’s health chatbot used in urban informal settlements in Mumbai, revealed that identical Hinglish queries about miscarriage produced contradictory AI responses across testing sessions, and that reports of heavy bleeding during pregnancy did not consistently trigger clear guidance to seek urgent medical care. In contexts where users may rely heavily on the system, these failures posed risks of emotional harm, misinterpretation, and delayed care-seeking - issues unlikely to be detected through standard benchmark-based evaluations alone.

[69] ‘LIST Pioneers AI Regulatory Sandboxes and Launches Ethical Bias Leaderboard’, EurekAlert!, 20 March 2024, https://www.eurekalert.org/news-releases/1038466.

[70] Pillai et al., ‘Make It Real - Mapping AI-Facilitated Gendered Harm’.

[71] Pillai et al., ‘Make It Real - Mapping AI-Facilitated Gendered Harm’.

[72] On February 10, 2026, the Ministry of Electronics & Information Technology notified amendments to the IT Rules, 2021, which require intermediaries to remove or disable access to unlawful “synthetically generated information” within three hours of receiving a takedown notice, reducing it from the previous thirty-six hour timeline. Ministry of Electronics and Information Technology, ‘Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026’, 10 February 2026, https://master-meity.digifootprint.gov.in/documents/act-and-policies/information-technology-intermediary-guidelines-and-digital-media-ethics-code-rules-2021-it-rules-2021-IjM5QjMtQWa.

[73] Ibid.

[74] ‘Documentation of Survivors of Gender-Based Violence (GBV)’, GSDRC, 22 July 2021, https://gsdrc.org/publications/documentation-of-survivors-of-gender-based-violence-gbv/.

LEAD
KREA UNIVERSITY

GxD hub
Digital Equity. Connectivity. Insights.

digital futures lab

www.ifmrlead.org

www.gxdhub.org

www.digitalfutureslab.in