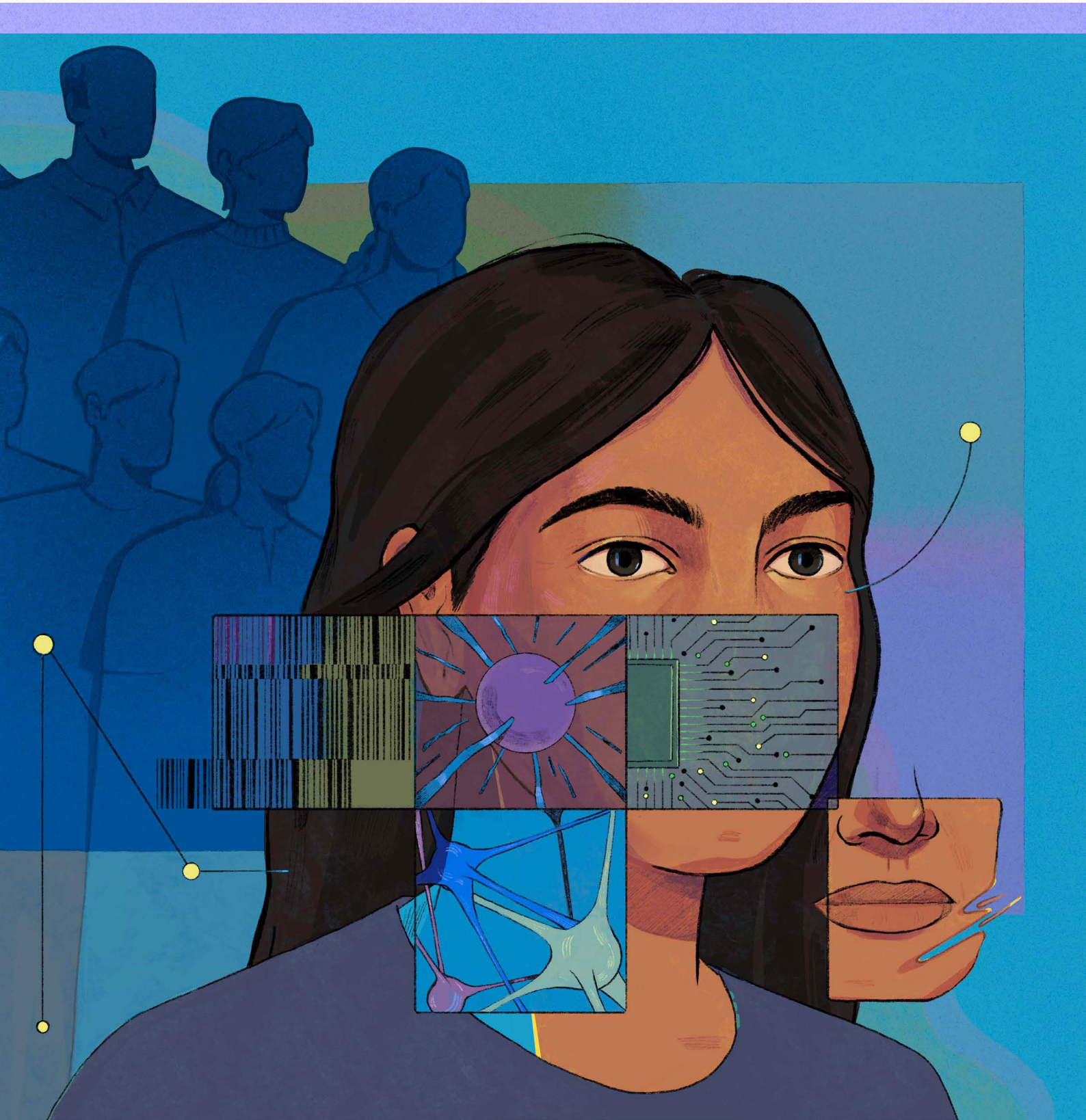
aapti institute

Funded by the European Union

B+HR BUSINESS AND HUMAN RIGHTS | ASIA

UNDP

Understanding Generative Artificial Intelligence's Implications on Gender Using a Value Chain Approach and a UNGP Lens



This report was produced by **Aapti Institute**, commissioned by the **United Nations Development Programme (UNDP)** under the Business and Human Rights in Asia programme, funded by the **European Union**. This document should not be considered as representative of the European Union's official position. The views expressed in this publication are those of the authors and do not necessarily represent those of the European Union, the United Nations, including UNDP, or the UN member States.

Aapti is a public research institute that works on the intersection of technology and society. It examines the ways in which people interact and negotiate with technology both offline and online.

ACKNOWLEDGEMENTS

In addition to contributions from the wider Aapti team, this report also draws upon the expertise of numerous academic and industry experts, practitioners, and workers. We are grateful for their input during interviews and feedback sessions.

Report design: Kartik Lav | **Cover illustration:** Joanna Davala | **Editor:** Yana Banerjee-Bey

ARTIST NOTE ON COVER PAGE ILLUSTRATION

The person in focus is being deconstructed by AI. Stripping down what makes them a human, losing nuance, and displacing personality. The faceless crowd in the background with individuals from different groups shepherded into queues to average out variation, sexuality, culture, and identity.

aapti institute

Funded by the European Union

B+HR BUSINESS AND HUMAN RIGHTS | ASIA

UNDP

Understanding Generative Artificial Intelligence's Implications on Gender Using a Value Chain Approach and a UNGP Lens

NANDINI JIVA, RITVIK GUPTA
AND KUNAL RAJ BARUA

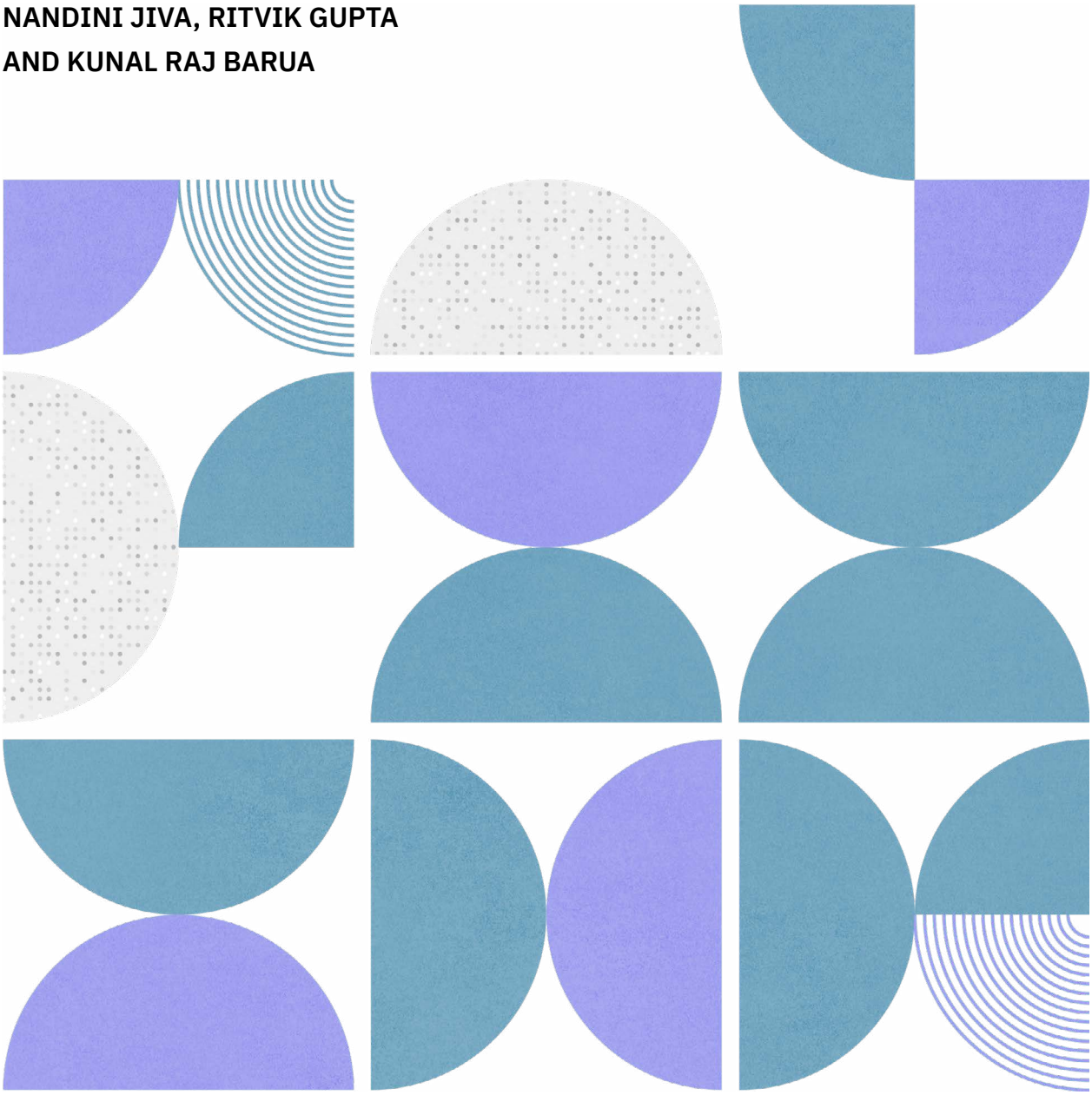


Table of Contents

OVERVIEW

Executive Summary

Executive Summary

Generative Artificial Intelligence (GenAI) technology has enjoyed widespread interest, experimentation, investment, and adoption since late 2022. GenAI has the potential to transform service delivery in a range of fields including law, healthcare, and social welfare. An analysis by EY-Parthenon estimates that the global economy could witness nearly USD 2 trillion GDP growth in the next decade, with various industries experiencing exponential growth in business operations and activities.[1] Another analysis by McKinsey Digital estimates that GenAI technology could absorb almost 70 percent of workers' time through automation, if businesses make the necessary investments.[2] Similarly, multiple estimates dub GenAI as a paradigm shifting force that could have tremendous impact on businesses and users. Yet, amid this continuing GenAI surge, states, businesses and society must grapple with the implications of scaling and expansion. The development and deployment of GenAI has also witnessed appreciable human rights consequences - with harms such as violent speech, bias and exclusion, and invasion of privacy of often vulnerable and marginalised populations.

In this context of dynamism and opportunity, with emergent harms, this report takes a systematic look at understanding the human rights risks that result from the use of GenAI for women and gender minorities and attempts to offer mechanisms for mitigation anchored in the United Nations Guiding Principles on Business and Human Rights (UNGPs), drawing from the Gender Dimensions of the UNGPs. For context, the UNGPs outline the roles and obligations of businesses, state, and society regarding the protection, prevention and remedy of business-related human rights violations.

[1] Boussour, L. (2024, March 22), "Harness the productivity potential of GenAI", EY-Parthenon

[2] Chui, M., E. Hazan, R. Roberts, A. Singla, K. Smaje, A. Sukharevsky, L. Yee, and R. Zemmel (2023, June 14), "The economic potential of generative AI: The next productivity frontier"

This study - an early inquiry into the intersection of GenAI, human rights, and gender - is anchored in a value chain approach. Use of a value chain approach allows unpacking of the processes, resources, and components involved in developing and deploying GenAI models, systems and services. This approach is particularly useful in gauging human rights risks, attributing agency and responsibility to actors, and devising solutions. Additionally, in an evolving global policy landscape, the study frames its understanding of human rights comprehensively, drawing from international and domestic (India) conventions and legislations.

At each stage of the GenAI value chain — dataset preparation, modelling, and model deployment — the study identifies specific risks that may be posed to women and gender minorities. The *dataset preparation* stage runs the risk of erasing women and gender minorities from datasets due to a 'gender-blind' approach. The absence of diverse gender identities in labour force data reduces meaningful representation within datasets. The lack of awareness around inclusionary data practices, and the paucity of training in data management practices further increases the potential harms that could emerge at this stage. Moreover, the dearth of privacy and user safety guardrails could compromise personal and sensitive data, resulting in unwanted bias and detrimental decision-making.

The *modelling* stage appraises a different set of human rights risks. For instance, 'gender-blind design' and assumptions restricted to binary or cis-male notions may translate to biases against women and under-represented gender groups in the operation of the technology. The absence of gender-sensitive training for GenAI developers further hinders inclusive development. This study also pinpoints the lack of a unified framework for inclusive GenAI development, and the inadequate screening of sensitive information and vague informed consent principles as modelling-related risks.

The final stage, *model deployment,* exhibits the risk of perpetuating gender stereotypes, and endangering human rights through AI-enabled misinformation, misgendering, and under-representation. Online harassment and gender-based violence have emerged as serious risks, with pornographic deepfakes being a case in point.

Use of pre-existing datasets with inherent biases also impedes objectivity in hiring — as highlighted by a case study detailed in Chapter 7. Lastly, the lack of specialised policies or guardrails for responsible use of GenAI is another area where human rights risks emerge.

Using the UNGPs, as well as the Gender Dimensions to the UNGPs, as the primary frameworks for analysis and assessment, this study proffers recommendations for businesses, suggesting pathways for mitigating human rights risks in their business actions. To guide this process, the report is accompanied by three risk assessment toolkits (attached as Asset 1 in the building blocks section of the report). This toolkit guides relevant stakeholders within businesses in self-assessing potential areas of human rights risks in their development and deployment of GenAI technology.

To work towards solutions, the study adopts an "un-risking" approach — with mechanisms to redress emerging human rights risks at each stage, specially focusing on the developers and actors closest to the technical artefacts. These mechanisms build on approaches that have previously worked in several contexts — overall mitigation mechanisms for data focus on building a wider range of open source models and small language models, while efforts are taken to enhance user trust. Mitigation at the model level will need to rely on improved model training, benchmarking, red teaming, and reinforcement learning. At the deployment stage, efforts need to focus on prompt engineering, provision of caveats and attribution in output, establishing feedback loops in addressing grievances, and working on public perceptions of technology.

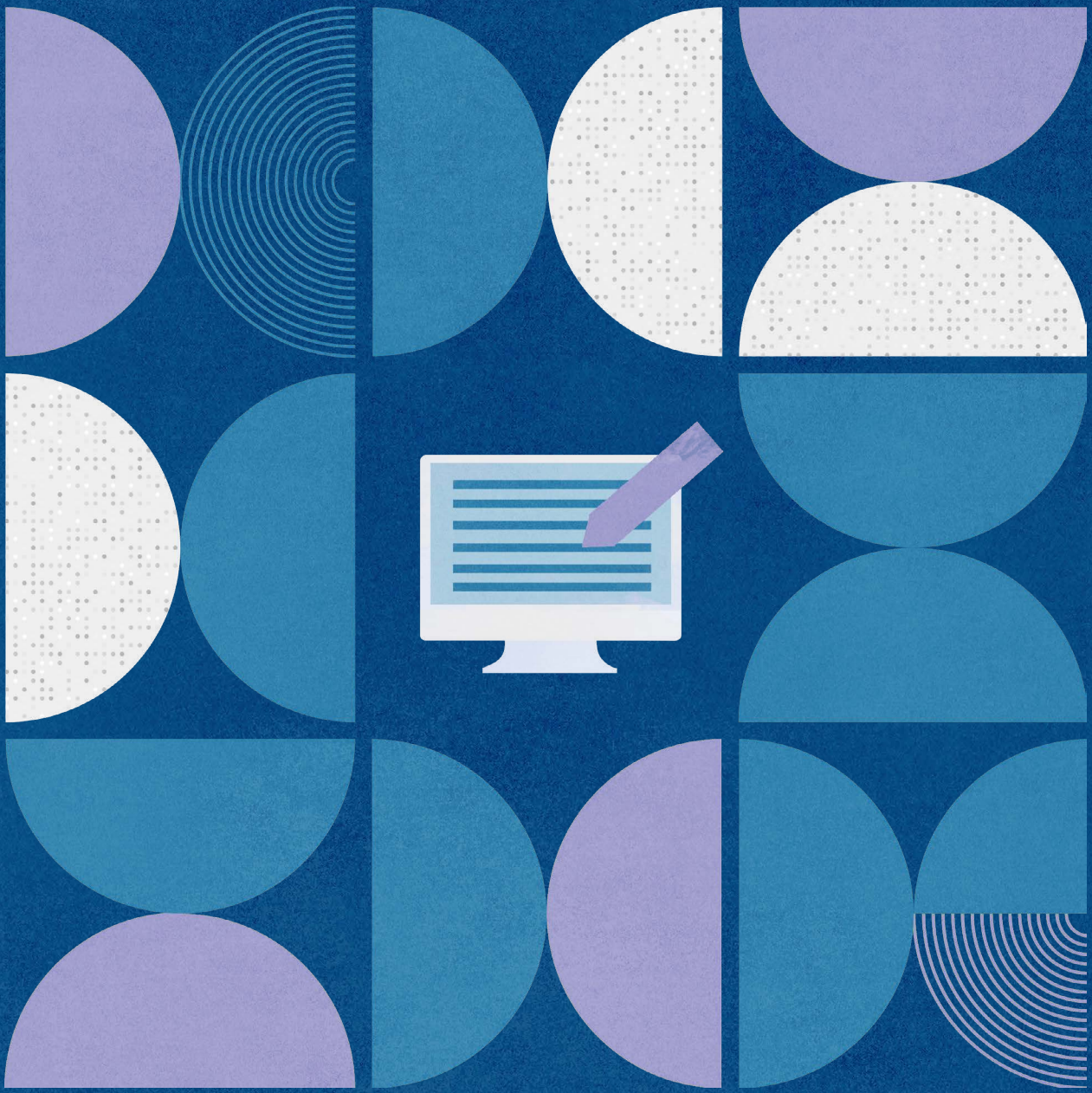
Based on UNGP Pillar 1, elucidating the state's duty to protect human rights, this study suggests pathways for state actors to be more involved in the responsible and effective development and deployment of GenAI based on the human rights risks highlighted in Chapters 5, 6, and 7. State actors can support the GenAI ecosystem by creating a contextualised gender inclusivity or gender fairness framework to guide business actions, allocating gender-responsive budgets, and supporting AI literacy efforts amongst developers and the workforce. Through collaborative approaches, state actors can enable businesses to build more socially responsible AI, co-create regulatory oversight mechanisms

— keeping in mind gendered dynamics and human rights considerations — and work with society and businesses to set up effective channels for redressal of harms and grievances. This report also proffers recommendations to state actors to guide their interface and regulation of GenAI technology (included in a policy brief attached as Asset 2 in the building blocks section of the report).

This structure of the report is as follows: **Part I** introduces the use cases, market adoption, and potential challenges of GenAI. It then introduces the value chain approach and identifies the UNGPs and the related Gender Dimensions as the primary framework steering the identification and assessment of human rights risks across the GenAI value chain. **Part II** articulates the need for the study and details the various points of analysis used to unpack the human rights implications of GenAI, relying on the UNGPs and the Gender Dimensions framework and the value chain approach. This section also expresses the need for an 'un-risking' approach. **Part III** dives into mapping the human rights risks to the established value chain and provides key recommendations for ecosystem actors. **Part IV** contains key tools developed, notably the risk assessment toolkits for businesses, the policy brief for state actors, a glossary of terms used in the study, and a detailed unpacking of the GenAI value chain. The report concludes (**Appendices**) by presenting the methodological approach and the research tools developed to conduct this study.

PART I

Introducing Generative Artificial Intelligence and Human Rights Risk



PART I

Introducing Generative Artificial Intelligence and Human Rights Risk

Possessing both extensive scope and a growing user base, GenAI is poised to transform systems, and arguably society, as we know them. However, the potential of GenAI should not overshadow the likely parallel impact on people. Thus, a close examination of its development and deployment is in order.

CHAPTER 1 | Introduction

The surge and potential of GenAI

Artificial Intelligence (AI), the "technology that enables computers and machines to simulate human intelligence and problem-solving capabilities",[3] has rapidly become a subject of discourse in recent years due to its adoption and expansion. Businesses like Google[4] and OpenAI[5] have shown interest in what AI can achieve for society, and governments globally are grappling with regulating the new frontier. AI has been perceived, understood, and touted as being capable of bolstering businesses and society in several ways.[6] Its potential has been reported through various industry estimates, with PrivewaterhouseCoopers (PwC) stating that AI could contribute more than USD 15 trillion to the global economy by 2030,[7] and a European Parliament briefing predicting worker efficiency to improve by nearly 40 percent.[8] While these estimates continue to evolve, the economic gains from AI technology are already immense.

[3] IBM (2024, March 19), "What is Artificial Intelligence (AI)?"
[4] Google AI. (n.d.), "Why we focus on AI"
[5] OpenAI. (n.d.), OpenAI Charter
[6] West, D.M., and J.R. Allen (2018, April 24), "How artificial intelligence is transforming the world"
[7] PwC Global, "Sizing the Prize, PwC's Global Artificial Intelligence Study: Exploiting the AI Revolution", PwC
[8] European Parliament briefing (2019), "Economic impacts of artificial intelligence (AI)"

A subset of AI, Generative Artificial Intelligence or GenAI is capturing the market at unprecedented rates, projected to grow to a USD 1.3 trillion market of its own.[9] While researchers point out that most, if not all, AI models in fact generate some form of content,[10] the study borrows from GenAI models and foundation models[11] to develop its understanding of the GenAI value chain. GenAI systems, often built on foundation models,[12] are trained on large volumes of data, and can generate outputs in the form of text, images, codes, audio, and video, from human instructions.[13] For instance, in November 2022, the world witnessed OpenAI launch the GenAI service ChatGPT – a chatbot that used OpenAI's Generative Pre-trained Transformer 3 (GPT-3)'s foundation model. The chatbot's launch was a watershed moment marking the surge of interest in, adoption of, and experimentation with GenAI.[14]

GenAI has much to offer - it has the potential to complement and potentially compensate human errors and biases and conduct a broad spectrum of business operations and activities. To ensure that its growth doesn't disproportionately impact specific population groups, debates and actions around timely and relevant governance of these systems should occur parallelly.

The emerging challenges of GenAI

The benefits of GenAI for business are undeniable but, equally, so too is its fallout regarding humans and human rights. It thus becomes crucial for GenAI systems to be inclusively designed and deployed. Given the rapid pace of this development and deployment, measures to ensure consideration for the human rights impacts of new technology need urgent attention. Emerging concerns in various industries where such impact could be considerable range from information manipulation, financial harm, identity theft to other techno-social harms.[15] Research on GenAI search engines indicates that if left unchecked, GenAI models could exacerbate biases, further the information divide, and generate misleading information, all while appearing to be a reliable source of information for users.[16]

More specifically, research has highlighted that GenAI could specifically harm women and under-represented gender minorities

[9] Bloomberg (2023, June 1), "Generative AI to Become a $1.3 Trillion Market by 2032, Research Finds"
[10] García-Peñalvo, F.J. and A. Vázquez-Ingelmo (2023, August), "What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI", ResearchGate
[11] Murphy, M. (2022, May 9), "What are foundation models?", IBM Research
[12] Amazon Web Services (AWS), (n.d.), "What are Foundation Models?"
[13] Stanford University, Stanford Institute for Human-Centered Artificial Intelligence (n.d.), section taken from 'What is a Foundation Model' Developing and understanding responsible foundation models", Center for Research on Foundation Models
[14] Marr, B. (2023, May 19), "A Short History Of ChatGPT: How We Got To Where We Are Today", *Forbes*
[15] Ferrara, Emilio (2024), "GenAI against humanity: nefarious applications of generative artificial intelligence and large language models", *Journal of Computational Social Science*, ResearchGate
[16] Memon, S.A. and J.D. West (2024, February 18), "Search engines post-ChatGPT: How generative artificial intelligence could make search less reliable", Center for an Informed Public, University of Washington,

more than men.[17] To understand this further, this study analysed a UN Women report, highlighting four areas of bias where AI systems impact women more than men - discrimination, stereotyping, exclusion, and insecurity.[18] Due to the type of outputs that GenAI systems create, they can replicate, perpetuate, and exacerbate gender inequities, lead to an increase in hypersexualised and homophobic content, and give rise to inaccurate and harmful information - such as deepfakes. Similar studies by the GDELT Project uncovered the role of GenAI in strengthening pre-existing gender stereotypes using the examples of male and female leaders in industry, and highlighted the hypermasculine stereotypes that currently inform the tech industry.[19] The following examples indicate ways in which GenAI systems could potentially lead to the abuse of under-represented gender groups, and the impact that such instantiations could have:

- **Reproducing Biases:** GenAI outputs have a propensity to reproduce biases, stereotypes, and sexism.[20] For example, DALL-E and Stable Diffusion have produced skewed representation for categories like occupations, turning out far more depictions of males as scientists and "IT Experts" than any other gender.[21] With the increased reliance on customised search results through GenAI search,[22] such results could lead to continued exacerbation and internalisation of harmful gender stereotypes.
- **Perpetuating harassment and discrimination:** Deepfake technology is a prominent form of GenAI that, when used maliciously, contributes to gender-related harassment and violence. GenAI's capability for visual content generation and modification has exacerbated the creation and circulation of non-consensual nudity and explicit content, particularly relating to women.[23]
- **Increasing exploitation and objectification:** It is a fact that bad actors exploit various technology systems.[24] In the case of GenAI where higher quality content can be produced quickly, and in the form of various output formats, these threats can severely impact women's safety and security online.[25] In addition to such misuse of GenAI systems, the capabilities and tendencies of the technology also hint at a less than equitable slant on outputs for non-majority gender identity groups such

[17] Lamensch, M. (2023, June 14), "Generative AI Tools Are Perpetuating Harmful Gender Stereotypes", Centre for International Governance Innovation

[18] Fournier-Tombs, E., J. Lee, M. Yang and P. Raghunath (2024), "Artificial Intelligence And The Women, Peace and Security Agenda in South-East Asia", Women, Peace and Cybersecurity: Promoting Women's Peace and Security in the Digital World project, UN Macau and UN Women Regional Office for Asia and the Pacific

[19] The GDELT Project (2023, August 5), "Gender & Race In ChatGPT's CEO Stories & How Embedding Models Rank White Male CEOs First And Women Last"

[20] Lázaro, Mónica Melero and F.J. Garcia-Uii (2023, August), "Gender stereotypes in AI-generated images",ResearchGate

[21] Nikolic, K. and J. Jovičić (2023, April 3), "Reproducing inequality: How AI image generators show biases against women in STEM", UNDP Serbia

[22] shelf (2023, November 3), "How Generative AI is Transforming Search Forever", Generative AI, shelf

[23] Chowdhury, Rumman, and D. Lakshmi (2023), "'Your opinion doesn't matter, anyway': Exposing Technology-Facilitated Gender-Based Violence in an Era of Generative AI", UNESCO Digital Library

[24] Daver-Massion, M., and J. Taylor (2023, August 29), "The role of GenAI in enabling online harms: How do new tools pose unique risks to online safety?", *PUBLIC*

[25] O'Neil, L. (2024, April 8), "Generative AI is making the online abuse of women as easy as point-and-click. Is there any way to stop it?", Harm reduction, *Compiler*

as women and minority gender groups. When attempting to use Lensa, a GenAI avatar generator, a reporter was provided with multiple overtly sexual representations, sometimes even nude images, of herself due to the models programming, despite explicitly seeking a male avatar.[26] Further exploration and interactions with experts helped conclude that an overabundance of sexualised images of Asian women on the internet had led to the technology's pattern recognition being influenced by the input data that consisted of inappropriate images and representations.

The existing digital divide and varied ways in which non-majority genders interact with platforms, models and systems have made digital technology a controversial space for women and non-binary genders. This aspect can be heightened by the emergence of an efficient tool like GenAI, with lower technical barriers to entry.[27] The instances cited above underline how valuable it is to grasp the implications of such transformative systems, especially for women and gender-minority groups. Ecosystem efforts such as Aapti Institute and UNDP B+HR Asia's previous work on 'Understanding the implications of AI systems on Human Rights'[28] coupled with initiatives like the 'B-Tech' project launched by the United Nations and Office of the High Commissioner of Human Rights, consolidate crucial evidence – making a strong case for states and businesses to view GenAI systems through the lens of human rights[29] – and are spearheads for this study.

Adopting a gender-first approach

As with most paradigm shifting forces, digital technology harbours the ability to empower and enable humankind.[30] However, alongside the promise of empowerment there lurk risks of widening inequalities and human rights abuses. Androcentric societies often exclude women from participation and and adequate representation in public spheres, including social, political, and economic positions of power. The internet has been widely perceived as an unsafe space for women and under-represented gender groups. A survey conducted by the World Wide Web Foundation found that more than 50 percent of young women respondents had faced some form of online abuse.[31]

[26] Heikkila, M. (2022, December 12), "The viral AI avatar app Lensa undressed me – without my consent", *MIT Technology Review*

[27] Daver-Massion, M., and J. Taylor (2023, August 29), "The role of GenAI in enabling online harms: How do new tools pose unique risks to online safety?", *PUBLIC*

[28] Rai, A., Vinay Narayan and S. Natarajan (2022), "Artificial Intelligence and Potential Impacts on Human Rights in India", UNDP India

[29] Supplemental Paper: "Taxonomy of Generative AI Human Rights Harms", B-Tech Project, United Nations Human Rights, Office of the High Commissioner

[30] United Nations (n.d.), "The Impact of Digital Technologies", United Nations

[31] World Wide Web Foundation (2020, March 12), "The online crisis facing women and girls threatens global progress on gender equality", Web Foundation

The gender-digital divide continues to widen with women in developing countries having limited access to digital tools.[32] Women continue to be under-represented in AI-development roles, and as end-users of AI-systems.[33] Furthermore, women and under-represented gender groups are predominantly misrepresented and hypersexualised in online spaces.[34]

The potential of GenAI systems remains vast, yet women could face a disproportionate fallout of poorly designed and deployed GenAI systems. Under-represented gender minorities, such as non-binary genders and persons of diverse SOGIESCs (Sexual Orientation, Gender Identity, Gender Expression, and Sex Characteristics), experience significant harms through poorly designed and deployed GenAI systems.[35] The impact of AI and GenAI models and services on gender could amplify further. These harms are often more nuanced when using a gender intersectional lens. GenAI's potential to create images, text, audio, and video based on word prompts is accompanied by the fact that these outputs do not borrow from a diverse gender representation.[36] These services exacerbate existing stereotypes and misrepresent women and gender minorities by limiting their representation to defined roles and responsibilities. Further, the impact this has on non-binary genders extends to hypersexualised content, deepfakes, harmful stereotyping, and homophobic content that are targeted towards the gender identity and/or sexual orientation of non-binary persons.

In the addressing of gender-related harms, there is a missing voice – the one recognising the experiences of women and non-binary genders. To create an equitable field that is just to such stakeholders, a more holistic approach is required to unpack the complexities of gender and proffer more equitable digital solutions. The absence of such an exercise can have far-reaching implications on GenAI services, ranging from education and employment to banking and healthcare. Reducing the divide between AI and women and gender minorities is crucial for progress towards an equitable society, creation of diverse workplaces, and mitigating bias and reinforcement of existing gender stereotypes.

[32] Ibid
[33] Aronsohn, I. D. (2024, February 22), "Why we need to act now on bridging the GenAI gender gap", Amdocs
[34] Lamensch, M. (2023, June 14), "Generative AI Tools Are Perpetuating Harmful Gender Stereotypes", Centre for International Governance Innovation
[35] Leufer, D. (2023, January 13), "Computers are binary, people are not: how AI systems undermine LGBTQ identity", Accessnow
[36] Lamensch, M. (2023, June 14), "Generative AI Tools Are Perpetuating Harmful Gender Stereotypes", Centre for International Governance Innovation

The necessity for such a gender lens becomes all the more imperative against the backdrop of cumulative historical and societal conventions and the resultant manifestation in contemporary digital tools. According to a 2023 UNESCO Report, only 22 percent of working professionals in AI are women, which can have detrimental consequences in the development of AI services in the form of bias, exclusion, and invisibilisation. This erasure is more pronounced with non-binary identities and under-represented gender groups of diverse SOGIESCs, with stereotyping that depicts non-majority gender identities with a binary perspective and a limited understanding of the non-binary experience.[37]

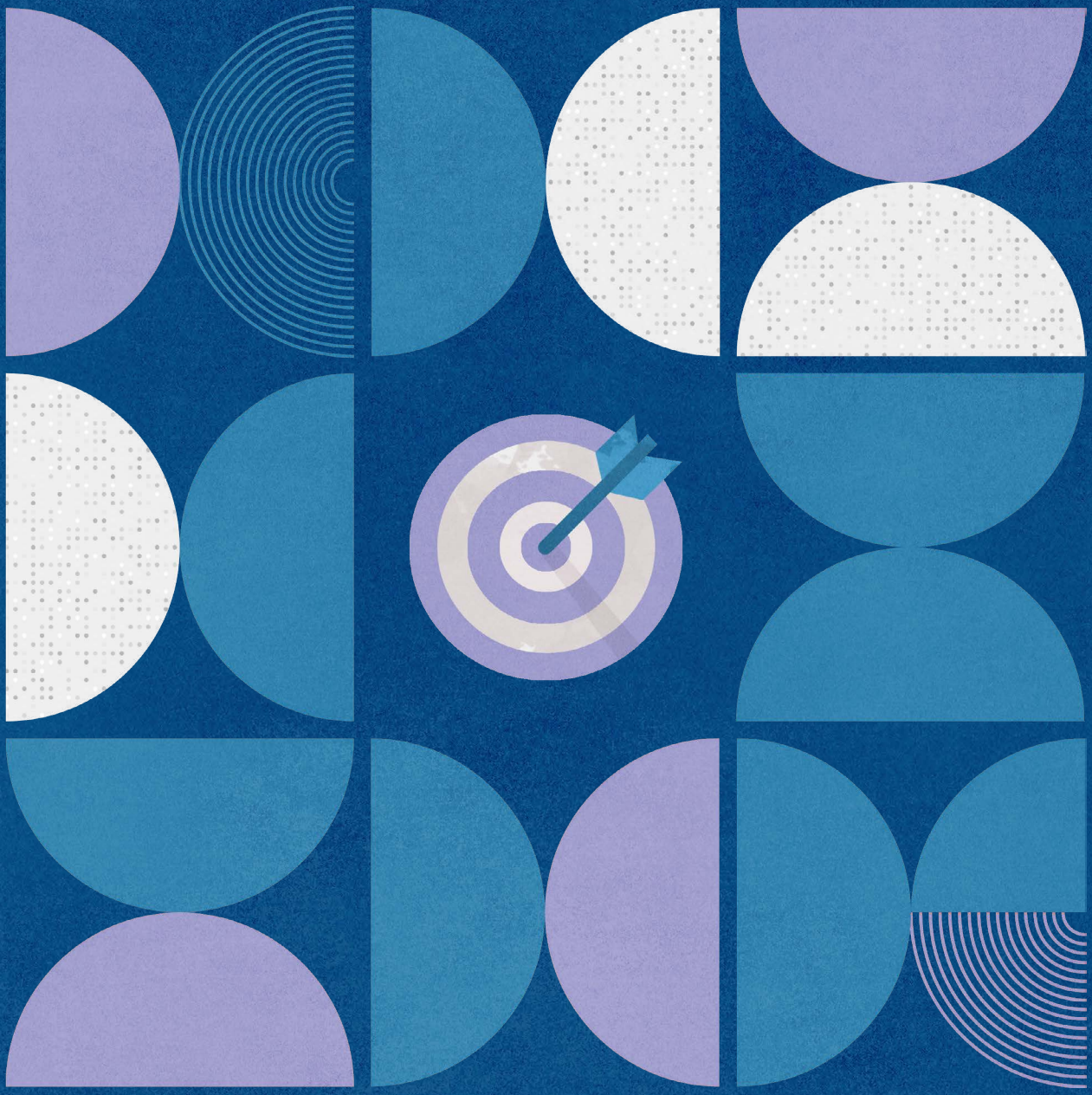
Adopting the gender lens is vital to ensure perpetual information assimilation and improvement of GenAI services' comprehension of the nuances of gender. Knowledge gaps in AI developer circles, along with the lack of gender diversity in the workforce, can be compensated for by greater interaction with perspectives focusing on accountability and transparency in AI services. Development and deployment of AI models that recognise women and gender minorities would yield obvious results: increasing active representation in GenAI outputs, strengthening gender-diverse leadership and creation of AI, and making the voices of women and gender minorities part of the development and decision-making process in AI outputs.[38]

[37] Rogers, R. (2024, April 2), "Here's How Generative AI Depicts Queer People, *WIRED*

[38] Chau, C. (2024, April 21), "Strengthening Gender Equality in Generative AI", American Medical Women's Association (AMWA)

PART II

Anchoring



PART II

Anchoring

The study is anchored in understanding the impact of digital technologies on people, by taking a gender lens to human rights frameworks. It employs the United Nations Guiding Principles on Business and Human Rights ('UNGPs') to identify human rights risks and frame solutions, incorporating key nuances from the Gender Dimensions of the UNGPs. The technical exploration of the study is grounded in a value chain approach to unpack GenAI systems. The report draws from global as well as Indian regulatory frameworks to unpack the policy perspectives governing the AI ecosystem.

This section dives into the nuances of these anchor points before articulating the potential human rights risks across the GenAI value chain.

CHAPTER 2 | The why and how: Rationale, Value Chains, and the UNGPs

This chapter outlines the rationale for the study, the value chain approach used to analyse GenAI's gender risks, and the UNGPs' involvement, with the relevant principles pertaining to gender, namely the Gender Dimensions on the Guiding Principles. This chapter lays the ground for the subsequent chapters which outline the risks that emerge from three stages of the value chain - dataset preparation (Chapter 5), modelling (Chapter 6), and model deployment (Chapter 7) stages.

Need for the Study

While AI has displayed indications of efficiency gains in numerous sectors, it has also been well established that use of AI in different spheres has significantly impacted human rights. This study is

constructed to rouse awareness around the gendered implications of GenAI systems and consolidate a definitive base for further inquiry into GenAI and gender. With respect to gender harms, current approaches have focused on inclusion through the involvement of women in Science, Technology, Engineering, and Mathematics (STEM)-related fields, sectors, and education, creating more representative datasets, and the installation of various forms of guardrails and prohibitive measures to inhibit malicious use of GenAI for gender harms. Discussions on the mitigation of gendered hate speech in online searches is another example of such collective movements towards gender equality.[39] While these attempts are significant, there is a case for a systematic review of gender harms arising from GenAI, which this report attempts.

Additionally, there is a need to have **specific recommendations for stakeholders, directly emerging from the UNGPs,** since states and businesses require clarity on harm prevention, mitigation, and redressal. Further, affected communities and societal players need ways to scrutinise GenAI-related produce, and highlight issues with the relevant businesses and state actors.

The United Nations Guiding Principles on Business and Human Rights (UNGPs) and the Gender Dimensions of the UNGPs

The UNGP framework helps direct responsibility and roles between states and private businesses. These frameworks are important to this study owing to its ability to help decide who should act and under what kind of obligations, bridging this inquiry's mapping of GenAI development and deployment to human rights framings, and drawing attention to the emergent risks of GenAI. Additionally, the UNGPs help in guiding actors around perceptions and responsibilities relating to the human rights framework and influence their responses to ensure a more rights-based approach.

The three pillars – Protect, Respect, and Remedy – alongside the principles on human rights, help connect the GenAI risks identified to human rights abuses, and eventually link to the obligations of business and government to act against such risks. The UNGPs also help with distribution of pathways to change and mitigation, and

[39] UN Women (2013, October 21), "UN Women ad series reveals widespread sexism"

key self-assessment considerations across the GenAI ecosystem's stakeholders.

The UNGPs have been interpreted for this study in the following manner:

- It is the state's *duty to* ***protect*** *human rights* by enacting effective policies and legislations and redress any human rights abuses caused due to business activity.
- Businesses have the *corporate responsibility to* ***respect*** human rights of all rights holders impacted by their business operations and supply chains and conduct due diligence to identify, prevent, and mitigate any human rights risks to rights holders.
- The state and businesses must **provide *access*** *to grievance redressal to individuals and communities for any adverse human rights impact, caused directly or indirectly by any business activity,* through effective and expedient judicial or non-judicial grievance redressal mechanisms.

Part III of this report outlines the stages of GenAI development and deployment, and links risks identified with each stage to human rights and to the UNGP directives, setting up tangible points for planning of solutions, redressal, and future prevention. Thus, the UNGPs serve as the intervening step between human rights and their more contextualised protection.

Gender Dimensions of the UNGPs

The Gender Dimensions of the UNGPs outline gender-sensitive interpretations of the 31 Guiding Principles. The framework identifies key principles and adds perspectives, guidance, and illustrative actions for states, businesses and other key stakeholders to consider the gendered impact of business-related human rights abuses on traditionally under-served communities.

While the UNGPs themselves provide clarity on stakeholder roles, principles to frame organisational values, and guidance on how to operationalise a rights-respecting business model, the Gender Dimensions add a much-needed layer of considerations and

recommendations to reduce, mitigate, remedy, and communicate the gender impact of businesses. The Gender Dimensions are essential in guiding actors through gender-relative nuances and can be applied across various industries and domains.

This study draws inspiration from the guidance provided under Principle 24 of the UNGPs – to provide evidence based research, tools, and recommendations to businesses and business stakeholders primarily, for identifying, mitigating, and remediating human rights imbalances that might emerge for women and minority gender groups.

At its core, this research applies the framing of the Gendered Dimensions of the UNGPs to the question and helps guide the growth of the GenAI industry. Leveraging these dimensions helps further consider the socio-normative dynamics that exist in this space, and also GenAI industry contributions to the erosion of rights of women and SOGIESC communities.

The value chain approach

For a closer look at the gender risks posed by GenAI technologies, it is crucial to map out the different stages of the AI value chain. An AI value chain is the "organisational process through which an individual AI system is developed and then put into use (or deployed)".[40] By breaking down GenAI models and deployed services into distinct but related contributors, the study identifies and connects the contexts, people, and practices that shape AI's ethical issues.[41] Each layer of (or element within) the value chain – dataset preparation, modelling, and model deployment – presents potential sites for specific human rights risks. These sites are where harms could emerge and where un-risking can occur.

- **The stages value chain:** GenAI's three stages of dataset preparation, modelling, and model deployment represent how key resources are gathered, the technology is developed, and how it reaches people, groups, and market sectors.
- **Sub-components within the value chain:** To meaningfully unpack the GenAI value chain, the study has identified key elements within the three stages to reflect the multi-directionality of data, value, and decision flows within. These

[40] Engler, A. and A. Renda (2022, September 30), "Reconciling the AI Value Chain with the EU's Artificial Intelligence Act", Centre for European Policy Studies (CEPS), page 2

[41] Attard-Frost, B. and D.G. Widder (2023, July 31), "The Ethics of AI Value Chains", arXiv, Cornell University, page 9

elements have been visually represented within the chapters and are the key sites of action in the risk assessment toolkit.

Appraisal of the components and processes within GenAI facilitates the linking back of human rights to entities, actors, or processes – which are very often distinct. This approach allows specific actors or entities engaged in the value chain to be supported and enabled to better tackle human rights risks. This approach is crucial when unpacking the stages of a GenAI system's life cycle, and to understand the elements that contribute to the larger stages.

The evolving policy landscape around AI

The landscape for regulation, accountability, and spreading best practices in the context of AI is still nascent, though several regulatory efforts can be considered noteworthy – the European Union's (EU) AI Act was passed in March 2024,[42] and the Executive Order on trustworthy AI signed by the US President in October 2023.[43] AI-related regulations are also emerging in various states of the US – for example, the Connecticut General Assembly has proposed an AI Act, part of which is meant to prevent the circulation of 'synthetic images'.[44]

India has been proactively engaging in the global conversations on AI regulations, while making efforts within the country to enable responsible AI and mitigate its ill effects. India is working on the Digital India Act (DIA), purported to be an umbrella legislation to deal with the harms of AI.[45] Some states – Karnataka for instance – have established and operationalised 'Fact Check Units' to deal with misinformation.[46] Equally, efforts to foster responsible development of AI are also underway. The Ministry of Electronics and Information Technology (MeitY) has created a dedicated division to AI and emerging technologies[47] in addition to various other initiatives to build a robust AI ecosystem.[48] The Responsible AI initiative, focused on compute, data and human capacity creates a policy environment for the just development of AI.[49]

As part of its G20 presidency, in the G20 New Delhi Leaders' Declaration, India stated its goal to drive inclusive and rights-protecting development and deployment of AI systems and to

[42] Future of life Institute, The Act Texts, EU Artificial Intelligence Act
[43] The United States Government (2023, October 30), Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, The White House
[44] An Act Concerning Artificial Intelligence: Substitute for S.B. No. 2 Session Year 2024, Connecticut General Assembly
[45] Press Trust of India, "Digital India Act will deal with ill effects of AI: MoS Chandrasekhar" (2023, November 24), *Business Standard*
[46] Pandey, K. (2024, May 1), "Karnataka's fact checking unit began operations in March: Report", MediaNama
[47] Government of India (n.d.), AI & Emerging Technologies Group, Ministry of Electronics and Information Technology,
[48] Raja, A. (2023, December 22), "Top AI initiatives by Meity in 2023", INDIAai
[49] Government of India (n.d.), Call For Expression of Interest on Responsible AI, Ministry of Electronics and Information Technology

harness AI systems responsibly, committing to the G20 AI Principles of 2019.[50] India also assumed the 'Council Chair' of the Global Partnerships for AI (GPAI), working with member states to create a safe and trustworthy AI environment for its citizens.[51] Further, several state adjacent bodies like The Indian Industry Association, and the National Association of Software and Service Companies (NASSCOM) have produced work on Responsible AI for businesses. India's momentum to drive the development of responsible and inclusive AI systems is visible in such initiatives. The installation of safeguards and regulation of systems with a human rights focus is crucial for extending the approach to further strengthen the protection of humans and make remediation more meaningful.

Tying it all together

This chapter has demonstrated the relevance for this research, while also outlining the key anchor points. The section has articulated the usefulness of the UNGPs and the Gender Dimension framework in understanding stakeholder roles and articulating mitigation pathways. The section also highlights the need for a human rights based approach and for a gender lens focus for this inquiry, to effectively shed light on the real-world implications of such systems.

The value chain approach allows the study to categorise the risks and solutions for relevant actors to take more meaningful and pointed action. The following chapters identify these gender-related risks based on the value chain stage and suggest mitigation pathways for GenAI developers and deployers to adopt.

[50] Government of India (2023, September 10), G20 New Delhi Leaders' Declaration, Ministry of External Affairs

[51] Government of India (2022, November 21), "India takes over as Council Chair of Global Partnership on AI (GPAI)", Press Information Bureau

Human Rights sources: Taking a wide view

For the purposes of this mapping of gender risks in GenAI technology, the study understands human rights and gender risks as potential harms arising out of design choices, practices, and processes feeding GenAI's development and deployment.

This study is anchored in a relatively wide pool of sources to understand and identify human rights risks. It relies on international and Indian human rights and regulatory frameworks to unpack the areas where human rights risks and violations emerge.

Some of the primary sources driving our thinking on human rights are listed below:

- The Universal Declaration of Human Rights **(UDHR)**
- International Covenant on Civil and Political Rights **(ICCPR)**
- International Covenant on Economic, Social and Cultural Rights **(ICESCR)**
- Convention on the Elimination of all Forms of Discrimination Against Women **(CEDAW)**

These frameworks, analysed in addition to the UNGP framework, enabled the study's attempt to map elements such as violations of India's Constitution and laws, and to look at local rights as well as legal contexts around GenAI's gender risks.

The Indian legal landscape

In the Indian context, the human rights regulatory framework is dominantly derived from the Constitution of India.[52] The study focuses on these constitutional rights, and other relevant legislations, to portray the human rights narrative in India. It also looks closely at the Digital Personal Data Protection Act, 2023 (DPDP Act) to examine the processing of and access to personal data.

[52] Government of India (2021, November 26), The Constitution of India

The study also delves into the IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011,[53] and the IT (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021[54] to trace clear liability when it comes to the ethical and safe processing and collection of sensitive and personal data. The proposed DIA[55] advocates a legal and institutional quality testing framework to examine regulatory AI models, algorithmic accountability, and vulnerability assessment through content moderation.

Apart from the digital protection regime in India, the study has also examined other statutory rights in Indian legislations that derive legitimacy as human rights from the Constitution. The study examines the Indian Labour Codes, 2020, specifically the Code on Wages, 2019,[56] and the Equal Remuneration Act, 1976,[57] which apply when discussing the role of GenAI in hiring systems, and the Right to Information Act, 2005,[58] which entitles all persons to seek and receive information in the spirit of transparency.

Delineating the need to look at GenAI in India through such a lens has its own set of legislative and commercial challenges. There is merit in noting that India has one of the most modern 'open' digital infrastructures in the world,[59] spearheading inclusivity and access. It is also one of the largest markets for AI applications to drive growth and productivity of enterprises. Moreover, India's International Human Rights Law (IHRL) commitments stem from the UDHR, and through its ratification of the ICCPR, ICESCR, and the CEDAW.

The Indian legal system has accommodated women's empowerment through a range of fundamental and legal rights under the Constitution and legislations dealing with skill development,[60] digital literacy,[61] and so on. Such initiatives and policies have propelled deeper thinking around gender; for this study, though, a more comprehensive approach has been adopted.

With India's efforts to build inclusive and responsible digital systems, and its human rights commitments, it would be apt for the country's developers and businesses to work towards building technology that is geared towards respecting rights.

[53] Ministry of Communications and Information Technology (Department of Information Technology) (2011, April 11), G.S.R. 313(E). IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules

[54] Ministry of Electronics and Information Technology. (2023, April 6). The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021

[55] Ministry of Electronics and Information Technology (2023, March 9), Proposed Digital India Act, 2023

[56] Government of India (2019), The Code on Wages, 2019, Bill No. 184 of 2019

[57] Government of India (1976), Equal Remuneration Act, 1976, Act 25 of 1976 amended by Act 49 of 1987

[58] Parliament of India. (2021. May 17), The Right to Information Act, 2005 (amended)

[59] Ernst & Young India, (n.d.), "The AIdea of India: Generative AI's potential to accelerate India's digital transformation"

[60] Vocational Training Programme for Women, Ministry of Skill Development and Entrepreneurship, Government of India

[61] National Institute of Electronics & Information Technology, Ajmer (2015, December 1), National Digital Literacy Mission (NDLM)

Using an "un-risking" approach and upholding Human Rights

Building on the legal principles, this study takes an "un-risking" approach to the question of GenAI's implications for non-cis-male gender identities. "Un-risking" refers to the identification and attempted mitigation of risks threatening human rights. This approach enables systems to work towards upholding human rights through the various rights declarations and conventions that countries, including India, are party to. Un-risking can be one route to improved gender outcomes and preservation of human rights.

Gender un-risking is in play, for instance, when developers create ways to have image-generating GenAI models refuse specific user requests such as those seeking content featuring nudity or sexual themes. This could be instantiated in specific ways: models could be trained to identify requests that go against a usage policy, where the policy explicitly articulates the themes that GenAI systems do not address or provide responses to. Creating and embedding robust usage policies could help fine-tune a GenAI model. The use of GenAI models could be monitored, as a service, and training data could be made more representative and diverse.The un-risking approach allows developers to put strategic efforts in place that address harms that emerge from GenAI systems.

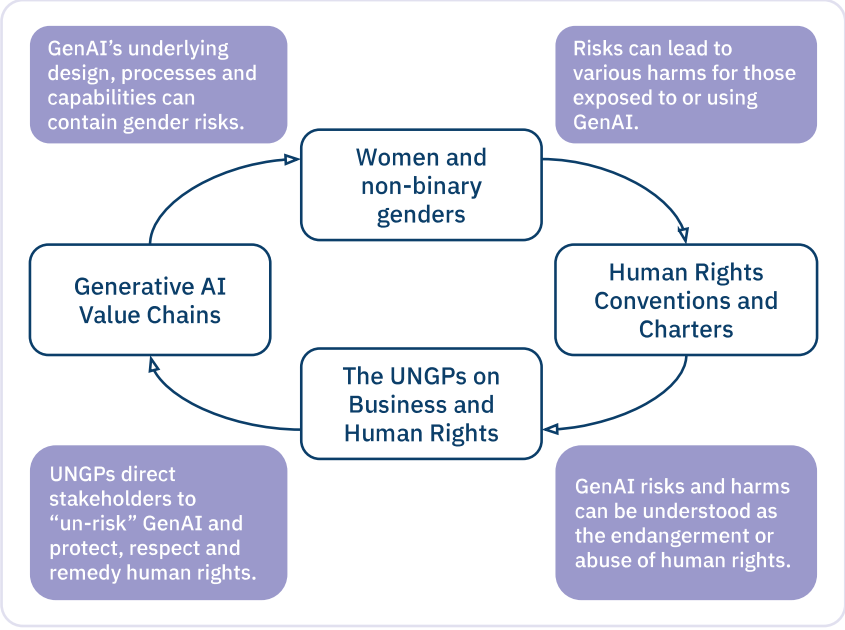


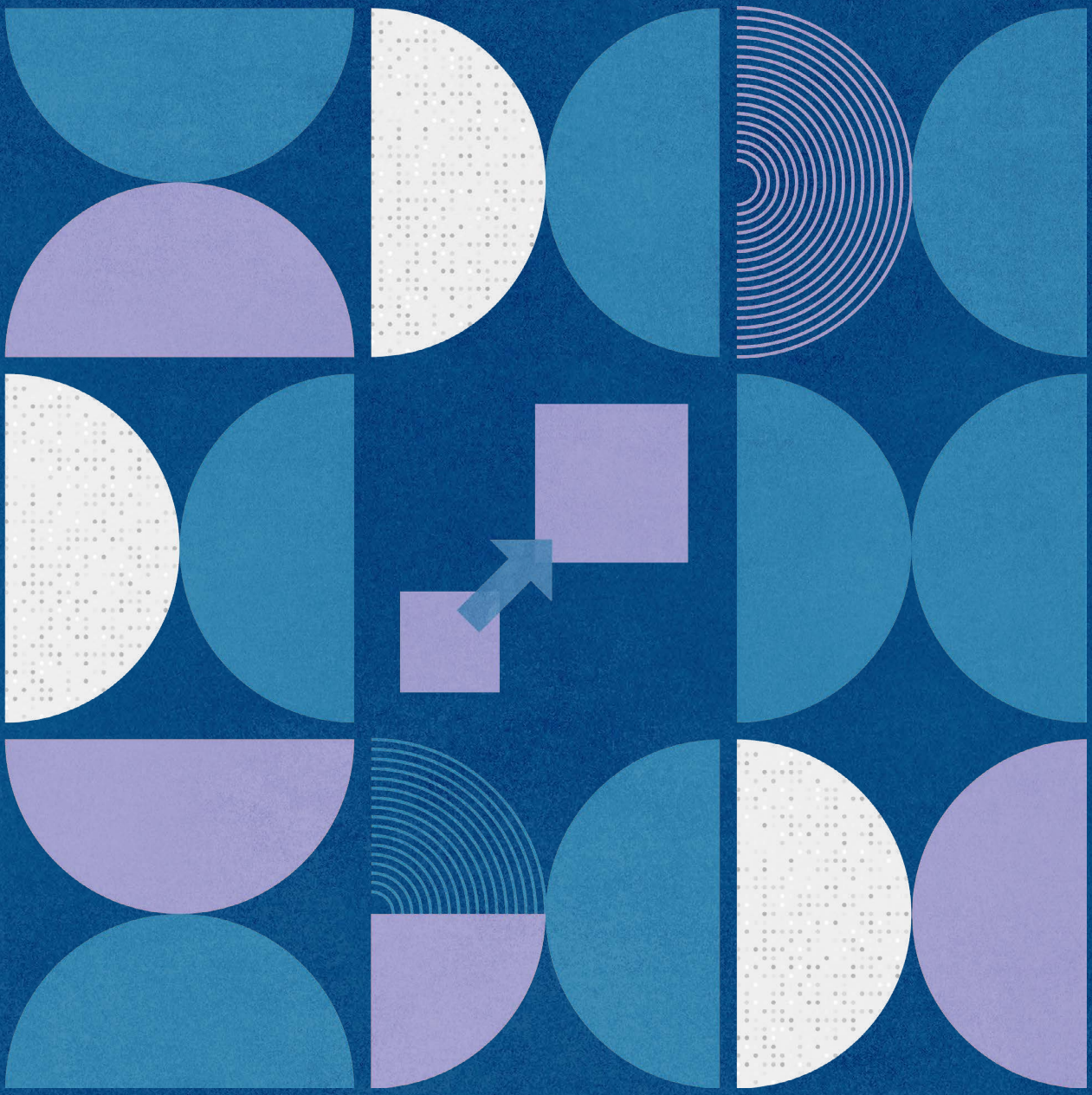
Figure 1. From Risks to Resolution via Value Chains and Human Rights

From “un-risking” to value chains

Before explaining the findings on the kinds of human rights and gender risks that inhabit GenAI, a short explanation of the value chain approach is presented. Chapter 4 delves into the value chain and provides stage specific details in Chapters 5 through 7. These chapters rely on the value chain mapping to present human right risks and recommend ways to ‘un-risk’ GenAI systems.

PART III

Risks and Solutions



PART III

Risks and Solutions

The risks presented by GenAI can be traced back to the different stages in developing GenAI models and services, and their release into, and circulation within, markets. This section analyses and documents the human rights and gender risks emerging from use of GenAI.

CHAPTER 4 | Breaking GenAI Down: The Value Chain Approach

AI value chains are the "organisational process through which an AI system is developed and then put into use (or deployed)".[62] This study has developed and makes use of the "stages" value chain that breaks GenAI into three broad stages, and then highlights "sub-components" within each stage.

[62] Engler, A. and A. Renda (2022, September 30), "Reconciling the AI Value Chain with the EU's Artificial Intelligence Act", Centre for European Policy Studies (CEPS), page 2

The Stages Value Chain

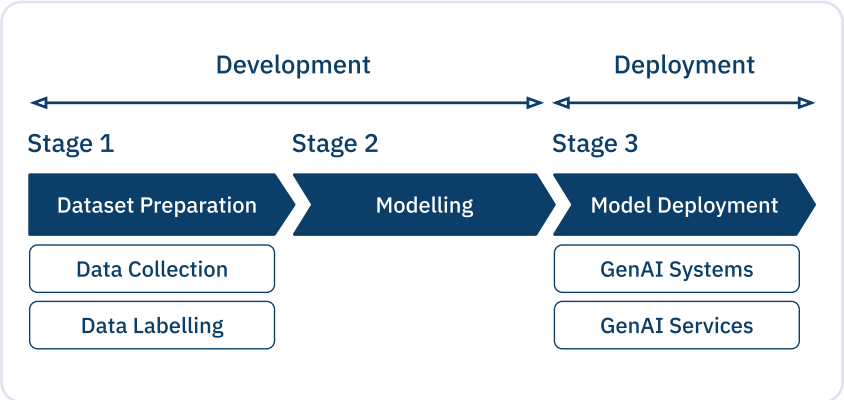


Figure 2. The Stages Value Chain

Figure 2 outlines the three stages of GenAI development and deployment in the stages value chain which understands them as reading various forms of data, producing a model capable of a wide range of responses and outputs based on human prompts, and integrating the model into systems or some form of a service.

Data, in forms like images and text, is required to develop a GenAI model's capability to generate content in response to human input. This corpus of data is assembled by wading through and gathering publicly available data on the internet, or data which is licensed and provided to a particular company or developer by a third party. For example, OpenAI's large language model (LLM), Generative Pre-Trained Transformer 4 (GPT-4),[63] uses publicly available data from the internet ("internet data") alongside third parties' licensed data to develop the foundation model's capabilities. However, a GenAI model's data needs do not end here.

The data involved in producing **models** can feed into human rights risks and harms associated with GenAI bias, output, and usage, necessitating scrutiny and caution. For instance, LAION-5B, a large dataset that has seen use in Midjourney and Stable Diffusion (GenAI models), was found to contain over a thousand images of child sexual abuse material.[64]

Besides gathering vast datasets for training GenAI model's capabilities, more curated data needs to be prepared for "fine-tuning". Fine-tuning refers to the enhancement of an AI model's capabilities at certain kinds of tasks and applications through the provision of particular "training data".[65] Thus, additional, smaller bodies of data that can contribute to improving model capabilities in specific ways are yet another requirement for GenAI development. The dataset preparation stage involves gathering data for the initial pre-training stage, and the fine-tuning that follows.

Eventually, models see **deployment,** when they are made available to a variety of industries and actors, leading to their adoption into various systems and services. The design and layout of AI services and their underlying models affect their scope of use and misuse, especially since various human elements (like GenAI developers) start intersecting heavily at this stage. For example, GenAI models

[63] Open AI (2024, March 4), "GPT-4 Technical Report", hosted by arXiv, Cornell University
[64] Buschek, C. and Jer Thorp, "Models All The Way Down", Knowing Machines
[65] OpenAI (n.d.), OpenAI Platform, "Fine-tuning"

have been found to be capable of producing sexualised content of women[66] and generating pornographic and homophobic material.[67] Research suggests that AI misuse begins at the data and modelling stage, but inevitably extends to the deployment stage, where outputs are generated for individual or group use.

This outline of the stages value chain provides an overview of the broad constituents of GenAI. Chapters 5, 6, and 7 discuss each stage, i.e., dataset preparation, modelling, and model deployment in detail. Un-risking GenAI with a gender lens can pinpoint sites of risk and, therefore, positive change – which can help transform GenAI.

Sub-components within the value chain

The secondary research and expert interactions led the study to a more detailed understanding of the GenAI value chain. This happened whilst diving into the broad stages of GenAI life cycles, particularly in the pursuit of gender un-risking. These sub-components have been articulated to magnify the three stages of GenAI's development and deployment into more distinct elements, with each piece reflecting a site for potential gender-positive change and care.

In addition to data, models, systems and services, this approach introduces new considerations in GenAI development and deployment. First, it tries to introduce how external stakeholders, who do not own or administer a given stage or sub-component of the GenAI value chain, can yet provide insight, feedback, scrutiny, alterations of labour for a particular tract of the value chain. For example, "AI red teaming" is a form of stress testing (to ensure resilience and robustness)[68] of a model to understand its ability and tendency to provide prohibited, dangerous, or harmful output.[69]

Second, this approach involves "policy" aspects that influence the development, deployment, and use of GenAI. Organisations can have policies on the use, abuse, and prohibited use of GenAI, that define boundaries demarcating what GenAI should or should not attempt. Such policies can affect the safety measures and decisions taken in building GenAI assets and processes, and collection of resources like data.

[66] Heikkila, M. (2022, December 12), "The viral AI avatar app Lensa undressed me – without my consent", *MIT Technology Review*
[67] Wiggers, K. (2023, July 21), "As AI porn generators get better, the stakes get higher", *TechCrunch*
[68] Chakravarty, A. (2010, January), "Stress Testing an AI Based Web Service: A Case Study", conference paper, ResearchGate
[69] Burt, A. (2024, January 12), "How to Red Team a Gen AI Model", *Harvard Business Review*

Putting it together: Human Rights, value chains, and risks

Chapters 3 and 4 attempted to anchor the inquiry in human rights risks using a value chain approach to unpacking AI development. With this understanding of human rights, and the AI value chain, the forthcoming chapters (5 - 7) explore the human rights and gender risks that the use of GenAI may pose, while linking the risks to the sub-components within the value chain. Following this risk analysis, the study pivots to mitigation mechanisms within these chapters and suggests sites (in the risk assessment toolkits) and measures for businesses to un-risk GenAI systems.

CHAPTER 5 | Dataset Preparation for GenAI

The **dataset preparation stage** is characterised by the initial, foundational training based on large amounts of data, like text and images. This stage is also called the 'pre-training stage', which denotes the routine pre-training of LLMs on billions of tokens and repetition of this process whenever new data becomes available.[70]

This data is typically sourced from third parties or exists in the form of internet data, entailing a process through which significant volumes of data are scraped from the internet. This data is often raw and is collected via a myriad of devices, from tactile sensors to system logs, used to record digital transactions of varying degrees. Internet searches, camera images, phone calls, social media posts, and credit card transactions are just some examples of these modes.[71] Such data can be collected from various publicly available datasets, code repositories, synthesised datasets, licensed data from corporations, and so on. These large swathes of data need to be customised and optimised to mature into datasets that can be used to develop LLMs that are the foundation for GenAI.[72]

The dataset preparation stage also comprises of extensive data collection and annotation labour force – which influences the kind of data that is ultimately fed into foundation models. Some organisations place special focus on female labour force

[70] Gupta, K., B. Thérien, A. Ibrahim, M.L. Richter, Q. Anthony, E. Belilovsky, I. Rish and T. Lesort (2023), "Continual Pre-Training of Large Language Models: How to (re) warm your model?", arXiv, Cornell University.

[71] Stanton, C., V. Lung, N. Zhang, M. Ito, S. Weber and K. Charlet (2019, August 5), "What the Machine Learning Value Chain Means for Geopolitics", Carnegie Endowment for International Peace

[72] Rey-Marston, M. and J.R. Lagunas (2024), "Supply chain networks in the age of generative AI: Turning promise into performance", Accenture

participation in mitigating gender bias and stereotypes in these existing Natural Language Understanding (NLU) models.[73]

Chapter 5 presents the human rights and gender risks this study has identified in the dataset preparation stage. The exploration of risks is followed by presentation of some mitigation pathways.

Unpacking dataset preparation

Before delving into Human Rights and gender risks and their potential un-risking, definition of the social and technical membership of dataset preparation is in order. The sub-components within the value chain (figure 3) help flesh out the dataset preparation stage.

Several different forms of data contribute to dataset preparation, becoming involved at different points in other stages of the value chain. The **pre-training data,** be it internet data or data licensed from some third party, helps develop GenAI model capabilities. **Training data** helps fine-tune specific capabilities of GenAI systems in desired ways. **Adversarial data** complements the training data by bolstering models' robustness against external misuse through methods like adding treated samples of adversarial uses to the training data processes.[74]

Data for GenAI refinement can also originate from the end-users: the people and groups using GenAI models. **Feedback data** is supposed to create GenAI feedback loops by collecting end-user inputs through various formats (for example, "thumbs up and thumbs down" ratings).[75] Data is thus involved in both building GenAI capabilities, as well as attempting to make models safer and more user-friendly.

Data for GenAI development involves labour, with people working to provide labelled, legible data that helps develop models' capabilities and responses. The workers and businesses involved in preparing data can affect the gender risks, as will be discussed further in this chapter.

[73] Karya (2023, March), "Collecting and annotating a corpora of 500,000 text sentences in 5 Indian languages for the Bill and Melinda Gates Foundation"
[74] Shekhar, R. (n.d.), "Tools for Responsible AI", INDIAai
[75] Government of the United Kingdom (2023, May 3) "AI Foundation Models: initial review

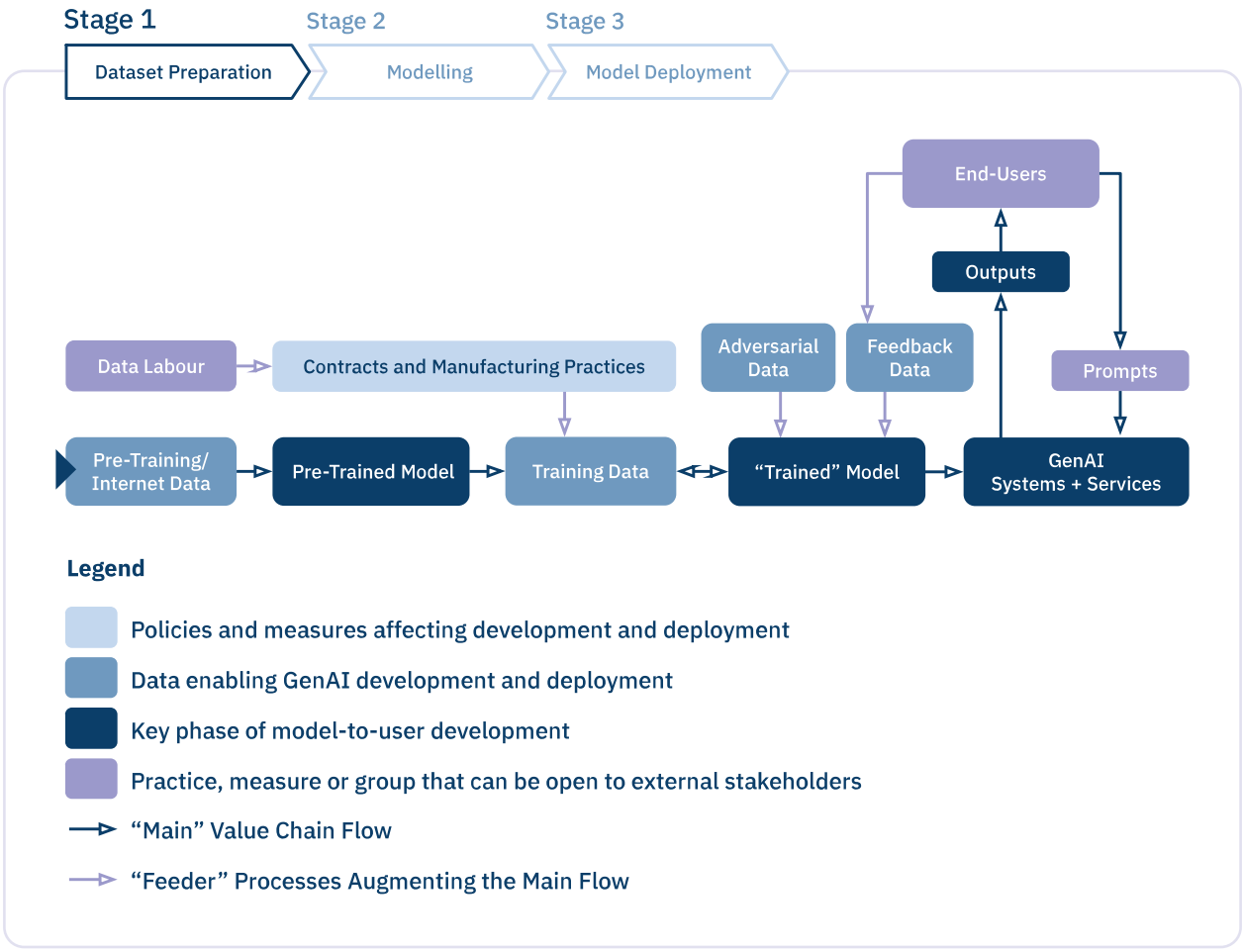


Figure 3. The GenAI Value Chain: The sub-components of the dataset preparation stage

Potential Human Rights and Gender Risks in Dataset Preparation

The dataset preparation stage presents specific risks to human rights from a gender lens. The risks relate to erasure of women and gender minorities from datasets (Risk 1), lack of representation from women in the data labour force (Risk 2), absence of guardrails for privacy and safety (Risk 3), a lack of awareness on data collection practices (Risk 4), and a lack of training in data collection and storage practices (Risk 5). These risks fall into two categories – Risks 1 and 2 emerge from the under-representation of women in dataset preparation, leading to consequences and harms from the AI at the deployment stage; Risks 3 - 5 are situations or consequences where women and gender minorities experience additional burdens and consequent human rights risks.

Risk 1: Erasure of women and gender identities from datasets

Women and gender minorities often risk being erased from datasets as there is a reliance on gender-blind design or on a binary understanding of gender. Various research efforts highlight that a digital divide exists between men and women's adoption of, affinity for, and usage of digital technology. Interactions with a data collection expert highlighted that women are often invisibilised or unaccounted for when collecting data from field surveys.[76] Moreover, depending on the socio-normative structures that exist in countries, representation of women is often limited in datasets due to their household roles.[77]

Efforts to go beyond the binary in the case of industry are nugatory and, when attempted, applied in limited contexts. 94.8% of research papers that focus on gender treat the concept as binary.[78] Dataset labels contain only 'male/female' or 'man/woman' options, without accounting for intersectionality.[79] Many studies also use the terms 'sex' and 'gender' interchangeably,[80] thus ***invisibilising a large section of women or gender minorities*** that may be categorised as non-binary as a result of both sexual orientation and gender identity.

Many Machine Learning (ML) techniques do not employ gender filters in pre-training data, whereas this practice can curb the erasure of women and non-cisgendered identities. Looking beyond erasure to affirmatively accommodate diverse gender identities, pre-trained data often brings a large variety of stereotypes with it. These stereotypes generate images that over-sexualise women and depict trans and non-binary identities as alien and explicit.[81] This leads to direct harm in wider society applications of AI services. While it was later found to have limited application applying only to the public-facing profile pages and news feeds,[82] Facebook's 2014 decision to add 56 gender options beyond the binary was a promising signal to change the standard programming of these platforms to prevent exclusionary practices.[83]

This risk is aggravated by lack of inclusivity in pre-training and training data that is procured from the internet and typically used

[76] Aapti primary research: Expert interaction with data expert
[77] Insight from key informant interview with data collection expert
[78] Keyes, O. (2018), "The Misgendering Machines: Trans/HCI implications of Automatic Gender Recognition", Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 88
[79] Ibid
[80] Ibid
[81] Ungless, E.L., B. Ross and A. Lauscher (2023), "Stereotypes and Smut: The (Mis)representation of Non-cisgender Identities by Text-to-Image models", arXiv preprint, arXiv:2305.17072
[82] Bivens, R. and O.L. Haimson (2016), "Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers", Sage Journals
[83] Ibid

to build the ability of the pre-trained model to perform the desired tasks or applications. The data labour force plays an instrumental role here: the people preparing these datasets must constitute a gender-diverse and sensitive workforce and be appropriately trained. Under Principle 14 of the UNGPs, businesses that may not be directly building datasets or engaging with the dataset preparation stages (for example model developers or deployers) retain a responsibility to protect human rights throughout their operations, including partnerships for dataset preparation. In the case of businesses connected more directly with the dataset preparation stages, regardless of their function, size, or context, Principle 13 of the UNGPs should be upheld – particularly to protect the rights of women and under-represented minorities when AI technologies are leveraged.

Further, this risk can translate into human rights abuses that intrinsically impact the rights of women and gender minorities and compromise their right to a life of dignity and equality. Under the Constitution of India, such abuses constitute violation of the fundamental rights of life, liberty, and security guaranteed under Article 21, which is complemented by Articles 14 and 15, mandating that all persons be treated equally before the law, and prohibiting discrimination on grounds of religion, race, caste, sex or place of birth. The lack of representation in datasets could lead to violation of the right to freedom of thought and expression accorded to all persons, which includes their right to express oneself without fear in public fora under Article 19(1)(a). The erasure and/or underrepresentation of women and gender minorities in datasets that often leads to stereotyping and targeted violence at the later stages of GenAI development invokes remedy under these constitutional rights. The Transgender Persons (Protection of Rights) Act, 2019 also elaborates on the prohibition of discrimination against a trans person. Erasure and invisibilisation of trans persons from datasets can manifest in denial of access to employment, education, healthcare, among other disadvantages.

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to life, liberty, and security	*This right recognises the personhood of women, trans and non-binary identities, thus making room for their inclusion in pre-training and training datasets.*	**International law:** UDHR: Arts. 3, 5, 7 | ICCPR: Arts. 6, 7, 16, 26 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21 | The Transgender Persons (Protection of Rights) Act, 2019: Section 4
Right to freedom of expression	*This right entitles women, trans and non-binary identities to express themselves without interference, and to impart information and ideas through any media.*	**International law:** UDHR: Arts. 18, 19 | ICCPR: Arts. 18, 19 | CEDAW: Arts. 1, 2 **Indian law:** The Constitution of India: Art. 19 (1)(a)
Right to equality before the law and protection from discrimination	*This right entitles women, trans and non-binary identities to be effectively protected against discrimination that comes with overlooking their identity in GenAI development.*	**International law:** UDHR: Arts. 2, 6, 7 | ICCPR: Arts. 3, 26 | ICESCR: Arts. 2, 3 | CEDAW: Arts. 1, 2, 7(c), 15 **Indian law:** The Constitution of India: Art. 14, 15 | The Transgender Persons (Protection of Rights) Act, 2019: Section 3
Duty to the community	*States and citizens have a duty towards enacting laws that promote the right of self-determination and ensure full development and advancement of women and gender minorities to participate in the inclusive development of GenAI.*	**International law:** ICCPR: Art. 1 | ICESCR: Art. 1 | CEDAW: Art. 3 **Indian law:** The Constitution of India: Art. 51 A(e)

Table 1. Human rights mapping for Risk 1 in dataset preparation

Risk 2: Lack of intentional gender representation in the data labourforce

A significant concern in the data stage emerges from the ***lack of intentional representation of women and gender minorities in the data collection, curation, and annotation stages.*** The collection of data through household surveys and other formats typically under-represents women and disregards gender minorities.[84] Moreover, datasets are often developed in a manner that excludes more data on women and gender-diverse identities. These datasets must not be constructed in an exclusionary manner, as elaborated below.

Many businesses and companies do not prioritise the intentional representation of women, and discount the inclusion of gender minorities in the data labourforce.[85] During the labelling and annotation phase, businesses typically categorise women and trans identities collectively and often non-intentional about women's representation in the project team.[86] The Gender Dimensions of Principle 11 of the UNGPs posits that businesses should intentionally create enabling environments and propel women and under-represented gender groups pursuing economic opportunities. Where such intentionality is missing, collective responsibility to include women and under-represented gender groups could fade over time. The same principle also states that businesses have a responsibility to not reinforce gender discrimination especially in relation to livelihood opportunities, a societal bias evident in the labour force participation figures of women as opposed to men.

Additionally, approximately 300 million fewer women than men access the internet, and women in low and middle income countries are 20 percent less likely than men to own a smartphone.[87] In combination, the under-representation of diverse gender groups in the data collection and annotation stages, and the gendered nature of digital literacy and digital usage could result in potential anti-civilisation of female and non-binary gender groups over time if not circumspectly monitored.

The absence of the role of gender in data collection has resulted in gender being deemed as "fixed", or assumed to be stagnant and based on physiology.[88] Just as focusing on the AI labourforce helps

[84] Aapti primary research: Expert interaction with data expert
[85] Key informant interview with a Gender and GenAI expert.
[86] Ibid
[87] Smith, G. and I. Rustagi (2021), "When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity", *Stanford Social Innovation Review*

to centre the important role played by human intervention in AI development, focusing on the representation of women and gender minorities helps to centre inclusivity through human intervention.[89]

The composition of the data labour force influences the quality and inclusivity during pre-training data refining, and how adversarial and feedback data are utilised to fine-tune the raw data. The absence of wider representation in the labelling workforce can have implications for users' enjoyment of the right to life and liberty under Article 21 of the Indian Constitution, to freedom of expression in the building of GenAI under Article 19(1)(a), and the right to equality before law and protection against discrimination for all genders in their access to GenAI services. Representation in the workforce mediates the right to life, liberty and security, which are determinants of the right to public participation and the right to find a voice in the public sphere. Similarly, representation also mediates the right to equality before law, and protection against discrimination can be affected by absence of diverse participation in the workforce — this assumes additional significance in contexts where the state is developing and deploying technology.

[88] Keyes, O. (2018), "The Misgendering Machines: Trans/HCI implications of Automatic Gender Recognition", Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 88

[89] Muldoon, J., C. Cant, B. Wu and M. Graham (2024), "A typology of artificial intelligence data work", Big Data & Society, 11(1), *Sage Journals*

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to life, liberty, and security	*This right guarantees that women and gender minorities be represented in data, so that they may participate and find a voice in the public sphere.*	**International law:** UDHR: Arts. 3, 5 | ICCPR: Arts. 6, 7, 16, 26 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21
Right to equality before the law and protection from discrimination	*This right enforces the belief that women and gender minorities, irrespective of their SOGIESC, are entitled to full enjoyment of the benefits of representation in GenAI applications.*	**International law:** UDHR: Arts. 2, 7 | ICCPR: Arts. 3, 26 | ICESCR: Arts. 2, 3 | CEDAW: Arts. 1, 2, 7(c), 15 **Indian law:** The Constitution of India: Arts. 14, 15

Table 2. Human rights mapping for Risk 2 in dataset preparation

Risk 3: Absence of guardrails for data privacy and user safety

Data privacy is an inescapable concern for GenAI models using consumer data.[90] An appraisal of open source datasets and data minimisation protocols highlights that a keen balance has to be struck when collecting data, gaining consent for it, and governing its dissemination and usage, to ensure that interventions such as anonymised data practices can be built upon to strengthen privacy and safety. This risk emerges in the data stage when pre-training data is inadequately filtered, leaving sensitive information exposed to the model to employ, often without the user's knowledge or consent.

A survey[91] in Europe showed that 45 to 60 percent of European respondents agree that AI will lead to more abuse of personal data. AI, data privacy, and user safety are inextricably interlinked and it is important that if a model requires private data, there are methods available to use it in secure and non-invasive ways.[92] Due to the nascency of GenAI and the diversity in adoption of digital technology in India, similar surveys are yet to be held in the country.

Preventing data misuse through self-created and -enforced responsible governance frameworks and guidelines is crucial in addressing such challenges. ***Compromising sensitive personal data can place women and gender minorities in a vulnerable position, leading to unwanted biases, confidential information leaks, or unfair and detrimental decision-making.***[93] Further, privacy breaches can include the misuse of medical information, sexual preferences, or location data, which can put an individual's privacy at risk, potentially exposing already vulnerable groups to greater danger and harm.[94]

Illustrative actions highlighted in the Gender Dimensions state that, under Principle 17 of the UNGPs, any gender-based violence should be treated with priority, triggering Principle 18, according to which businesses have a responsibility to conduct meaningful impact assessments to gauge areas of gender-specific violations. In the absence of privacy and security guardrails to protect sensitive and personal data, women and gender minorities are at risk of endangering their dignity and security.

[90] Von Gravrock, E. (2022, May 31), "Why artificial intelligence design must prioritize data privacy", World Economic Forum
[91] BEUC (2019), "Artificial Intelligence: what consumers say – Findings and policy recommendations of a multi-country survey on AI", BEUC
[92] Von Gravrock, E. (2022, May 31), "Why artificial intelligence design must prioritize data privacy", World Economic Forum
[93] Villegas-Ch, W. and J. García-Ortiz (2023), "Toward a Comprehensive Framework for Ensuring Security and Privacy in Artificial Intelligence", Electronics, 12(18), 3786, Multidisciplinary Digital Publishing Institute (MDPI)
[94] Ibid

Section 3 of the Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, recognises any information relating to physical, physiological and mental health conditions as well as sexual orientation as sensitive personal data. Under these rules, this data must be handled lawfully and responsibly by businesses. In cases where such data is not handled appropriately, breach of law could significantly impact businesses.

Under Sections 5 and 11 of the DPDP Act, accessing such sensitive information requires a notice to be sent to the user on the nature of the data and the purpose for which it is being processed, along with how such rights can be exercised and the manner in which any violation can be redressed. Datasets for GenAI must mandate consent that is free, specific, informed, unconditional, and unambiguous — under Section 6 of the Act.

Privacy allows for the undisputed protection of gender-sensitive data so that when women, and especially marginalised gender minorities, wish to volunteer information on their gender identity, they can do so without fear.

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to life, liberty, and security	*This right entitles women and gender minorities to a life of dignity, where they can enjoy the freedom of their sensitive data being handled with care and caution.*	**International law:** UDHR: Arts. 3, 5 | ICCPR: Arts. 6, 7, 26 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21
Right to privacy	*Privacy is an inalienable, fundamental right that includes the preservation of personal and sensitive data in a manner that protects user safety and anonymity, keeping in mind informed consent.*	**International law:** UDHR: Art. 12 | ICCPR: Art. 17 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21 | The IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011 | The DPDP Act, 2023: Sec 5, 6, 11

Table 3. Human rights mapping for Risk 3 in dataset preparation

Risk 4: Lack of awareness on inclusive data collection practices

Inclusive data collection practices are rare when interacting with stakeholders within this value chain. Questioning the origins and composition of a training dataset plays a crucial role in avoiding or exacerbating under-representation and misrepresentation of women and gender minorities. There is often a ***lack of awareness among technologists on how to create inclusive datasets or evaluate whether a particular dataset is influenced by AI fairness and ethics principles.***

The datasheet for a dataset, which is usually a document containing the demographic details that went into creating it, can reflect the intention and the composition of the dataset. This can act as a check on data quality standards. Data collectors and annotators are typically not mindful of the importance of gender dimensions in originating quality datasets.[95] This includes questions about who created the set, how it was annotated, who funded it, and what their intentions were.[96]

Guidance available from the Gender Dimensions of Principle 12 of the UNGPs states that businesses should be responsible for all business activities and processes and ensure gender equality. As dataset preparation plays a foundational role when developing and deploying models, building awareness on mitigating potential risks at this stage should be a priority for all business types building GenAI technology. While there might be a dearth of information, it remains the responsibility of business stakeholders to create equitable policies and conduct training for their personnel.

The importance of training the labour force vis-a-vis inclusive data collection practices lies in that it can create gender-centricity in both pre-training and training data used in the value chain. This risk mitigation aspect goes hand in hand with the representation of women and gender minorities in the data labour force.

The absence of inclusive datasets can hinder the extent to which women and gender minorities derive equal representation; thus, ensuring their presence is an affirmative step towards protection against gender-based discrimination as outlined under Articles 14

[95] Aapti primary research: Expert interactions with AI developers and deployers

[96] "A Critical Field Guide for Working with Machine Learning Datasets" (n.d.), Knowing Machines, last retrieved April 1, 2024

and 15 of the Constitution of India. There is a certain opacity to these datasets that opens up the conversation around women and gender minorities having the right to be informed of what constitutes ethical collection of their data and its processing, which invokes the right to information that every citizen has when debating their right to privacy and security. Taking reasonable steps to ensure that the data collected is accurate and does not contain any discrepancies is essential to processing information in a secure and responsible manner, as endorsed by Section 5 of the IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011.

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to equality before law and protection against discrimination	*This right entitles adequate representation of women and gender minorities in ethical data collection on the road to creating inclusive datasets.*	**International law:** UDHR: Arts. 2, 7 | ICCPR: Arts. 3, 26 | ICESCR: Arts. 2, 3 | CEDAW: Arts. 1, 2, 7(c), 15 **Indian law:** The Constitution of India: Arts. 14, 15
Right to information	*Users have the right to understand how their data is being collected, and the intention and composition of such datasets.*	**International law:** UDHR: Art. 19 | ICCPR: Art. 19 **Indian law:** The Constitution of India: Art. 19(1)(a) | The Right to Information Act, 2005: Sec. 3
Right to privacy	*Privacy is an inalienable, fundamental right that includes the preservation of personal and sensitive data in a manner that protects user safety and anonymity.*	**International law:** UDHR: Art. 12 | ICCPR: Art. 17 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21 | The IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information Rules, 2011

Table 4. Human rights mapping for Risk 4 in dataset preparation

Risk 5: Lack of training in data collection and storage practices

A consequence of the vast quantity of data being processed at this stage is that there is a ***lack of awareness and understanding of***

how data should be collected, stored, and processed by technology companies. Traditionally, labelling data has been a natural focus of research for ML tasks.[97] Semi-supervised learning includes model training that is a melange of a small amount of labelled and a large amount of unlabelled data. However, with the increase in volumes of data being used to train models, a long list of considerations emerges. Data management issues such as acquiring large and representative datasets, performing data labelling at scale, and improving the quality of large amounts of existing data become more relevant.[98]

The data labour force at this stage is often under-trained for these tasks due to a lack of awareness of good practices. They rarely receive special training on particular requirements for AI systems.[99] This ultimately leads to poor quality datasets that may adversely affect the accuracy and quality of AI services,[100] including but not limited to the collection of inclusive data encapsulating women and gender minorities.

This could be seen as acting against Principle 13 of the UNGPs, where the illustrative examples of the Gender Dimensions elucidate that it is the responsibility of businesses to provide advice, build capacity, and offer incentives to business partners and internal stakeholders. This risk further goes against Principle 16 of the UNGPs which states that it is the responsibility of businesses to dedicate specific funds to ensure implementation of gender-inclusive policies.

This lack of training has an additional impact on the value chain: the substandard quality of the datasets further leads to discrepancies in training data, and pre-trained and trained models — manifesting as harms when they translate into GenAI systems and services.

Beyond the inclusive collection of data, it is imperative that women and gender minorities be informed of how their data is processed. This is in line with respecting the right to liberty and security when it comes to sensitive personal information, as stipulated under Article 21 of the Constitution of India.

[97] Roh, Y., G. Heo and S.E. Whang (2019), "A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective", IEEE Transactions on Knowledge and Data Engineering, 33(4), 1328-1347, arXiv:1811.03402v2

[98] Ibid

[99] Muldoon, J., C. Cant, B. Wu and M. Graham (2024), "A typology of artificial intelligence data work", Big Data & Society, 11(1), *Sage Journals*

[100] Ibid

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to life, liberty, and security	*This right allows for users to seek protection from data collection practices that could compromise their personal information.*	**International law:** UDHR: Arts. 3, 5 | ICCPR: Arts. 6, 7, 16, 26 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21

Table 5. Human rights mapping for Risk 5 in dataset preparation

Mitigating risks in Dataset Preparation

The identified risks paint a picture of how elements, humans, and processes contribute to GenAI design. These risks, if left unchecked, impact an array of human rights as identified in the previous section and can also lead to further compounding effects. The development stage is characterised largely by the tasks of collecting, storing, and processing often sensitive and personal data for use by GenAI models and services. Therefore, solutions attached to this stage involve building trust, transparency, and inclusion within the dataset preparation ecosystem.

Proposed mechanisms for businesses to consider at the dataset preparation stage include - instantiating collaborative and diverse data collection practices, building open-source datasets, creating Small Language Models (SLMs), and employing focused social efforts to build trust.

PATHWAY	RELEVANT UNGP REFERENCED
Collaborative data collection practices	11
Open-source databases	15, 16
Building datasets for SLMs	14
Building trust	17, 18

Table 6. Consolidated recommendations for the dataset preparation stage

1. Instantiate collaborative, diverse, and inclusive data collection practices:

The need for a diverse and representative cohort of people collecting and labelling pre-training data is indispensable but is often missing from standard business practice. Businesses that work on 'humans-in-the-loop' technology advocate demographic diversity in the cohort of data collectors and annotators at this stage.[101] This diversity includes marital status, educational status, and whether the pool of candidates includes persons from economically weaker sections.[102] Differences in model type also prompt the question of which GenAI technologies require gender diversity and inclusion. For example, speech-based AI is often developed through gender-agnostic methods of collection, with its own benefits and harms, while image and text generation might require more consideration toward gender.

Businesses should work towards building collaborative and inclusive data collection practices, allocate resources for this and monitor progress to create GenAI applications that understand the nuances of the problem in a more inclusive manner.[103] Karya, a dataset-building organisation, has incorporated gender intentional design principles in its data collection practices to ensure that 50 percent of the cohort is always female. This mandate does not change with the gender-centric nature of any project.[104] OpenNyAI, another initiative that provides AI development tech and know-how to other players in the ecosystem, labels the data for its legal justice chatbot, Nyaayabandhu, by involving law students in validating the datasets.[105] Such practices ensure that models are equipped with understanding and are representative of the larger portion of the demographic they are meant to cater to. Businesses should actively work towards bringing about this gender inclusivity, as laid down in Principle 11 of the UNGPs.

A study by the University of Georgia found that utilising mixed training datasets has the potential to improve equality and fairness in AI tools, specifically in sectors such as education, where fairness occupies an important place in evaluation.[106] The study compared mixed datasets (relating to both men and women), and found that balanced training procedures offer a workable solution to the

[101] Key informant interview with 2 Gender and GenAI experts
[102] Key informant interview with ecosystem expert working in a CSO and AI development organisation
[103] Key informant interview with experts from CSO organisation and GenAI development organisation
[104] Key informant interview with expert from CSO organisation
[105] Key informant interview with expert with GenAI/AI system developer
[106] Latif, E., X. Zhai and L. Liu (2023), "AI Gender Bias, Disparities, and Fairness: Does Training Data Matter?", arXiv preprint, arXiv:2312.10833

widespread malady of AI systems reproducing societal prejudice.[107] This is further proof of how existing dataset construction processes are representative of cultural norms and stereotypes.[108] It also paves the way for the understanding that gender-based representation in the value chain can lead to differential outcomes for a model at the deployment stage.

Moreover, involving gender minorities and non-binary gender identities in this process can bring about broader questions of representation and inclusivity in datasets as well. While it can often be difficult to define gender representation in datasets, working towards making data universally applicable by using gender-neutral language as much as possible, is a much needed starting point for businesses to adopt. Gender sensitisation exercises and foundational learning about gender for the data collection workforce are essential for inculcating this collaborative environment.

2. Prepare Open-source databases:

Processes and the type of data used for AI systems are often opaque and are scraped from private datasets. While disclosing proprietary information about processes might be tougher for businesses, bringing transparency around datasets, or at least their nature and origin, could help increase user trust. Open-source (OS) AI is a shift from the black box models that are often seen circulated in the GenAI ecosystem. OS codes are a way to mitigate discrimination and bias in foundation models due to the convergence of AI and transparency of code.[109] OS data is built on contributions of the broader developer community and are available for anyone to download and utilise. This ensures that there is an expanding user base for a dataset along with developers trusting the business due to the collaborative nature of the dataset.[110] In turn, this diversity in building datasets invites gender sensitivity through a broader array of inclusive language, user engagement, and code contributions.[111]

While building datasets, businesses should have policies in place that govern collection, consent seeking, and sharing of utilisation-related information with data principals. This is

[107] Ibid
[108] Myers West, S. (2019), "Discriminating Systems: Gender, Race, and Power in AI – Report", AI Now Institute, last retrieved April 1, 2024
[109] Theben, A., L. Gunderson, L. López-Forés, G. Misuraca, and F. Lupiáñez-Villanueva (2021), "Challenges and limits of an open source approach to Artificial Intelligence," Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, last retrieved April 1, 2024
[110] Sahu, A. (2022, August 17), "Why making your product's code free is a competitive advantage", World Economic Forum, last retrieved April 1, 2024
[111] Ball, B. (2023, May 31), "Creating Safe Spaces for Underrepresented Individuals in Open Source Communities", Meta, last retrieved April 1, 2024

reflected in Principle 15 of the UNGPs which urges businesses to be responsible for human rights implementation by having policy commitments in place to identify and mitigate adverse human rights issues. Principle 16 also suggests embedding these responsibilities through a statement of policy that is informed, publicly available, and reflected in operational policies.

3. Build datasets for Small Language Models (SLMs):

LLMs typically require immense computing power and are expensive to operate but continue to be favoured when building GenAI models. With GenAI's potential to address a diverse spectrum of use cases, defining its purpose and the effort required to achieve this purpose could be key. As part of the primary research explorations conducted for this study, SLMs have proved to be feasible alternative to LLMs. Though SLMs operate on a smaller scale, they are less likely to generate bias or offensive text.[112] SLMs can be open-source, allowing for greater degrees of transparency and customisation,[113] and are also trained on smaller, curated datasets, making them more interpretable. Depending on the purpose, **businesses should consider investing in leveraging SLMs to improve transparency and reduce operating costs.** Even if using SLMs, it remains imperative for businesses to continue upholding human rights, particularly of those who have been disproportionately and historically impacted. This is endorsed by Principle 14 of the UNGPs which holds businesses responsible for respecting human rights irrespective of their size, operational context, and sector — which includes all kinds of GenAI manufacturers building foundation models.

Given these benefits, existing studies usually opt for a relatively small gender-neutral database for their models.[114] This reduces the cost and complexity of building a dataset that does justice to the rights and lived experiences of women and gender minorities. However, most general-purpose AI (GPAI) requires large datasets that are beyond the stage of gender-sensitive curation to improve representation in existing datasets or strengthen gender representation when building new datasets. Ultimately, SLMs merely proffer an alternative for specific use cases.

[112] Benny Mathew, A. (2023, October 30), "Demystifying the Benefits of Compact Language Models for Specific Business Needs", E2E Cloud, last retrieved April 1, 2024
[113] Ibid
[114] Fatemi, Z., C. Xing, W. Liu and C. Xiong (2021), "Improving Gender Fairness of Pre-Trained Language Models without Catastrophic Forgetting", arXiv preprint, arXiv:2110.05367

4. Work on establishing trust:

For the general public to place faith in any GenAI software, there must be a process of trust-building embedded. Principles 17 and 18 of the UNGPs provide for a due diligence process that involves internal and external stakeholders assessing actual and potential human rights impacts, in this case on women and gender minorities, along with integration and acting upon such findings through operational and structural changes. This can help protect against potential human rights risk and requires ensuring transparency, accountability, and ethical considerations in AI development.

Confusion often reigns in the ecosystem regarding the entity responsible for building transparency and accountability in GenAI systems: the state, the businesses, or a third party.[115] In such instances, **investment in an implementable and trackable GenAI governance framework built by businesses and shared publicly is recommended.** This framework should operate on the pillars of resilience, inclusion, explainability, transparency, and performance.[116] Even if the dataset created is utilised for the public good, maintaining transparency is key to strengthen trust in the system. Clearview AI, a facial recognition software maker, was caught in a lawsuit over its facial recognition database. Though it was selling its software to the local police and government agencies in the US, its technology was deemed illegal in Canada, Australia, and other parts of Europe due to its database of over 20 billion facial photos scraped from the internet and popular social media sites such as Facebook, LinkedIn, and Instagram.[117] While Clearview's datasets have helped law enforcement in its work, its use of facial recognition with loose, if not absent, privacy considerations led to it being faced with multiple lawsuits.[118] Examples of GenAI models emerging from India are still nascent and, according to this EY report, Indian developers have only recently graduated from demo development to proof of concept, with a few released GenAI products.[119] Given the pace of this technology's evolution, we can hope to learn from various GenAI models in the near future.

This makes the case for trustworthy GenAI systems moving to allow the lived experiences of women and gender minorities to

[115] Key informant interview with ecosystem expert in gender and GenAI technology.
[116] Godhrawala, A. and A. Kumar (2024, January 5), "How GenAI governance framework can help build trust in tomorrow's tech", EY, last retrieved April 1, 2024
[117] Mac, R. and K. Hill (2022, May 9), "Clearview AI settles suit and agrees to limit sales of facial recognition database.", *The New York Times*, last retrieved April 1, 2024
[118] Hill, K. (2023, June 21), "Clearview AI, Used by Police to Find Criminals, Is Now in Public Defenders' Hands", *The New York Times*, last retrieved April 1, 2024
[119] Ernst & Young India, (n.d.), "The AIdea of India: Generative AI's potential to accelerate India's digital transformation"

permeate their current black box habitat. In the Indian context, there is a lack of instances of that bring to light emerging concerns around regulating GenAI. The regulatory framework for governing GenAI services is still in its nascency. The Indian government's response to Google's Gemini and Ola's Krutrim has led to increasing scrutiny and discussions on how these platforms can be regulated to deal with misinformation.[120]

To prevent trust breakdown and foster accountability, businesses need to monitor themselves for any human rights abuses stemming from their work. Principle 17 of the UNGPs urges businesses to conduct due diligence to identify, track, and mitigate activities that can impact human rights.

CHAPTER 6 | GenAI Models

The characteristic element of the modelling stage is the creation of 'foundation models', sometimes called general-purpose AI or GPAI.[121] This term was coined by Stanford's Center for Research on Foundation Models, to denote AI models trained by multiple datasets for a set of diverse tasks.[122] These models are capable of various general tasks ranging from text synthesis to audio and image generation. OpenAI's GPT-3 and GPT-4 are foundation models that work at the backend for ChatGPT.

In this chapter, on the modelling stage of the GenAI value chain, the study details its identification of human rights and gender risks relevant to models.

Unpacking the modelling stage:

The actions and choices of developers while building the models shape the human rights risks and biases that may manifest later during use of the GenAI technology.[123] In addition to data and training, there is also stress testing, finding ways to measure model performance in various tasks, and the development of safety mechanisms to curb GenAI's malicious use and manifestation of risks as gender-based harms.

[120] Agrawal, A. (2024, March 2), "Under-testing AI models must get govt permission before deployment: MeitY", *Hindustan Times*

[121] Jones, E. (2023, July 17), "Explainer: What is a foundation model?", Ada Lovelace Institute, last retrieved April 1, 2024

[122] ITI (2023), "Understanding Foundation Models & The AI Value Chain: ITI's Comprehensive Policy Guide", Information Technology Industry Council Report, last retrieved April 1, 2024

[123] IBM Data and AI Team (2023, October 16), "Shedding light on AI bias with real world examples"

This second stage of the GenAI value chain features models in two subsequent states:

- **The pre-trained model** and the capabilities it acquires from being trained with large amounts of data from sources like the internet.
- **The trained model** that is prepared by receiving data curated to develop its capabilities in specific ways.

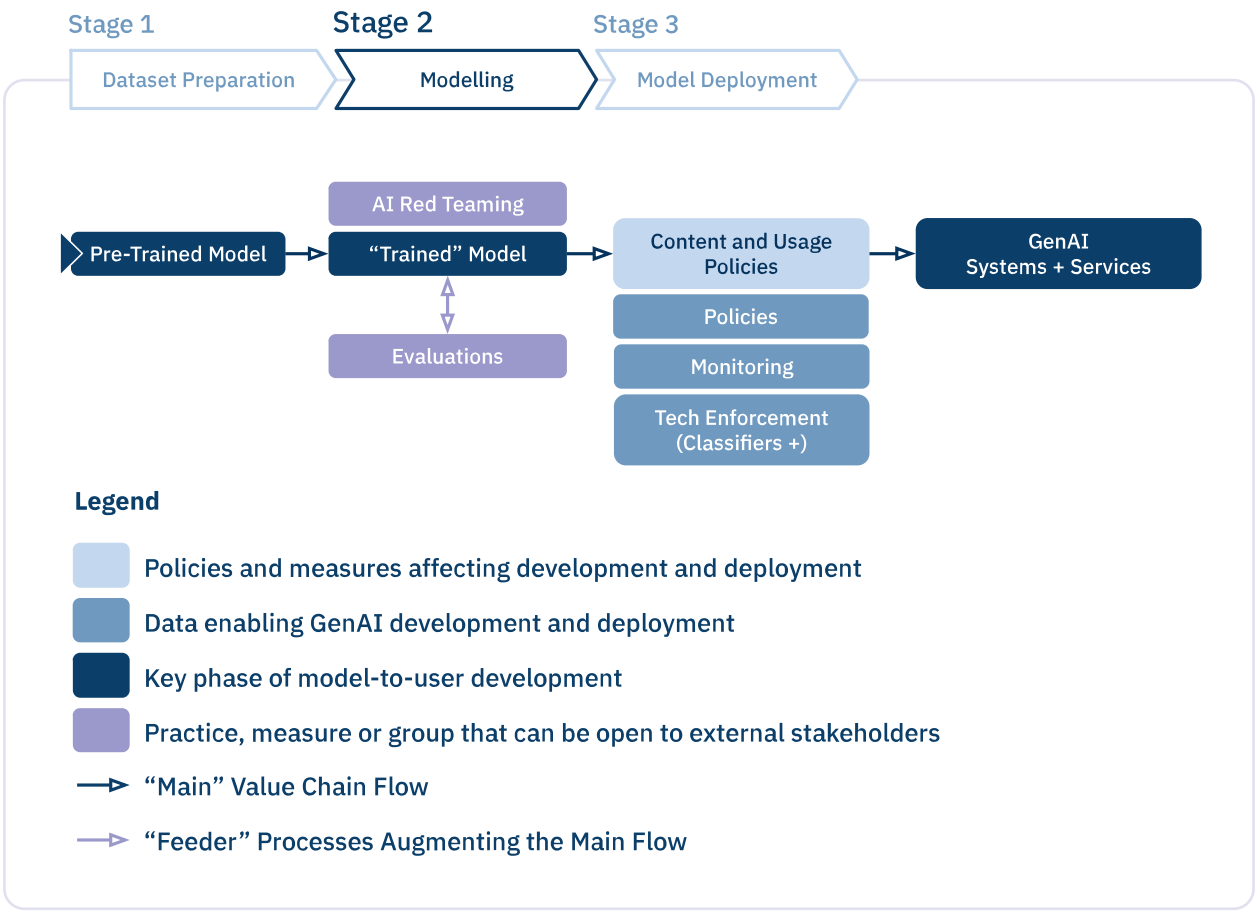


Figure 4. The GenAI Value Chain: The sub-components of the modelling stage

The modelling stage includes several stages of testing. Developed models are tested for efficacy through AI red-teaming processes that involve stress testing to gauge a model's ability and tendency to provide prohibited, dangerous, or harmful outputs.[124] AI red teams can also allow stakeholders outside the organisation that own or administer a value chain component, to contribute to making GenAI less harmful.[125] Evaluation, another space for testing

[124] Burt, A. (2024, January 4), "How to Red Team a Gen AI Model", *Harvard Business Review*
[125] OpenAI (2023, September 19), "OpenAI Red Teaming Network"

model capabilities, is "the process of validating and testing the outputs that your LLM applications are producing".[126] Evaluations provide developers and adjacent parties the opportunities to gauge models' performance in responding to tasks.

Beyond understanding models' limits to respond well and to respond safely, the modelling stage focuses on putting safety features and behaviours in place. To this end, models are governed by content and usage policies (CUPs) that define harmful and prohibited usage of GenAI models and services, and lay down the terms of use for models. CUPs can be operationalised into technological measures that are part of the model development to help prohibit specific forms of GenAI use. For example, if a model's policies state that seeking explicit, sexual material is prohibited, this declaration would entail the inclusion of technological measures, like classifiers, that prevent the production of sexual, homophobic, or harmful stereotype-enforcing content. CUPs can help make GenAI safer and less prone to harm and abuse by:

- Informing model evaluations
- Aiding the development of classifiers to find and track harmful content
- Strengthening efforts towards monitoring model use
- Developing a model's capacity for refusing prompts for particular kinds of tasks.

Monitoring is a crucial part of GenAI safety efforts and is present in models as well as services built on them. It refers to observation of the use of GenAI for preventing and stopping misuse of the model. Both humans and automated mechanisms can be involved in monitoring and responding to misuse.[127] OpenAI makes use of both to review material that goes against its CUPs, to mitigate its use, and to improve the firm's policies and how they are upheld.[128]

Thus, data helps train models' capabilities and measures like AI red teaming, evaluations, CUPs and their technical implementation, and monitoring builds GenAI systems' resilience against abuse, misuse, and production of harmful material. However, despite these measures — which are mainly oriented towards enhancing efficacy and efficiency, and preventing known

[126] Ziv, R. and S. Anadkat (2024, March 21), "Getting Started with OpenAI Evals", OpenAI Cookbook
[127] OpenAI (2023, March 23), GPT-4 System Card
[128] Ibid

harms — choices and actions during the modelling stage may still produce risks to normative values, and other socio-technological unknowns.

The following sections explore the human rights and gender risks present across the modelling stage. These risks include biased design assumptions (Risk 1), lack of gender-diverse participation (Risk 2), lack of gender-sensitisation training for GenAI developers (Risk 3), absence of a uniform framework for responsible GenAI construction (Risk 4), and inadequate screening of sensitive and personal data when training models (Risk 5).

Risk 1: Design assumptions restricted to cis-male or binary gender notions; gender blind design

Design assumptions in GenAI are often human-made choices influenced by a person's conscious or unconscious biases and decision-making. Design assumptions can emerge through algorithmic or cognitive bias, or both. ***Adopting a gender-blind notion could in turn lead to a reduced understanding of harms that might occur for women and gender minorities.*** Guidance provided by the Gender Dimensions related to Principle 11 of the UNGPs urges businesses to place more emphasis on human rights of gender identities, to reduce gender-based harms and employ a 'range of measures to ensure equal representation and participation' of historically under-served population groups.

Another form of bias, algorithmic bias, is the manifestation of flawed training data that repeatedly produces errors and unfair outcomes. This slant can also be caused by programming errors, when developers weigh in on algorithmic decision-making based on their personal biases.[129] This leads to further cognitive bias, where people process information and make judgements on the choice of information dissemination through their often skewed opinions favouring certain datasets over others.[130]

Design assumptions become a risk when steps are not taken to ensure the mitigation of these often skewed assumptions which can seep into GenAI models. To this end, the U.S. Department of Commerce's National Institute of Standards and Technology (NIST) recommends widening of the scope of the sources of these biases

[129] IBM Data and AI Team (2023, October 16), "Shedding light on AI bias with real world examples
[130] Ibid

beyond ML data and training to include societal factors and human intervention which is accompanied by its own set of influences.[131]

Several businesses argue for gender-blind design principles when creating GenAI services. While these services might obviate some human rights abuses, they may ultimately be unhelpful in representing women and gender minorities, and strengthen existing assumptions about gender being strictly binary. Moreover, most of these models do not usually provide explanations on the decision-making process they employ, and are inherently complex, "black box models".[132] In general, GenAI models suffer from a lack of explainability, with weak evaluation frameworks for error mitigation.[133]

The risk of biased design assumptions can manifest in both pre-trained and trained GenAI models. Content and usage policies that address high-risk categories and have clauses on gender-based violence, risks, and harms which are unacceptable for the model to engage in, can allow for these design assumptions to reduce and eventually be removed.

Altering design principles to expand the understanding of gender as strictly binary violates Article 21 of the Indian Constitution, set in place to preserve the dignity and rights of women and gender minorities. Moreover, the issue of representation is also dealt with by constructing models that obligate all genders receiving equal protection of the law, invoked through Articles 14 and 15.

[131] "There's More to AI Bias Than Biased Data, NIST Report Highlights" (2022, March 16), NIST, last retrieved April 1, 2024
[132] Hassija, V., V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud and A. Hussain (2023), "Interpreting Black-Box Models: A Review on Explainable Artificial intelligence", Cognitive Computation, 16, 45–74, Springer
[133] Ibid

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to life, liberty, and security	*This right guarantees the adequate representation of the lived experiences of women and gender minorities in the design of GenAI models.*	**International law:** UDHR: Arts. 3, 5 | ICCPR: Arts. 6, 7, 16, 26 | ICESCR: Art. 3 | CEDAW: Art. 1 **Indian law:** The Constitution of India: Art. 21
Right to equality before law and protection against discrimination	*This right entitles women and gender minorities to adequate representation such that it is equally indispensable in the GenAI modelling stage.*	**International law:** UDHR: Arts. 2, 7 | ICCPR: Arts. 3, 26 | ICESCR: Arts. 2, 3 | CEDAW: Arts. 1, 2, 7(c), 15 **Indian law:** The Constitution of India: Arts. 14, 15

Table 7. Human rights mapping for Risk 1 in GenAI models

Risk 2: Lack of gender-diverse participation in AI model development

AI models may be overwhelmingly narrow in their understanding of the gender spectrum and its various nuances. This stems from the ***limited representation of women and gender minorities in the workforce, particularly at the development stage.*** The Gender Dimensions of Principle 11 of the UNGPs place the responsibility of providing enabling environments for non-majority genders and gender equality in economic opportunities on businesses — and this constitutes an area where businesses could be lacking. A 2020 World Economic Forum report found that women occupy only 26 percent of data and AI positions in the global workforce.[134] This phenomenon is also present in ancillary industries. For instance, in 2022, only one in four researchers publishing on AI worldwide was a woman.[135] Additionally, the 2021 AI Index Report by the Stanford Institute for Human-Centered AI found that women constitute 16 percent of tenure-track faculty dedicated to AI in the global workforce.[136]

These figures represent a deeper issue relating to the detrimental effects of the lack of gender participation in the GenAI sector. Coupled with the gender digital divide that skews access to digital awareness and education, this gender disparity creates uneven distribution of power and leadership in the AI sector.[137] This, in turn, affects how much of a role representation plays in datasets and algorithmic products.[138] Not embedding this representation at the modelling stage can have far-reaching effects on not just the trained models, but also on red-teaming and monitoring efforts.

The rights mapping of this risk explains a pertinent point about how a gender-inclusive workforce in the modelling of GenAI systems can allow women and gender minorities to have their dignity, equality, and expression validated on larger platforms fuelled by GenAI. The intentional and unintentional consequences of under-representation or non-representation violate Articles 14, 15, 19, and 21 of the Indian Constitution. In situations where these risks are not addressed, the right to accurately represent, to become an equitable member, are not fulfilled, and the inability to contribute meaningfully to the larger digital ecosystem can impact basic human rights.

[134] Global Gender Gap Report 2020, World Economic Forum, last retrieved April 1, 2024
[135] Gangadharan, S. (2024, February 25), "Need more women in AI for inclusive development of the technology", *The Economic Times*, last retrieved April 1, 2024
[136] Artificial Intelligence Index Report 2021, Stanford University Human-Centered Artificial Intelligence, last retrieved April 1, 2024
[137] Ramos, G. (2022, August 22), "Why we must act now to close the gender gap in AI", World Economic Forum, last retrieved April 1, 2024
[138] Ibid

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to life, liberty, and security	*This right entitles women and gender minorities to be equitably represented in the GenAI workforce.*	**International law:** UDHR: Arts. 3, 5 | ICCPR: Arts. 6, 7, 16, 26 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21
Right to equality before law and protection against discrimination	*This right allows for women and gender minorities to have a say in GenAI modelling, as well as a voice in the way in which inclusivity can be meaningfully inculcated.*	**International law:** UDHR: Arts. 2, 7 | ICCPR: Arts. 3, 26 | ICESCR: Arts. 2, 3 | CEDAW: Arts. 1, 2, 7(c), 15 **Indian law:** The Constitution of India: Arts. 14, 15
Right to freedom of expression	*Women and gender minorities are guaranteed the freedom to express themselves and their differential lived realities in GenAI modelling.*	**International law:** UDHR: Arts. 18, 19 | ICCPR: Arts. 18, 19 | CEDAW: Arts. 1, 2 **Indian law:** The Constitution of India: Art. 19
Duty to the community	*The State has a duty to its citizens to ensure that technology built for public use includes diverse gender representations.*	**International law:** UDHR: Arts. 3, 5 | ICCPR: Arts. 6, 7, 16, 26 | ICESCR: Art. 3 | CEDAW: Art. 1 **Indian law:** The Constitution of India: Art. 21

Table 8. Human rights mapping for Risk 2 in GenAI models

Risk 3: Lack of gender-sensitive training for GenAI developers

In the contemporary GenAI ecosystem, due to insufficient collaboration between developers and researchers in the field, and the low numerical strength of women in the AI workforce, ***GenAI models often lack diversity in design.***[139] As the role of developers is critical to this system, the understanding of diversity is an important factor. The lack of diversity results in AI systems mirroring a certain homogeneity that lacks inclusivity and further intersectionality.[140] For instance, security robots are primarily male, but service and sex robots are represented as primarily female.[141] Alternatively, female gendering increases a bot's perceived humanity and acceptability in the ecosystem, because of the notion

[139] Shams, R.A., D. Zowghi and M. Bano (2023), "AI and the quest for diversity and inclusion: a systematic literature review", *AI and Ethics*

[140] Ibid

[141] Kumar, S. and S. Choudhury (2022), "Gender and feminist considerations in artificial intelligence from a developing-world perspective, with India as a case study", Humanities and Social Sciences Communications, 9(1), 31, *Nature*

that female AI is more humane and reliable.[142] The presence of LGBTQIA+ communities in the design of AI and robots remains largely unexplored, though prioritisation of this could bolster inclusive design.[143]

The illustrative actions recommended in the Gender Dimensions of Principle 12 of the UNGPs require businesses to invest and build capacity of personnel and business partners and, if not prioritised, could result in human rights risks. The Global Dialogue on Gender Equality and AI organised by UNESCO highlighted that third-party intersectional evaluations of AI models are still limited to the research community and not adopted for consumer-facing applications (or for developers).[144] Most training and skilling programmes are focused more on improving the quality of STEM access for women and girls. While this is indispensable to increase representation, it is seldom coupled with conversations and interventions around upskilling and gender-sensitive training and awareness for established AI developers. Integrating Diversity, Equity, and Inclusion (DEI) principles into AI and GenAI has the potential to mitigate challenges arising through fairness and bias concerns.

Human rights abuses that emerge when an entire ecosystem repudiates the inclusion of women and gender minorities can affect their right to live with dignity, afforded to them under Articles 14, 15, and 21 of the Indian Constitution. Additionally, the IT (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 also provide, under Section 3(1)(b)(ii), that due diligence must be conducted to ensure that intermediaries take responsibility to notify users not to host, display, publish, store, or share any information that is defamatory, obscene, pornographic, invasive of another's privacy, and harassing on the basis of gender. Clause (x) of the same provision also includes within its ambit information that is false and untrue, written with the intent to harass or mislead a person, or to cause injury to a person. Violation of this provision invites developers to take responsibility for the sensitivities of gender-based content in their AI models and services.

[142] Ibid

[143] Poulsen, A., E. Fosch-Villaronga and R.A. Søraa (2020), "Queering machines", Nature Machine Intelligence, 2(3), 152, *Nature*

[144] Nandi, A. (2024, March 7), "From data to deployment: Gender bias in the AI development lifecycle", Observer Research Foundation, last retrieved April 1, 2024

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to life, liberty, and security	*This right urges GenAI developers to look at gender sensitivity in training their models.*	**International law:** UDHR: Arts. 3, 5 | ICCPR: Arts. 6, 7, 16, 26 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21
Right to equality before law and protection against discrimination	*This right protects persons that are affected by GenAI developer businesses that overlook gender dimensions in their models.*	**International law:** UDHR: Arts. 2, 7 | ICCPR: Arts. 3, 26 | ICESCR: Arts. 2, 3 | CEDAW: Arts. 1, 2, 7(c), 15 **Indian law:** The Constitution of India: Arts. 14, 15
Duty to the community	*States are duty bound to mandate businesses to undertake gender-sensitive training programmes while developing foundation models.*	**International law:** ICCPR: Art. 1 | ICESCR: Art. 1 | CEDAW: Art. 3 **Indian law:** The Constitution of India: Art. 51 A(e) | The IT (Intermediary Guidelines and Digital Media Ethics Code) Rule, 2021: Sec. 3(1)(b)(ii)

Table 9. Human rights mapping for Risk 3 in GenAI models

Risk 4: Absence of a uniform framework for inclusive GenAI construction

The lack of training programmes for gender sensitivity is also a result of the general ***absence of operationalisation of AI principles.***[145] ***The dearth of both guidelines and actionable steps for concrete implementation of fairness and transparency principles has hindered gender-responsiveness in GenAI.***[146] Existing literature reveals that AI projects do not consistently or adequately take into account concerns relating to bias, equity, diversity, and inclusion.[147] This can be attributed primarily to a lack of practical and customisable set of tools to operationalise these principles and guidelines through checklists, definitions, design pattern templates, questionnaires, and requirement guidelines.[148] There is also confusion or ambiguity about who is responsible for DEI in AI development processes — regarding both overall responsibility and oversight of DEI considerations in any AI project.[149] The Gender Dimension illustration provided in Principle 15 of the UNGPs places responsibility for creating gender

[145] UNESCO (2020), "Artificial Intelligence and Gender Equality: Key findings of UNESCO's Global Dialogue", UNESCO Digital Library
[146] Ibid
[147] Zowghi, D. and F. da Rimini, F. (2023), "Diversity and Inclusion in Artificial Intelligence", arXiv preprint, arXiv:2305.12728
[148] Ibid
[149] Ibid

frameworks, and their implementation, on businesses. The detrimental effects of not having uniform frameworks can lead to skewed effects across the ecosystem, where due diligence appears to differ for different businesses.

Several expert interviews yielded a similar conclusion: the AI workforce is not designed to consider diverse perspectives.[150] Co-creator diversity in the process is important, as well as building these frameworks through appraisal of existing AI models and their learnings.[151]

The absence of frameworks to map gender biases and design flaws can hinder the efficacy of evaluations and red teaming efforts to ensure fairness and accountability in GenAI models. Frameworking inclusive participation in GenAI affirms the principles of equity and inclusion, forming the core of the right to life under Article 21 of the Indian Constitution and to equality before the law under Articles 14 and 15. Equity allows women and gender minorities to redress the consequences of skewed datasets, and inclusion is necessary for them to sustain a life of dignity.

[150] Key informant interview with ecosystem expert
[151] Key informant interview with ecosystem expert

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to life, liberty, and security	*The absence of a uniform framework can hinder the growth of DEI principles in GenAI models, causing skewed representation of women and gender minorities.*	**International law:** UDHR: Arts. 3, 5 | ICCPR: Arts. 6, 7, 16, 26 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21
Right to equality before law and protection against discrimination	*This right allows for women and gender minorities to demand that GenAI models live up to a standard of DEI recognition.*	**International law:** UDHR: Arts. 2, 7 | ICCPR: Arts. 3, 26 | ICESCR: Arts. 2, 3 | CEDAW: Arts. 1, 2, 7(c), 15 **Indian law:** The Constitution of India: Arts. 14, 15

Table 10. Human rights mapping for Risk 4 in GenAI models

Risk 5: Inadequate screening of sensitive information and vague informed consent principles

Principle 12 of the UNGPs, along with the illustrative examples provided in the Gender Dimensions, leverages international human rights models to place the responsibility of collecting meaningful consent with businesses. While meaningful consent can be tough to acquire, businesses must consider the importance of seeking consent from historically marginalised population groups as numerous models are trained using personal or sensitive data. In an effort to improve security and customised solutions, ***GenAI systems create human rights risks by collecting personal data, often without consent.*** With increased data collection mechanisms and processes becoming a norm to provide more customised solutions, data practices often disregard data minimisation protocols.

Training models without the ability or understanding to process personal data can lead to privacy concerns that manifest in the breakdown of trust in these models. Content and usage policies must reflect terms that address the screening of sensitive personal data and data provenance for how the data is to be utilised. Evaluation and monitoring of these models must ensure that data minimisation is adopted in handling sensitive information like sexual orientation and gender identity.

Screening sensitive information is essential to respecting the privacy of women guaranteed under Article 21 of the Indian Constitution, and for gender minorities to be able to access GenAI systems without fear of risking sensitive data or other facets of their identity. It also goes towards ensuring that women and gender minorities are able to enjoy the security of their personal data being used responsibly, as laid down under Sections 3 and 4 of the IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011.

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to privacy	*Privacy is an inalienable, fundamental right that includes the preservation of personal and sensitive data in a manner that protects user safety and anonymity.*	**International law:** UDHR: Art. 12 | ICCPR: Art. 17 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21 | The IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011
Right to life, liberty, and security	*This right entitles women and persons of diverse SOGIESCs to a life of dignity, where they can enjoy the freedom of their sensitive data being handled with care, caution, and informed consent.*	**International law:** UDHR: Arts. 3, 5 | ICCPR: Arts. 6, 7, 16, 26 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21
Duty to the community	*States are duty bound to mandate businesses to put mechanisms in place that can screen sensitive data and private information.*	**International law:** ICCPR: Art. 1 | ICESCR: Art. 1 | CEDAW: Art. 3 **Indian law:** The Constitution of India: Art. 51 A(e)

Table 11. Human rights mapping for Risk 5 in GenAI models

Pathways for mitigating risks

To address gender biases at the modelling stage, developers have evolved varied strategies, including benchmarks, red teaming and stress-testing, Reinforcement Learning from Human feedback (RLHF) and Constitutional AI (CAI), jailbreaks, and profanity filters, along with gender-sensitive training and awareness programmes, inclusive design guidelines, and overarching representation in GenAI developer communities. GenAI Developers also follow a series of tests and benchmarks to maintain the accuracy of LLMs and to avoid hallucinations that may ensue at the deployment stage. While the un-risking approach as a mitigation pathway is a good start to reducing human rights-related risks, larger ecosystem alignment, knowledge sharing, collaborative efforts, and value chain thinking are crucial overall initiatives.

PATHWAY	RELEVANT UNGP REFERENCED
Gender-sensitive training and awareness	16
Benchmarks and evaluation technique to test inclusivity	19
Red-teaming and stress-testing GenAI models	13, 15, 17
RLHF and constitutional AI	15, 16, 21
Profanity and privacy filters in models	11, 18
Design guidelines for gender-sensitive models	16, 18, 20
Representation of gender in leadership roles at big tech companies	11, 12

Table 12. Consolidated recommendations for the GenAI modelling stage

1. Conducting Gender-sensitive training and awareness:

Awareness and training programmes for GenAI developers must be omnipresent throughout this stage of the value chain, enabling developers to formulate inclusive metrics, improve explainability of the GenAI application's decision-making, and allow for greater representation of the concerns of women and gender minorities in their content and usage policies. GenAI businesses are often focused on services' fast deployment, which in turn deprioritises gender awareness. Principle 16 of the UNGPs addresses this aspect of GenAI development by embedding responsibility to uphold human rights through personnel training and policy statements that can reflect in operations. **Businesses should invest in building capacities and, according to the guidance provided by the Gender Dimensions of the UNGPs, prioritise this training to reduce human rights risks for women and gender minorities.**

A study identified the way in which gender awareness programmes can be structured to bring about an ability and willingness to tackle the issue.[152] This study included tutorials as part of its awareness effort; initially, the focus was on identifying gender bias in LLMs through exercises where developers had to spot gender bias. This was done through two tutorial models, one involving bias in recruitment and in autocomplete.

The second tutorial was to gauge willingness to temper gender bias in AI. This experiment led to the conclusion that most male engineers did not concern themselves with gender in AI bias as it did not affect them directly.[153] However, after each experiment, developers showed increased interest in mitigating bias in models. The takeaway was that a hands-on, interactive training module that made developers aware of the lived experiences of women and gender minorities as a result of gender bias in AI, along with education on how debiasing can work, would significantly increase chances of developers taking their own initiatives in their work. Beyond this, toolkits like the one by Knowing Machines[154] for those working with pre-training and training data, are crucial for developers to ask the right questions.

2. Building benchmarks and evaluation technique (and frameworks) to test inclusivity:

It is essential that models be trained to enhance and optimise for privacy.[155] The concept of data sovereignty is a critical element that often gets overlooked in AI systems. It covers the "use, management and ownership of AI to house, analyse, and disseminate valuable or sensitive data".[156] Data sovereignty should be explored through the lens of whose data is being used.[157]

Inspired by Principle 19 of the UNGPs, **businesses should deploy evaluations, fulfilling their responsibility of conducting impact assessments to mitigate adverse human rights impacts.** Evaluations of GenAI models should be **done through community-in-the-loop methodologies** which invite participation from a diverse array of stakeholders. WinoQueer, a benchmark specifically designed to measure whether LLMs encode biases harmful to the LGBTQIA+ community, is a community-sourced and -surveyed tool for such purposes.[158] This template is a product of sentences,

[152] Zhou, K. Z., Cao, J., Yuan, X., Zhou, K.Z., J. Cao, X. Yuan, D.E. Weissglass, Z. Kilhoffer, M.R. Sanfilippo and X. Tong (2023), " 'I'm Not Confident In Debiasing AI Systems Since I Know Too Little': Teaching AI Creators About Gender Bias Through Hands-on Tutorials", arXiv preprint, arXiv:2309.08121

[153] Ibid

[154] "A Critical Field Guide For Working With Machine Learning Datasets", (n.d.), Knowing Machines, last retrieved April 1, 2024

[155] Koerner, K. and B. LaLonde (2023, February 28), "Federated learning: Supporting data minimization in AI", IAPP, last retrieved May 13, 2024

[156] Zowghi, D. and F. da Rimini, F. (2023), "Diversity and Inclusion in Artificial Intelligence", arXiv preprint, arXiv:2305.12728

[157] Ibid

[158] Felkner, V.K., H.C.H. Chang, E. Jang and J. May (2023), "WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models", arXiv preprint, arXiv:2306.15087

names/pronouns, identity descriptors, and predicates that is entirely human-created and human-audited.[159] Such a benchmark can ensure that models are non-exclusionary regarding women and persons of diverse SOGIESCs.

3. Incorporating robust red-teaming and continuous stress-testing GenAI models:

GenAI developers have ramped up their stress-testing mechanisms through red-teaming efforts and hope to address the risks of toxic content and biased language in GenAI content by combining human expertise with ML to filter out harmful content.[160] Red-teaming efforts usually involve rigorous testing of models through simulations that incorporate real-world scenarios. In doing so, red-teaming measures ensure protection against undesirable or malicious behaviours that could compromise the integrity of a model's output.[161]

Principle 13 of the UNGPs allocates responsibility to businesses to take measures that can prevent or mitigate adverse human rights impacts linked to their products and services. Principles 15 and 17 also allow for businesses to conduct due diligence to assess actual and potential human rights impacts. Red teaming can identify vulnerabilities in GenAI that are not apparent during the development phase. It can improve performance and enhance reliability, while being a cost-effective way to test GenAI models.[162] **Red-teaming efforts that focus on identifying how gender-intentional questions and evaluations can mitigate biases and harms are a way to create inclusive foundation models.**

Many businesses have created red-teaming technology that can increase efficiency gains in completion of tasks.[163] However, it must be recognised that these automated red-teaming software cannot replace manual community-in-the-loop processes where women and gender minorities can be a part of the feedback loop; rather, it must augment and simplify tedious tasks for them.[164]

Beyond this, developers can also stress-test GenAI models to identify how they react to questions that can potentially exacerbate biases. One study advocates for a 'Fairness Auditor' to evaluate any bias mitigation algorithm by conducting a thorough, in-depth

[159] Ibid
[160] Appen (2023, April 4), "Red Teaming: Why It's Critical for Accurate and Reliable Generative AI", last retrieved April 1, 2024
[161] Ibid
[162] Ibid
[163] Kumar, R.S.S. (2024, February 22), "Announcing Microsoft's open automation framework to red team generative AI Systems", Microsoft Security blog, last retrieved April 1, 2024
[164] Ibid

analysis of training models.[165] Such an evaluation plays a key role in building trust in bias mitigation methods before they are deployed.

Additionally, many developers also employ the more extreme method of 'jailbreaking' their own GenAI models. Jailbreaking is an approach to identify how GenAI models respond to potentially harmful prompts. For instance, jailbreak prompts may include questions around how to pick a lock or build a bomb. The responses that these jailbreaks generate can further be used to mitigate and correct gender biases that might creep in during deployment.[166] Developers have built repositories of jailbreaks to keep a check on these models that can be leveraged by other developers and deployers of such systems.[167]

These methods can collectively act as safeguards against any gender biases that creep into models. A repository for gender-specific jailbreaks or red-teaming efforts would facilitate developers in creating a GenAI service that is safe and inclusive for women and gender minorities.

4. Embedding Reinforcement learning from human feedback (RLHF) and Constitutional AI (CAI) approaches:

While GenAI continues to evolve, innovations should be explored in the technical and non-technical elements. Inspired by Principles 15 and 16 of the UNGPs, businesses should embed gender inclusivity in operational policies and procedures. **Businesses can leverage RLHF processes and CAI methods to strengthen feedback and improve learning methodologies to address emergent gender harms from GenAI development.** RLHF is a method used to train more helpful, honest, and harmless AI systems.[168] Reinforcement learning occurs when AI models use human learning to understand what the closest-to-accurate answer is for a particular prompt.

Examples of such mechanisms can be seen in numerous GenAI models. The approach follows a three-step feedback process where the AI starts acting randomly in an environment. Following this, two video clips of its behaviour are provided to a human, and human learning enables the AI to conclude which of the two clips

[165] Bhanot, K., I. Baldini, D. Wei, J. Zeng and K. Bennett (2023, August), "Stress-Testing Bias Mitigation Algorithms to Understand Fairness Vulnerabilities", Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (pp. 764-774)
[166] "Jailbreaking ChatGPT: how AI chatbot safeguards can be bypassed" (2023, April 10), *The Economic Times*, last retrieved April 1, 2024,
[167] Ibid
[168] Bai, Y., S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, ... and J. Kaplan (2022), "Constitutional AI: Harmlessness from AI Feedback", arXiv preprint, arXiv:2212.08073

is closest to fulfilling its goal.[169] It then uses reinforcement learning (RL) to find the reward function to justify the human's judgements, while continuing to ask for human feedback to refine its understanding of what it is supposed to generate.[170] This process is only as good as the human's grasp of the task at hand. If the human does not understand the lived experiences of women or gender minorities, the GenAI model could likely learn to justify and imbibe gender biases and harmful judgements. RLHF is a process that allows for diverse feedback providers, making the data more holistically representative and fine-tuned.

CAI has emerged as an upgraded method that provides for training without any human feedback labels for harms.[171] This method partially replaces RLHF with a "constitution".[172] This constitution is a set of principles or instructions that AI models can use to supervise other AI models, resulting in scaling of supervision, elimination of evasive responses, and creation of transparency in what constitutes responsible AI behaviour.[173] This principle also aligns with the concept of explainable AI (XAI), where users can understand the internal workings and reasoning behind GenAI predictions.[174] XAI can manifest through model interpretability, regulatory requirements, trust and adoption, model validation, as well as debugging and improvement in decision-making.[175]

The premise of developing CAI and XAI stems from the concept of AI being both helpful and harmless. While an AI assistant answering with "I don't know" would be harmless, it would also be useless.[176] Therefore, a part of RLHF would be to constitute a set of instructions that can be encoded into the training goals of LLM-building, such as CAI.[177] **Building CAI to address gender concerns and equality can help improve language model capabilities significantly.**[178]

5. Incorporating profanity and privacy filters in models:

Life cycles of GenAI models can vary depending on purpose and format of output. However, various processes can be incorporated into these services to reduce the emergence of harms. Inspired by Principles 11 and 18 of the UNGPs, under which the responsibility for mitigating gender bias and gender-based harassment devolves

[169] Amodei, D., P. Christiano and A. Ray (2017, June 13), "Learning from human preferences", OpenAI, last retrieved April 1, 2024
[170] Ibid
[171] Bai, Y., S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, ... and J. Kaplan (2022), "Constitutional AI: Harmlessness from AI Feedback", arXiv preprint, arXiv:2212.08073
[172] Ibid
[173] Ibid
[174] Hassija, V., V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud and A. Hussain (2023), "Interpreting Black-Box Models: A Review on Explainable Artificial intelligence", *Cognitive Computation*, 16, 45–74
[175] Ibid
[176] Bai, Y., S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, ... and J. Kaplan (2022), "Constitutional AI: Harmlessness from AI Feedback", arXiv preprint, arXiv:2212.08073
[177] Ibid
[178] Ibid

upon businesses, **GenAI models should incorporate profanity and privacy filters within their models.**

Most GenAI models already have filters in place to screen content that uses inappropriate or harsh language. The makers of Call of Duty have announced that the game's next title will make use of an AI-powered voice chat moderation tool called ToxMod.[179] It will identify toxic speech in real time, including any discriminatory language or harassment. While this is a step forward in addressing gender-based harassment that often ensues in online fora, the question does emerge as to how in touch with community-in-the-loop mechanisms this intervention might be.

Additionally, employing data minimisation techniques to restrict access to data using security mechanisms such as encryption and access-control lists, as well as ensuring that policies are put in place to ensure the removal of personal and proprietary data, is essential for ensuring data security.[180] Minimisation also avoids storing data on cloud servers for longer than is necessary. An Indian AI platform, Agami's OpenNyAI, built Nyaayabandhu, an SLM to bridge the gap between justice and technology. It follows a collaborative process of data collection that understands the need to protect sensitive information from being used by the model.[181] The model has filters in place that check for profanity as well as sensitive information that users may have provided the chatbot. The sensitive information is screened and redacted as well, disabling the chatbot for generating content beyond its stipulated usage.

6. Adhering to contextually created design principles for gender-sensitive models:

Design principles to build GenAI representative of women and gender minorities should follow a set of non-negotiable guidelines. Principles 16, 18, and 20 of the UNGPs recommend that businesses lay down the process of embedding policies for mitigation as well as involving diverse stakeholders for consultation processes.

[179] Wodecki, B. (2023, September 1), "Call of Duty Turns to AI to Stop Players Using Profanities In-Game", AI Business, last retrieved April 1, 2024

[180] Koerner, K. and B. LaLonde (2023, February 28), "Federated learning: Supporting data minimization in AI", IAPP, last retrieved May 13, 2024

[181] Key informant interview with AI system developer and deployer.

The Inter-Agency Working Group on Artificial Intelligence formulated the 'Principles for the Ethical Use of Artificial Intelligence in the United Nations System',[182] which provide several rules of thumb for design of GenAI models. They include defining purpose and necessity of usage, safety and security, fairness and non-discrimination, sustainability, privacy, human oversight, and explainability, to name a few.[183] The principles also employ the pillars of inclusion and participation to promote gender equity, wherein consultations with women and gender minorities as primary stakeholders and aggrieved parties should be held to identify risks, harms, and adverse human rights impacts.

7. Prioritising representation of gender in leadership roles at big tech companies:

The responsibility of businesses to uphold human rights through inclusive hiring policies supporting women and gender minorities traces its foundations to Principles 11 and 12 of the UNGPs. While biases and representation-related harms can be detected in various stages of the value chain, **the inclusion of diverse perspectives through inclusive hiring practices could reduce gender-based harms.**

Recruitment of women in GenAI businesses is more an issue of concentration of power than it is about representation.[184] While having diverse voices in the workforce can introduce nuance, the question remains as to whether this step would stand the test of an industry that is male-dominated and supported by similar tangential industries.[185]

[182] "Principles for the Ethical Use of Artificial Intelligence in the United Nations System" (2022), UNSCEB, last retrieved April 1, 2024
[183] Ibid
[184] Key informant interview with expert with specialty in understanding from a multilateral organisation
[185] O'Mara, M. (2022, August 11), "Why can't tech fix its gender problem?", MIT Technology Review, last retrieved April 1, 2024

After dataset preparation and modelling comes model deployment, where trained and tested models get integrated into various systems and services that see usage from a wider pool of actors. An example of a model reaching deployment would be ChatGPT, the chatbot service that uses the GPT-3.5 and GPT-4 models. Model deployment is the most people-facing stage of this study's value chain, as it represents the point where access to GenAI technology and its output-generating capabilities opens to a large number of people and groups.

This chapter covers the third and final section of our stages value chain and risk mapping for model deployment.

Unpacking the deployment stage:

The model deployment stage represents a trained model becoming integrated into GenAI systems and services, with chatbots being an example of deployment. Similar to the modelling stage, the deployment stage also involves content and usage policies (CUPs) and monitoring that shape prohibitions and measures that penalise or barricade behaviours and capabilities deemed to be harmful and dangerous.

As services and systems, GenAI technology faces end-users, a blanket term for the various kinds of stakeholders who make use of GenAI to varying degrees of success. Prompts are the requests users make to GenAI services and the underlying model capabilities which, mediated by CUPs and related measures, then yield outputs.

Output can then see various kinds of application and circulation, which can have consequences of their own. For example, deepfakes can see extensive circulation across the internet and social media with serious consequences, from the humiliation of people to the spread of misinformation. The potentially adverse consequences of GenAI-produced content make it critical to think about ways to detect, impede, and eliminate harmful material originating out of GenAI.

Prompt-based outputs can be a site for creating feedback loops in GenAI development and deployment. Integrating measures for procuring users' feedback, ratings or indications for what could make for better output is an important part of model deployment, as are openings for flagging issues.

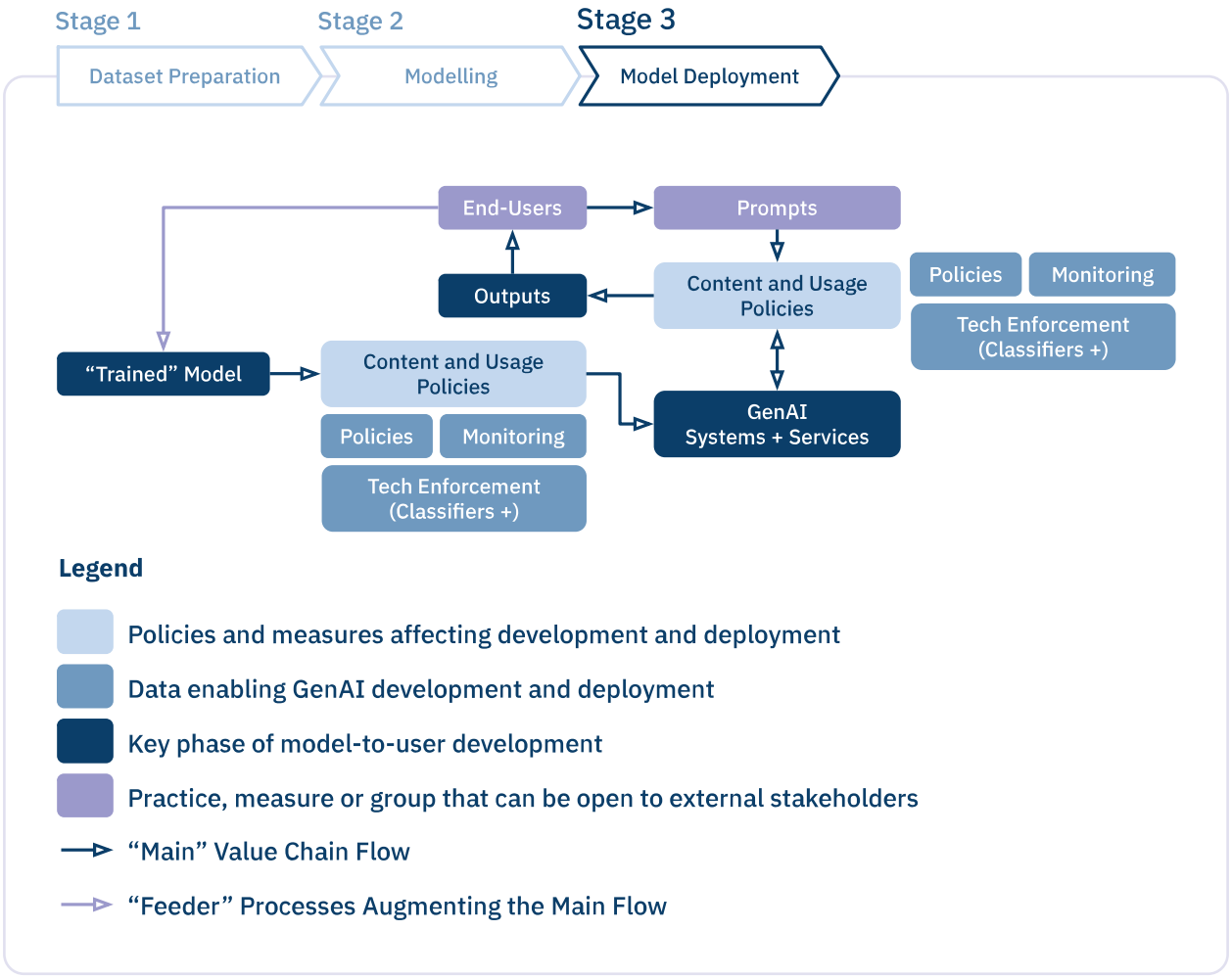


Figure 5. The GenAI Value Chain: The sub-components of the model deployment stage

The figure above outlines the model deployment stage. The risks identified in this stage include the perpetuation of gender stereotypes (Risk 1), ensuing misinformation, misgendering, and under-representation of women and gender minorities (Risk 2), online harassment and gender-based violence (Risk 3), the lack of specialised policy guardrails to regulate GenAI systems and services (Risk 4), and bias in GenAI-based hiring systems (Risk 5).

Risk 1: Perpetuation of gender stereotypes

The ***manifestation of gender stereotypes*** is through a generalised view or preconception of attributes or characteristics that are viewed as possessed by men, women, and other genders. ***These stereotypes become harmful when they make restrictive assumptions about any gender and are incorporated into GenAI systems and services.*** GenAI deployment sees the emergence of gender stereotypes, and harmful and homophobic content in an overwhelming number of case studies. According to Principle 11 of the UNGPs, the responsibility for mitigation of such harms lies with the businesses themselves.

The perpetuation of gender stereotypes finds its way into GenAI services and systems due to a lack of gender-friendliness and inclusiveness in CUPs as well as in prompt engineering efforts. Models that have been trained on adversarial data are far more likely to categorise "raciness" through a gendered lens, often detrimental to women and gender minorities. A *Guardian* investigation used AI tools developed by big tech companies including Google and Microsoft to detect "raciness" and its determinants.[186] Upon analysing hundreds of photos of men and women in underwear, working out, partaking in medical tests that involve partial nudity, the investigation found that the models tagged photos of women in these everyday situations as being sexually suggestive.[187] AI algorithms tested on images released by the US National Cancer Institute demonstrating the technique of a clinical breast examination gave these images the highest score for "raciness" when put through Google's AI.[188] Microsoft's AI was 82% confident that these images were "sexually explicit in nature", and Amazon tagged them as "explicit nudity".[189]

Considering that experts believe GenAI will artificially generate as much as 90 percent of content on the internet in the future, ensuring non-proliferation of bias becomes a bigger issue.[190] Stable Diffusion, another GenAI application, also categorised images of different occupations according to their gender-based stereotypes: where men were more likely to be perceived as engineers, CEOs, doctors, and politicians, and women as lower-paid teachers, social workers, and housekeepers, among others.[191]

[186] Mauro, G. and H. Schellmann (2024, January 9), " 'There is no standard': investigation finds AI algorithms objectify women's bodies", *The Guardian*, last retrieved April 1, 2024
[187] Ibid
[188] Ibid
[189] Ibid
[190] Nicoletti, L. and D. Bass (2023), "Humans Are Biased. Generative AI Is Even Worse", Bloomberg Technology + Equity, Bloomberg, last retrieved April 1, 2024
[191] Ibid

Stereotyping is harmful to marginalised and vulnerable genders because they undermine the basic tenets of the right to a life with dignity, security, and absence of discrimination — guaranteed by Articles 14, 15, and 21 of the Indian Constitution. Manifestation of this risk in the public sphere leads to manifold consequences that can affect public opinion on the lived experiences of women and gender minorities.

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to life, liberty, and security	*This right allows for gender to be accurately represented in GenAI applications that involve public usage.*	**International law:** UDHR: Arts. 3, 5 | ICCPR: Arts. 6, 7, 16, 26 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21
Right to equality before law and protection against discrimination	*This right allows for women and persons of diverse SOGIESCs to a life of dignity.*	**International law:** UDHR: Arts. 2, 7 | ICCPR: Arts. 3, 26 | ICESCR: Arts. 2, 3 | CEDAW: Arts. 1, 2, 7(c), 15 **Indian law:** The Constitution of India: Arts. 14, 15

Table 12. Human rights mapping for Risk 1 in model deployment

Risk 2: Misinformation, misgendering, and underrepresentation

Misinformation can thrive, if not actively contained by businesses, in the GenAI ecosystem. The spread of false or inaccurate information, which is fuelled by misgendering and under-representation, can be used to wilfully abuse the human rights of women and gender minorities. This phenomenon particularly affects women due to the overwhelming rise in pornographic deepfakes that should be considered manipulated content.[192] Principle 11 of the UNGPs and its Gender Dimension repose the responsibility of undoing stereotypes and hyper sexualisation in businesses. The major hurdle posed is that such content is so convincing that it becomes extremely challenging to differentiate it from actual instantiations. Without guardrails to regulate this industry, misinformation can have exponential consequences.

[192] Goodchild, J. (2023, September 27), "Gender bias in AI: 'Where are all the women?' ", SC Media, last retrieved April 1, 2024

User interfaces worldwide have made attempts to reshape interface design to accord people more agency in defining their own gender identity beyond the binary.[193] In contrast, the practice of automated gender recognition (AGR) in GenAI services can potentially remove the opportunity to self-identify, leaving room for the application to infer this sensitive information from existing datasets.[194] When a community or population group is not represented, or disproportionately represented, they lose the ability to effectively advocate their right to fundamental rights and freedoms guaranteed to all, including essential services such as housing and healthcare.[195]

A study on trans representation in AGR literature found that 94.8 percent of papers treated gender as binary.[196] Over 72 percent of these papers also believed gender to be immutable (the idea that gender cannot change post-assignment). AGR is likely to misclassify trans and non-binary persons due to the presumption that gender is binary. This understanding of gender does not allow space for women of colour and non-binary identities to be accurately classified. For instance, in bathrooms labelled for usage by either men or women, non-binary and agender people are often forced to misclassify themselves, causing substantial discomfort in the form of dysphoria. Even when opting for a gender, there is always a risk of hostility, rejection, and assault from other cis users of the facilities. [197]

Questions of accurate representation plague every sub-component of the value chain, but in this stage the absence of gender-friendly design in the GenAI service, or any gender advisories reflected in the usage policies, can manifest in applications that do not cater to the lived realities of women, and gender minorities. Misinformation can also be tackled through prompt engineering and feedback loops that allow businesses to flag and address harms that may emerge.

Aspects of misgendering, misinformation, and under-representation can violate a slew of human rights under the Indian Constitution. This could range from how women and gender minorities navigate the public sphere with dignity (Article 21) to equality before the law (Article 14) and right to freedom of expression (Article 19(1)(a)) and to be identified accurately and

[193] Leufer, D. (2023, January 13), "Computers are binary, people are not: how AI systems undermine LGBTQ identity", Accessnow
[194] Ibid
[195] Ibid
[196] Keyes, O. (2018), "The Misgendering Machines: Trans/HCI implications of Automatic Gender Recognition", Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 88
[197] Ibid

respectfully. Freedom of expression comes with a reasonable justification that encompasses gender and sexuality sensitivities in its decision-making. This is laid down in Part II of the Schedule in the IT (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, and can extend to GenAI-based services that involve creative output. The strength or impact of their inclusion must also be taken into account in this regard. Women and gender minorities must also have freedom from fear of physical or psychological harm that emerges from biases creeping into GenAI systems. The Transgender Persons (Protection of Rights) Act specifically calls out any such harm or injury to the life, safety, health or well-being of a transgender person, including acts of physical, sexual, verbal, emotional, and economic abuse, which are punishable with imprisonment.

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to life, liberty, and security	*This right entitles women and gender minorities to have more agency over how they wish to be identified in the public sphere.*	**International law:** UDHR: Arts. 3, 5 | ICCPR: Arts. 6, 7, 16, 26 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21
Right to equality before law and protection against discrimination	*This right entitles women and gender minorities to protect themselves against gender-based violence and harm that emerges from misinformation and misleading gender data.*	**International law:** UDHR: Arts. 2, 7 | ICCPR: Arts. 3, 26 | ICESCR: Arts. 2, 3 | CEDAW: Arts. 1, 2, 7(c), 15 **Indian law:** The Constitution of India: Arts. 14, 15
Right to freedom of expression	*This right allows for women and persons of diverse SOGIESCs to express their identity in an accurate manner.*	**International law:** UDHR: Arts. 18, 19 | ICCPR: Arts. 18, 19 | CEDAW: Arts. 1, 2 **Indian law:** The Constitution of India: Art. 19 | The IT (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021: Schedule, Part II (a), (d)

Table 13. Human rights mapping for Risk 2 in model deployment

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Duty to the community	*States are responsible for ensuring misinformation of misrepresentation caused by GenAI services is regulated and aggrieved persons are redressed.*	**International law:** ICCPR: Art. 1 | ICESCR: Art. 1 | CEDAW: Art. 3 **Indian law:** The Constitution of India: Art. 51A(e) | The IT (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021: Sec. 3(1)(b)(ii)
Right to freedom from physical and psychological harm	*Women and gender minorities are entitled to a life of dignity and security when they use GenAI services.*	**International law:** ICESCR: Art. 12 **Indian law:** The Constitution of India: Art. 21 | The Transgender Persons (Protection of Rights) Act, 2019: S. 18(d)

Table 13. Human rights mapping for Risk 2 in model deployment

Risk 3: Online harassment and gender-based violence

Misinformation often leads to online harassment and violence — especially for women and other gender minorities. Pornographic deepfakes, or AI-generated synthetic media, have primarily been used to create pornographic representations of women.[198] Sensity AI estimates that 90 to 95 percent of online deepfake videos are non-consensual, and around 90 percent of them feature women.[199] While Principle 23 of the UNGPs states that businesses should not contribute to gender violence based on socio-normative contexts, the responsibility of reducing these harms in the digital space should also lie with business-sided stakeholders.

The results of these deepfakes are far from perfect. However, the trajectory of deepfakes in the GenAI landscape indicates they can soon become indistinguishable from genuine content. Arguably, the quality of these deepfakes is inconsequential when considering the significant psychological toll that such technology exerts on the victim.[200]

Pornographic content-generating platforms provide subscriptions through anonymous 'Discord' or 'Patreon' accounts, giving users access to create deepfakes that can be tailored to ethnicities, body types, and other features. There is no requirement for disclosure

[198] Hao, K. (2022, February 4), "A horrifying new AI app swaps women into porn videos with a click", *MIT Technology Review*, last retrieved April 1, 2024
[199] Ibid
[202] Ibid

on how the material has been obtained, apart from a perfunctory disclaimer intended to absolve the platform in case of any conflict regarding the content.[201]

The aftermath of such experiences can be harrowing for victims. People viewing explicit images of a person without their consent, whether real or fake, is a form of sexual violence.[202] Victims often face judgement and confusion from the community around them. A schoolteacher in the US lost her job after pupils' parents learned about a deepfake with her likeness; one that was made without her consent.[203] Due to the lack of understanding of what this technology entails, victims can often be ostracised and isolated more than usual. An even trickier segment is child pornographic deepfakes. In these cases, no physical harm is caused during the creation of the content but its availability may lead to real-life victimisation and psychological trauma.[204]

The harassment and violence factors can be tackled at the output stage of the value chain through constant updation and through markers to identify deepfakes as well as bridge the gaps in gender safety.

Owing to the design failures in GenAI systems, the harassment and violence often directed towards women and gender minorities in varying degrees constitute a sustained attack on their right to dignity, to a life without fear of violence, under Article 21 of the Indian Constitution. Misgendering and misinformation, especially when categorising and representing trans and non-binary identities, can have long-term effects on the right to privacy, which encompasses prevention of use of their personal data for non-consensual output like deepfakes and gender recognition.

Under Section 18(d), the Transgender Act does specify acts of such harassment and harm as offences punishable with imprisonment. Further, under Section 4 of the DPDP Act, the entity that processes the personal data of a user must do so for a lawful purpose which involves informed consent and legitimate use. This safeguards women and gender minorities in cases where their personal or non-personal information is being used with the intent of harassing them or committing acts of violence against them.

[201] Kamdar, H. (2023, May 18), "With Deepfake Porn and God Chatbots, AI Dystopia is Here", *The Swaddle*, last retrieved April 1, 2024
[202] Hunter, T. (2023, February 14), "AI porn is easy to make now. For women, that's a nightmare.", *The Washington Post*, last retrieved April 1, 2024
[203] Ibid
[204] Ibid

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to life, liberty, and security	*This right entitles women and gender minorities to a life of dignity and safety in a situation where GenAI services have had a negative or violent effect on their person.*	**International law:** UDHR: Arts. 3, 5 | ICCPR: Arts. 6, 7, 16, 26 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21 | The Transgender Persons (Protection of Rights) Act, 2019: S. 18(d)
Right to privacy	*Women and gender minorities are entitled to privacy as an inalienable right that must be guaranteed by GenAI service providers in the safe deployment of their services to all persons.*	**International law:** UDHR: Art. 12 | ICCPR: Art. 17 | ICESCR: Art. 3 **Indian law:** The Constitution of India: Art. 21 | The IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011 | The DPDP Act, 2023: Sec. 4
Right to freedom from physical and psychological harm	*Women and gender minorities are entitled to a life of dignity and security when they use GenAI services, without any danger to their person due to their SOGIESCs.*	**International law:** ICESCR: Art. 12 **Indian law:** The Constitution of India: Art. 21 | The Transgender Persons (Protection of Rights) Act, 2019: S. 18(d) | The IT (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021: Sec 3(1)(b)(ii), (iii)

Table 14. Human rights mapping for Risk 3 in model deployment

Risk 4: Lack of specialised policy or guardrails for responsible usage

Regulation has its own set of stakeholder perspectives. Among GenAI developers, the risks of AI are often overlooked as not requiring rigorous oversight. But ***GenAI can spread misinformation, which has seen manifestations in many forms in mainstream media.*** Those arguing for AI regulation point out that the impact of GenAI could be profoundly damaging if left unregulated.[205] Principles 13 and 15 of the UNGPs, drawing from the Gender Dimensions, accord the responsibility for mitigating gendered human rights violations and the responsibility of due diligence, respectively, to businesses. These arguments shed light on what responsible innovation constitutes.

[205] Pimentel, B. (2024, February 1), "Why AI still needs regulation despite impact", Thomson Reuters law blog, last retrieved April 1, 2024

The intent behind AI regulation is simple: its impact must not cause the kind of disruption that results in the destruction of businesses, privacy, or government programmes.[206] Generally, a policy regulatory framework would answer multiple questions regarding the potential lack of transparency in the functioning of GenAI systems, the training datasets, bias and fairness considerations, intellectual property infringement, and privacy as well as security concerns.[207]

Policymakers have begun rigorously appraising ***GenAI and AI policy guardrails and regulations*** to understand, control, and guarantee the safety of this technology while encouraging innovation.[208] The need to regulate this space stems from an economic incentive for businesses to escalate GenAI adoption. However, in the absence of such regulatory protections, businesses should be expected to sufficiently refine these models so that they do not hallucinate or perpetuate biases.

The role of regulation in the GenAI value chain is multifaceted, and comes into play at each step of the development cycle to ensure that innovation and modelling follow the responsible AI principles of fairness, accountability, and safety.

The absence of policy-based guardrails is an important factor in grievance redressal in cases of gender bias. The right to effective remedy stems from the constitutional rights intrinsic to human life and dignity, and is a fundamental right under Articles 32 and 226 of the Indian Constitution.

[206] Ibid
[207] Kremer, A., A. Luget, D. Mikkelsen, H. Soller, M. Strandell-Jansson and S. Zingg (2023, December 21), "As gen AI advances, regulators – and risk functions – rush to keep pace", McKinsey & Company, last retrieved April 1, 2024,
[208] Ibid

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to effective remedy	*Any person who is negatively affected upon using a GenAI service has the right to seek remedies in various forms as prescribed by the law and non-judicial grievance redressal mechanisms.*	**International law:** UDHR: Art. 8 | ICCPR: Art. 2(3) | CEDAW: Art. 23 **Indian law:** The Constitution of India: Arts. 32, 226

Table 15. Human rights mapping for Risk 4 in model deployment

Risk 5: Gender bias in GenAI-powered hiring systems

When GenAI technology entered the realm of hiring and talent management, the product-market fit was apparent — GenAI could sift through heaps of job applications, picking out suitable candidates through predictive analyses and ML mechanisms.[209]

This practice is, in reality, a double-edged sword. The adoption of automated systems should result in objective decision-making; however, the following case study revealed that systems are built on biased datasets and processes. Looking at the Amazon recruitment AI case study in the previous stage brings the study to a key moment in understanding AI ethics. Integrating GenAI into decision-making appears an antidote to bias in hiring, but unconscious bias tends to exist because of the effects of historical data and humans-in-the-loop. ***Despite the potential for GenAI to reduce discrimination through filters and improved training, gender biases could continue to emerge.*** Drawing from the Gender Dimensions guidance for Principle 11 of the UNGPs, organisations must remain vigilant about ethical and compliance implications of using this technology in areas where economic opportunities are inequitable, such as Human Resource practices.[210]

In similar situations, it is important to maintain transparency about what parts of the recruitment process that are being handled by a GenAI service, and the AI tools being employed for this purpose. GenAI decision-making is often opaque, and can potentially become a legal quagmire, with employment-related decisions not being backed by rationale from a GenAI application.[211]

Gender bias and exclusion in hiring systems and leadership roles are a result of value chain failings at each stage. Historical data that is fed into models influences the deployment effects of how women and gender minorities are seen in hiring systems. The absence of well-tested prompts to improve gender-friendliness and inclusion is also an aggravating factor in the way GenAI services screen applicants.

Differential treatment enabled by biased GenAI systems obfuscates transparency in how GenAI sorts through applications for recruitment, based on gender. The principle of explainability in

[209] Malik, A. (2024, February 20), "AI Bias In Recruitment: Ethical Implications And Transparency", *Forbes*, last retrieved April 1, 2024
[210] Ibid
[211] Ibid

GenAI decision-making is essential to protect women and gender minorities from discrimination, as laid down under Articles 14 and 15 of the Indian Constitution. Under India's labour protection laws, discrimination between men and women workers while recruiting for the same work or work of similar nature is prohibited in accordance with Section 5 of the Equal Remuneration Act, 1976. The Code on Wages, 2019 provides the same stipulation under Section 3(2)(ii).

HUMAN RIGHTS RISK	APPLICATION	RELEVANT STATUTORY PROTECTIONS
Right to equality before the law and protection from discrimination	*Women and gender minorities have the right to fair and just recruitment processes; differential treatment should be acceptable only when there is a reasonable justification for the same.*	**International law:** UDHR: Arts. 2, 7 | ICCPR: Arts. 3, 26 | ICESCR: Arts. 2, 3, 7 | CEDAW: Arts. 1, 2, 7(c), 15 **Indian law:** The Constitution of India: Arts. 14, 15 | The Equal Remuneration Act, 1976: Sec. 5 | The Code on Wages, 2019: Sec. 3(2)(ii)
Right to information	*This right allows for women and gender minorities to have transparency in how recruitment processes are being conducted, and whether it is fair and equitable.*	**International law:** UDHR: Art. 19 | ICCPR: Art. 19 **Indian law:** The Constitution of India: Art. 19(1)(a) | The Right to Information Act, 2005: Sec. 3

Table 16. Human rights mapping for Risk 5 in model deployment

Pathways to mitigating risks

Strategies to tackle GenAI at the deployment level are usually public facing and often involve the services of third parties as well as civic awareness to address issues at various stages. These services range from prompt engineering models to caveats and attributions, technological impact assessments, creation of efficient feedback loops, and altering perceptions of the anthropomorphisation of GenAI systems. A collaborative approach would benefit the ecosystem variously and could help address the human rights risks emerging from these systems in a more meaningful manner.

PATHWAY	RELEVANT UNGP REFERENCED
Prompt engineering and prompt tuning responses	13
Caveats and attributions in generated output	15
Technological impact assessments	19
Feedback loops for addressing user grievances	20
Public perception of GenAI technology	13

Table 17. Consolidated recommendations for the GenAI model deployment stage

1. Strengthening prompt engineering methodologies and prompt tuning responses:

Principle 13 of the UNGPs stipulates that businesses have a responsibility to prevent or mitigate adverse human rights impacts that are directly linked to their operations, even if not caused by them. **Businesses should adopt innovative measures and strategies, or strengthen existing ones, to ensure that models are not biased or become sites for gender-based human rights violations.** Prompt engineering is a method of fine-tuning training models that users can customise to get the best result out of the language models.[212] Prompt engineering also gives developers a lot more control over user interactions with the GenAI service. Effective usage of prompts can act as a crucial corrective tool to mitigate bias in the training data. This process, when conducted at the deployment stage, feeds back into the GenAI value chain to create better datasets that are equipped with these questions.

As usage of GenAI services increased, questions on the potential dangers of GenAI capabilities emerged.[213] Several trains of thought emerged — for instance, there were those who used such services to recreate malware strains, and at the other end of the spectrum were security professionals whose usage was aimed at identifying potentially exploitable code that could heighten security.[214] This practice of reverse engineering models to ramp up security and create applications for other outputs can also be handy in combating gender biases as and when they crop up.

[212] "What is Prompt Engineering?" (n.d.), Amazon Web Services, last retrieved April 1, 2024
[213] Carney, R. (2023, June 11), Generative AI – A Creator of Malware or Defender of Cybersecurity?", *The Times of India*, last retrieved April 1, 2024
[214] Ibid

2. Communicating caveats and attributions in generated output:

Guided by the Gender Dimensions, Principle 21 of the UNGPs places the responsibility of communicating strategies adopted to address human rights risks with businesses. GenAI outputs align with training data and how it has been collected and processed. Concepts like attribution are important at this stage and are legally upheld.[215] Attribution is essential to ensure that if an LLM is generating plagiarised content, complete credit for the research or content is appropriately given. Attribution is also a combative tool against the circulation of illicit content, misinformation, and pornographic deepfakes generated to target women particularly, and gender minorities. Finding a balance between trust in attribution and human oversight will be an ongoing challenge.

Providing caveats that generated output may not be completely accurate, or provide the entire picture of the problem is essential; however, this comprises one approach to increasing explainability in GenAI models.[216] It is in line with building transparency and credibility in ongoing efforts to make GenAI inclusive, as Principle 15 of the UNGPs states that **businesses should be transparent in their communications when dealing with concerns around human rights, especially if impacting specific gender identities, appropriately attribute, and provide disclaimers and caveats wherever relevant to build user awareness and understanding of GenAI models.**

3. Frequently conducting technological impact assessments:

Principle 19 of the UNGPs urges businesses to take the responsibility to conduct technological impact assessments and integrate the findings for appropriate action to curb or mitigate any adverse human rights impacts. It lies within the purview of a business's operations to **conduct technical impact assessments, communicate its findings, work towards addressing areas of concern, and share these practices publicly.**

An impact assessment is a risk management tool that seeks to answer a simple question: along with the system's intended use, for whom could it fail?[217] This is an extremely relevant question

[215] Deloitte AI Institute (n.d.), "Proactive risk management in Generative AI", Deloitte, last retrieved April 1, 2024
[216] Ibid
[217] Long, S., J. Pesner and T. Romanoff (2022, November 9), "Explainer: Impact Assessments for Artificial Intelligence", Bipartisan Policy Center, last retrieved April 1, 2024

that most systems seek to answer when their models negatively impact women and gender minorities. Impact assessments for GenAI services allow for a benefits and risk analysis that is broader than a risk assessment, where an AI service can be evaluated for its data protection, environmental impact, and human rights impact.[218] This also allows an organisation to compare and optimise the costs and benefits of a system through a specific impact lens.

4. Incorporating feedback loops for addressing user grievances:

Data feedback loops for LLMs are largely determined by "in-house" pre-launch training and testing.[219] However, post-deployment, feedback loops can broadly include user feedback by liking or disliking a generated answer, or by users asking follow-up questions that signal whether the answer was satisfactory or not. The ways to look at user feedback could include a variety of natural data feedback loops. The Gender Dimensions for Principles 16 and 18 posit that **businesses should adopt consultative processes that strengthen gender-inclusive understanding and feedback mechanisms in business operations.** Through this method, LLMs could add new features where users can save and organise responses they find helpful and delete the rest.[220] They could also create challenges where users compete with the GenAI service to evaluate answers. All this is to create a system of signals that gauges the quality of the LLM's answers, which can in turn be used to improve algorithms.[221]

User feedback can also involve humans-in-the-loop. After a question is submitted, the response is reviewed and amended by the service's in-house team that constitutes the human in the loop. Thus, an AI service benefits by learning from both the in-house human intervention and through implicit user feedback.[222] Including third-party feedback and evaluation loops in GenAI applications can contribute to varied perspectives in tracking the effectiveness of any human rights impacts. This tracking mechanism through internal and external feedback is laid down and suggested under Principle 20 of the UNGPs.

[218] Ibid
[219] Hagiu, A. and J. Wright (2023, July 11), "To Get Better Customer Data, Build Feedback Loops into Your Products", *Harvard Business Review*, last retrieved April 1, 2024
[220] Ibid
[221] Ibid
[222] Ibid

5. Strengthening public perception of GenAI technology:

Principle 13 of the UNGPs requires businesses to prevent or mitigate adverse human rights impacts directly linked to their operations. Further, Principle 18 urges businesses to strengthen transparent communication with its users to build trust. **Identifying human rights risks, conducting the required due diligence through audits, a consultative process and impact assessments, and creating communication channels that transparently convey the impact of such mitigation measures should be encapsulated as part of a business's operations.**

Trust can also be strengthened by building a framework for tackling anthropomorphisation of GenAI; narratives have been branching into discussions around artificial general intelligence (AGI), which is termed as "superhuman AI" that can overtake humans in the loop.[223] Despite such arguments, models still show basic errors in understanding certain concepts. This presents us with a paradox: how can seemingly superhuman capabilities persist with errors that few humans would make?[224] The phenomenon of anthropomorphising GenAI, where it is deemed to have human capabilities, is becoming more pervasive, and can be misleading for users regarding its ethical capacities.[225] The challenges posed through such anthropomorphisation must be tackled through designing for inclusion of women and gender minorities, educating users on the consequences of such a title, and, most important, maintaining human control.[226] Providing meaningful support to human creators rather than a body that has its own intent and authorship should be the priority throughout.[227]

[223] Meacham, S. (2023, September 8), "A Race to Extinction: How Great Power Competition Is Making Artificial Intelligence Existentially Dangerous", *Harvard International Review*, last retrieved April 1, 2024

[224] West, P., X. Lu, N. Dziri, F. Brahman, L. Li, J.D. Hwang, ... and Y. Choi, Y. (2023, October), "The Generative AI Paradox: 'What It Can Create, It May Not Understand', The Twelfth International Conference on Learning Representations

[225] Gupta, A. (2024, January 31), "Navigating the AI Frontier: Tackling anthropomorphisation in generative AI systems", Observer Research Foundation, last retrieved April 1, 2024

[226] Ibid

[227] Ibid

In May 2024, OpenAI released GPT-4o, with the 'o' standing for "omni". According to the firm, the model was capable of accepting input and preparing outputs in "any combination" of pictures, sounds, and text as input, and had been trained on all three mediums on the same neural network.[228] February 2024 saw the announcement of Sora,[229] a video-generating GenAI model, and the release of Google's "light-weight" Gemma open models.[230] Research released in March 2024 suggested that Apple was also working on GenAI and "multimodal large language models (MLLMs)".[231] Such accounts show that the interest in, investments into, and the exploration of GenAI has remained robust.

As the experimentation and adoption continue, so must good governance innovation, circumspect scrutiny, and the pursuit of bettering GenAI systems to make it something that does not put people in harm's way. This study had set out to investigate the kinds of gender risks prevalent in the field of generative artificial intelligence, and attempt an understanding of the social risks in GenAI development and deployment— with priority to gender. To this end, the study took GenAI value chains, human rights, the United Nations Guiding Principles on Business and Human Rights (UNGPs), and the Gender Dimensions, as guiding principles and frameworks to protect human rights, especially of women and gender minorities.

GenAI value chains were understood as comprising three stages: dataset preparation, modelling, and model deployment. First, the procurement and curation of data from sources like the internet and from dedicated labour, followed by the development of GenAI models and their capabilities through models' training on the prepared data corpus. Finally, trained models see integration into systems or development into services that are made available to large groups of people.

As the study sought to find pathways towards the gender "un-risking" of GenAI, AI value chains were conceptualised in a more detailed manner. This approach moved from the broad stages to a more detailed presentation of the "sub-components" that sought

[228] OpenAI (2024, May 13). "Hello GPT-4o"
[229] OpenAI (n.d.), "Creating video from text"
[230] Schmid, P., O. Sanseviero and P. Cuenca (2024, February 21), "Welcome Gemma – Google's new open LLM", The AI community building the future, Hugging Face blog
[231] MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training", Apple Machine Learning Research

to represent sites of gender risk, development-related processes, and assets like AI red teaming, content and usage policies, and training data.

Our research identified gender risks within each stage of the articulated GenAI value chain. The dataset preparation stage risks include the erasure of women and gender minorities in GenAI models' underlying data needs, and the lack of training for collecting and storing data. The modelling stage hosts risks such as absence of gender diversity in the workforce behind GenAI models, as well as the paucity of gender-sensitive training for the GenAI landscape's existing pool of developers. Problems persist into the model deployment stage, as seen in the case of risks like GenAI-enabled harassment and gender-based violence, and the spread of misinformation, misgendering, and misrepresentation. Each risk in each stage is not only a prelude to some kind of gender harm, but also an endangerment and potential abuse of human rights, which the study has mapped alongside the risks it identified.

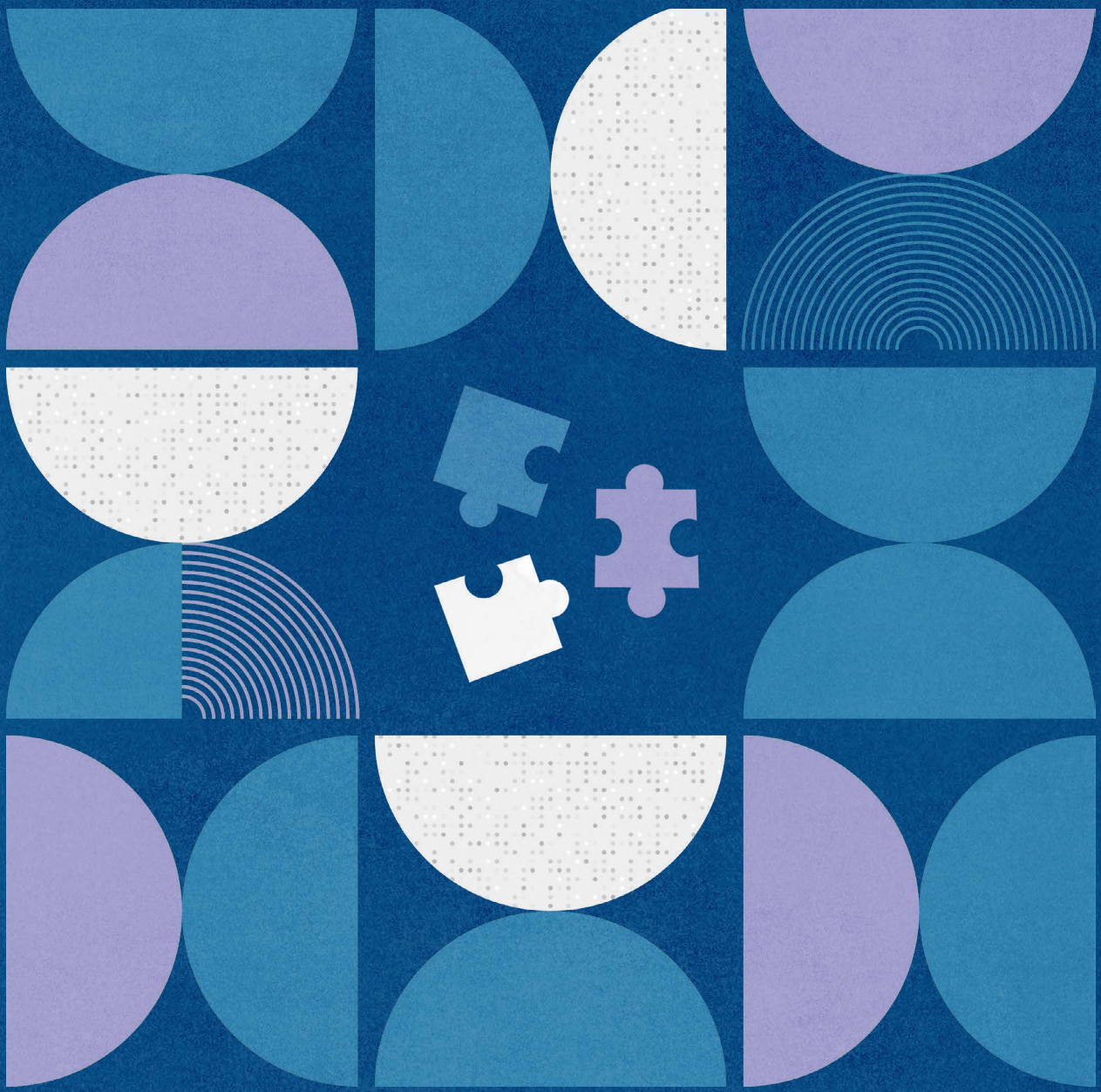
The study's identification of gender risks across the GenAI value chain is accompanied by potential pathways to "un-risking". Some examples of the pathways discussed in this report include open-source databases for the dataset preparation stage, AI red teaming for models, and feedback loops for user grievances in model deployment.

From the articulation of these findings, the study mapped possible areas of concern — primarily for businesses and states in the form of ecosystem assets. For businesses, the team developed a self-assessment risk toolkit, that uses the mapping of human rights to specific sub-components of the value chain to highlight areas or sites of possible mitigation. Similarly, using the regulatory lens applied to unpack the GenAI landscape, the team also drafted a policy brief for policymakers and state actors to support businesses, and to help protect human rights. These assets have been included within this report (as Assets 1 and 2) for usage by these stakeholders. While these assets should help businesses and states to understand how to address these human rights concerns, the study urges stakeholders to nuance these assets according to the context they are used in.

As a field and as a market, GenAI enjoys interest and use from businesses, regulatory attempts from state figures, and growing reach within contemporary society. However, at the very least, GenAI needs to be seen as a work-in-progress. Progress on GenAI is meaningfully possible only when its risks have been addressed, to allow people to be exposed to the technology without their rights and well-being endangered. This study has highlighted the potential of a value chain and human rights framing for addressing risk areas in GenAI technology and hopes that this will lead to a more equitable and inclusive approach when designing, developing, deploying, and managing such critical digital systems.

PART IV

The Building Blocks of Our Research



PART IV

The Building Blocks of Our Research

Over the course of this study, the team developed a number of assets to understand and map the risks, harms and human rights implications of GenAI. Risk assessment toolkits built to identify the potential areas of risk are housed in this section, primarily for businesses to self-administer. A policy brief that highlights suggestions for policymakers is also included. Finally, this section consolidates relevant explanations and definitions on topics like the UNGPs and the segments of the value chain.

ASSET 1 | Risk Assessment Toolkits (RATs): Primarily for Businesses

Context: The risk assessment toolkits (RATs) attempt a gender un-risking of GenAI value chains. Drawing inspiration from Principle 19 of the UNGPs — an operational principle outlining how businesses can leverage the results of diligently conducted internal assessments — the team has created a trio of RATs for businesses to take stock of their approach to gender-related GenAI risks. The RATs allow stakeholders working on GenAI development and deployment to consider the impact and value chain-level implications of their approaches and decisions.

Principle 18 of the UNGPs suggests that businesses should conduct such evaluations prior to the launch of a business offering. The RATs have been designed to also be considerate of systems that might already have been developed or deployed.

Frequency and scope of usage: Additionally, due to the nature of human rights and the evolution of GenAI systems, it is advisable for businesses to visit these self-assessments frequently, and update and address the risks accordingly. With the complexity and manifestation of GenAI systems, these RATs do not provide suggestions on how to improve systems, as **individual evaluations are necessary** to understand the (i) risks created, (ii) resources required, (iii) devising the mitigation, remediation, and remediation strategies, (iv) and the impact of addressing.

Navigating the toolkits: The RATs are provided as 'yes-no' questions in the following sections, in the order of the value chain stages: dataset preparation, modelling, and model deployment. The self-assessment questions are presented in table format, alongside specifications on which value chain sub-component is involved and which of the three stakeholder types — private business, state/state actors, society and civil society organisations (CSOs) — the question primarily addresses. Some of the self-assessment questions can also refer to all three stakeholder types (the Trifecta). Trifecta questions are denoted by an asterisk(*) at the end of such questions.

The team is hopeful that these RATs will help businesses explore areas where potential human rights concerns can be addressed. The team also requests the relevant people within the stakeholder groups to share with it the results of these self-assessments to enable a more comprehensive RAT to be evolved and more specific needs to be addressed.

Question Set 1: Data Labourforce

For: Business and Trifecta

QUESTIONS	RESPONSE
If you **use contractors for procuring data** for model development, then do you look for data firms with gender-diverse workforces?	**Yes / No**
As a ***business that prepares datasets*** as per client specifications, is your labourforce made up of gender identities beyond cis-men?	**Yes / No**
As a ***business that prepares datasets,*** do you have measures in place that ensure the continued participation and inclusion of the different communities producing data for the firm?	**Yes / No**
As a ***business that prepares datasets,*** do you have measures that help data workers improve capacity to fulfil newer and more complex data requirements while ensuring gender-based representation and inclusion?	**Yes / No**
As a ***business that prepares datasets,*** have you iteratively developed and implemented **sensitisation measures and training** that help data workers minimise gender biases in data preparation?	**Yes / No**
Does the ***business preparing training data*** make its **sensitisation measures and training** available for public input, scrutiny and feedback?	**Yes / No**
Does the ***business preparing training data*** directly involve and represent data workers, locals, and affected communities in developing its **sensitisation measures and training?**	**Yes / No**
When expanding the data labourforce, have you made **concerted efforts to make more long-term inroads with different local communities** to include more diverse representation in data work?	**Yes / No**
Has worker-side capacity-building for adopting and using digital technologies been **tested and arranged?**	**Yes / No**

Question Set 2: Contracts and Manufacturing-Side Practices

For: Business and Trifecta

QUESTIONS	RESPONSE
As a ***business contracting data firms*** to prepare datasets, do your **contracts include requirements for gender representation** beyond cis-men and cis-women?	**Yes / No**
As a ***business that prepares curated datasets,*** do you have screening, filtering, and removal measures for any **personally identifiable information (PII)** that may have reached the datasets furnished?	**Yes / No**
In addition to preparing accessible information on how data is prepared, does an active, persistent line of communication exist between the data generators and the model developers?	**Yes / No**
Have you prepared easily accessible documentation (for parties further along the value chain) that includes information that can be used to make decisions about using data in their respective development and deployment processes?	**Yes / No**
If you are making the shift to improved representation in your datasets, have you planned out the ways the production process will change and in what increments?	**Yes / No**
Have you dedicated people and resources to information-sharing across the ecosystem and its stakeholders? *Information-sharing includes notes and records on challenges, experimentation and best practices related to GenAI-relevant data.*	**Yes / No**
If you are a Global North-based business, then have you developed and implemented mechanisms to include and adapt to Global South contexts and variations?	**Yes / No**
Are you consistently watchful and mindful of the emergence and adoption of new kinds of data and new data requirements?*	**Yes / No**

QUESTIONS	RESPONSE
Have you planned and arranged methods and processes that help the business adapt its gender fairness and inclusion approaches to new technologies, capabilities, and requirements?	**Yes / No**
Have you committed people and resources to monitoring the production and circulation of GenAI-related data that your business produces?	**Yes / No**
Have you committed people and resources to your GenAI data production's government compliance? *Government compliance entails understanding, operationalising, and ensuring the business' adherence to state-issued regulations and policies.*	**Yes / No**
Does the organisation have people and resources dedicated to the systemic monitoring of data production to understand, operationalise and maintain compliance with industry standards and guidelines?	**Yes / No**
Does the organisation have people and resources dedicated to the systemic monitoring of the use of data to consistently collect and operationalise feedback and inputs from external stakeholders?	**Yes / No**

Question Set 3.1: Pre-Training Data

For: Business and Trifecta

QUESTIONS	RESPONSE
Do you document movements in the automation, training, and models that go into assembling the pre-training data corpus?	**Yes / No**
Have you prepared a mechanism and format for disclosing how the pre-training dataset(s): was assembled, the type of methods and sources used, and the automation?	**Yes / No**

QUESTIONS	RESPONSE
Have you disseminated documentation around the technological means that help build the pre-training data corpus available to scrutiny from external stakeholders?	**Yes / No**
Are you transparent about the language(s) prioritised and prevalent in the training dataset?	**Yes / No**
When crawling for publicly available data, have you developed blacklists and countermeasures for the kind of material and information that needs to be discarded and filtered from the pre-training data corpus? *Reasons could range from the potential for harm to potential privacy risks.*	**Yes / No**
Are content blacklists and filtering mechanisms built using insights and consultation from stakeholders (outside the business), and from a wide variety of contexts (like affected communities and experts working in gender inclusion)?	**Yes / No**
Have people and resources been committed to the long-term, dynamic, and participatory development of gender-inclusive pre-training data?	**Yes / No**
Are the attempts at filtering harmful material from pre-training data documented and made publicly available in an accessible and detailed format?	**Yes / No**
Have there been attempts to eliminate material whose outdated nature makes it a risk for some form of gendered harm or discrimination? *For example, books from the nineteenth century may be at greater risk of extensive sexism and misogyny.*	**Yes / No**
Are the sources used to develop the pre-training data corpus publicly disclosed in adherence with global **data provenance standards?**	**Yes / No**
Do you make pre-training datasets publicly available, while protecting Personally Identifiable Information (PII)?	**Yes / No**

Question Set 3.2: Pre-Training Data

For: The State/ State Actor

QUESTIONS	RESPONSE
Have you produced standards or guidelines on businesses' obligations to be transparent about the pre-training data that is prepared?	**Yes / No**
Have you released **data provenance standards** focusing on pre-training data for GenAI value chains?	**Yes / No**
Have you released guidelines or regulations for businesses making pre-training datasets available to external, public scrutiny?	**Yes / No**

Question Set 3.3: Pre-Training Data

For: Society and CSOs

QUESTIONS	RESPONSE
Have you developed methods and frameworks for dissecting datasets and identifying problems and fault lines for different gender identities, as well as other social groups?	**Yes / No**

Question Set 4.1: Training Data

For: Business and Trifecta

QUESTIONS	RESPONSE
As a ***firm requiring training datasets,*** do you include gender representation requirements for identities beyond cis-men and cis-women?	**Yes / No**
When preparing training data, do you have procedures to ensure representation and safe inclusion of gender identities who have struggled with being meaningfully included, represented, and recorded in data?	**Yes / No**

QUESTIONS	RESPONSE
If you have processes for including low-resource populations and gender identities, then do they include measures for locating and eradicating PII?	**Yes / No**
Have you made the methods and sources used to develop training data for underserved populations available to external stakeholders for public scrutiny?	**Yes / No**
Have attempts been made to make available publicly **benchmarks and representative datasets** that can act as representation and fairness baselines for other datasets in a given context?*	**Yes / No**
Have information and accounts of the processes and methods behind developing a particular context's **benchmark and representative datasets** been disclosed for public consumption?*	**Yes / No**
Have you considered and put in place some measures to address gender implications and requirements of datasets when material is being reused?	**Yes / No**
Do your data practices account for intersectional representation of different gender identities, accounting for identities beyond gender, to include things like caste, location, religion?	**Yes / No**
When preparing datasets, have you accounted for the gender differences in the different languages' structure and use?	**Yes / No**

Question Set 4.2: Training Data

For: The State/ State Actor

QUESTIONS	RESPONSE
Have **state guidelines** outlining the nature and kinds of gender biases, as well as the harms they perpetuate, been prepared?	**Yes / No**
Have **state guidelines** on gender biases and harms involved affected communities, veteran experts, and grassroots players in their development?	**Yes / No**

Question Set 5: Adversarial Data

For: Business and Trifecta

QUESTIONS	RESPONSE
Are processes like 'red teaming' used to cultivate, strengthen, and improve incorporation of adversarial data?	**Yes / No**
Do you consistently conduct red teaming exercises and expand stockpiles of adversarial data?	**Yes / No**

Question Set 6: Feedback Data

For: Business and Trifecta

QUESTIONS	RESPONSE
Is the feedback data collection process designed to help the GenAI asset's gender fairness improve?	**Yes / No**
Is the feedback data collected (or a record of it), the measures put in place to strengthen system design, and the impact of those changes collected, disclosed and shared publicly?	**Yes / No**

RAT TWO | The Model Toolkit

Question Set 1: Pre-Trained Model

For: Business and Trifecta

QUESTIONS	RESPONSE
Have you tested and discerned the pre-trained models' overall capabilities to determine the areas where it can provide viable answers versus where it struggles to generate outputs?	**Yes / No**
Have you developed metrics for understanding the gender-related behaviours, risk, and harmfulness for the pre-trained model?	**Yes / No**
Have you completed an exhaustive review of approaches to GenAI models before choosing this development path?	**Yes / No**
Is it viable for you to develop models that are more curated in the training data used, like small-language models (SLMs)?	**Yes / No**

Question Set 2: Trained Model

For: Business and Trifecta

QUESTIONS	RESPONSE
Have you tested the pre-trained model's abilities to depict different gender identities without stereotyping or providing hate content?	**Yes / No**
Have you developed metrics for understanding the gender-related behaviours, risk, and harmfulness regarding trained models?	**Yes / No**
Have you ensured that your gender-related metrics for the pre-trained and trained models provide a usable and reliable account of the changes in the behaviours and capabilities after training?	**Yes / No**
Have you developed an additional layer of mitigation and safety-oriented prompts that work alongside the end-users' prompts?	**Yes / No**

Question Set 3: Content and Usage Policies (CUPs)

For: Business and Trifecta

QUESTIONS	RESPONSE
Have you developed **model policies and terms** that define the types of requests and uses the model is not to be used for under any circumstances?	**Yes / No**
Do your **model policies and terms** have a gender component, focusing on the different forms of gender-based violence, risks, harms, and unfairness that the model cannot be allowed to engage in?	**Yes / No**
Have the **model policies and terms** been made publicly available and diligently presented to the user base or the larger GenAI ecosystem?	**Yes / No**
Do **monitoring** efforts inform the continual development of **model policies and terms** and is the impact of this disclosed to the GenAI ecosystem?	**Yes / No**
Do feedback and input mechanisms exist to allow external stakeholders to comment on and add to the **model policies and terms** transparently and easily?	**Yes / No**
Have you developed and documented a constantly-evolving list of high-risk situations and use cases where GenAI is not to be used, and should refrain from acting on requests?	**Yes / No**

Question Set 4: Tech Enforcement of CUPs

For: Business and Trifecta

QUESTIONS	RESPONSE
Do the **model policies and terms** result in the development and integration of mechanisms that prohibit and refuse the enumerated harmful and risky uses of the model?	**Yes / No**

QUESTIONS	RESPONSE
Are the **model policies and terms** used to inform the processes by which pre-training and training data get filtered for prohibited and harmful content?	**Yes / No**
Do you have the ability to remotely disable the use of your model(s) and any relevant assets for parties and use cases violating your **model policies and terms?**	**Yes / No**
Is the list of high-risk GenAI uses that should not be engaged in integrated into the model's responding behaviours?	**Yes / No**

Question Set 5.1: Evaluations

For: Business and Trifecta

QUESTIONS	RESPONSE
Have you made it possible for external stakeholders to build and run evaluations of your models?	**Yes / No**
Have you developed evaluations that gauge models' behaviours for different gender identities?	**Yes / No**

Question Set 5.2: Evaluations

For: The State/ State Actor

QUESTIONS	RESPONSE
Have pre-mass-release guidelines or benchmarks reflecting basic requirements for models' gender fairness and inclusion been developed?	**Yes / No**

Question Set 6.1: Red Teaming

For: Business and Trifecta

QUESTIONS	RESPONSE
Is there an AI red teaming process in place for the models being developed?	**Yes / No**
Does the red-teaming include people and groups from outside the developer firm?	**Yes / No**
Do AI red teams actively include different gender identities and have a gender-oriented stress testing programme?	**Yes / No**
Is your AI red teaming an iterative process, where people test the model as new measures and controls get integrated?	**Yes / No**
Have you opened the model up to **unrestricted** stress testing, experimentation, and jailbreaking attempts from a self-selecting but transparent cadre of experts, researchers, and journalists?	**Yes / No**

Question Set 6.2: Red Teaming

For: The State/ State Actor

QUESTIONS	RESPONSE
Do basic requirements for what red teaming needs to involve, in terms of stress testing and cross-stakeholder testing, exist?	**Yes / No**

Question Set 7: Monitoring and Adaptation

For: Business and Trifecta

QUESTIONS	RESPONSE
Have you attempted to measure the differences between pre-trained and trained models through the quality of answers that involve different gender identities?	**Yes / No**

QUESTIONS	RESPONSE
Do you have methods and processes in place to compare the differences in the pre-trained and trained models' gender inclusivity and fairness?	**Yes / No**
Have you dedicated people and resources to **monitoring** the emerging use cases, trends, and scenarios for the model?	**Yes / No**
Is the information gained from **monitoring** utilised to alter the model and develop countermeasures to maintain or enhance the models' gender fairness, and to reduce its gender harms?	**Yes / No**
Is the information gained from system monitoring shared with policymakers, regulators, and the wider GenAI community and user base in a way conducive to making GenAI safer vis-a-vis gender-related harms and misuse?*	**Yes / No**

RAT THREE | The Deployment Toolkit

Question Set 1.1: GenAI Systems and Services

For: Business and Trifecta

QUESTIONS	RESPONSE
Do you have the remote ability to disable your system or services, both selectively and wholly, to disable harmful capabilities until meaningfully robust solutions are found?	**Yes / No**
Do opportunities for the sharing of research findings and risk-harm concerns with businesses or the larger GenAI landscape exist?	**Yes / No**
Have you set up a reporting or feedback system for affected individuals and parties to flag issues and report problems?	**Yes / No**
Do you have a process to report safety incidents to external stakeholders as and when they occur?	**Yes / No**
On finding risk and harm areas, do you proactively attempt to search and rectify them or similar hazards in other areas of your GenAI operations?	**Yes / No**
Do your auditing, testing, remediation efforts, and feedback loops help shape the organisation-level approaches and procedures for GenAI development and deployment?	**Yes / No**
Have you tested and gathered evaluation information regarding the GenAI model(s) you will use to develop a comprehensive understanding of the model's gender risks, harms and failings?	**Yes / No**
Have you developed and tested measures to manage the GenAI model's potential for gender-related harms before releasing the system or service to a large audience?	**Yes / No**
Have you looked into measures that attempt to directly integrate legal and human rights frameworks into output generation, through approaches like **Constitutional AI?**	**Yes / No**

QUESTIONS	RESPONSE
Do you share information on how you manage gender risks and harms with the wider public or the industry?	**Yes / No**
Have you fine-tuned the model(s) underlying your services for the domains you intend to expose them to?	**Yes / No**
Have you committed people and resources to continuously developing safeguards and countermeasures against jailbreaking your service/system's underlying models?	**Yes / No**
Before you take up an AI model for the solution, service, or system for its intended purpose, have you extensively explored and discarded all other alternatives to involving GenAI?* *Base question: Is the GenAI you are choosing, the route to solving your problem statement?*	**Yes / No**
Are you developing literacy and educational programs that help people handle, use, and understand the responsible use and implications of GenAI?*	**Yes / No**
Do large, accessible, multi-stakeholder gatherings exist, where different actors can present challenges, dangers, solutions in development, and the potential for shared practices, standards and norms?*	**Yes / No**

Question Set 1.2: GenAI Systems and Services

For: The State/ State Actor

QUESTIONS	RESPONSE
Have you considered developing and introducing sector-specific regulations for GenAI?	**Yes / No**
Have you committed people and resources to forming sites and methods through which people can report the harms they experience or witness? *Grievance redressal mechanisms such as hotlines and helplines could be considered basic examples of such commitments.*	**Yes / No**

QUESTIONS	RESPONSE
Do you have requirements from services to report technical and safety information for protecting citizens' dignity, privacy, and fair treatment? *This could include service policy and terms enforcement measures and PII erasure strategies involved.*	**Yes / No**
Has some kind of **"safe heaven"** or **"investigator's protection"** been arranged to aid publicly inclined investigations and research into GenAI by researchers and investigations?	**Yes / No**
Does the mechanism to compel or order a business to remotely shutdown its GenAI service's operations and public availability exist?	**Yes / No**
Have you developed, or started developing, distributions for responsibilities and obligations members of different value chain components must shoulder and abide by?	**Yes / No**

Question Set 1.3: GenAI Systems and Services

Society and CSOs

QUESTIONS	RESPONSE
Do you have a strategy to legally protect your researchers, journalists, and technical experts as you work to test and dissect GenAI to understand its potential dangers and harms for the public?	**Yes / No**
Is it regular practice to convey findings and concerns on GenAI to businesses?	**Yes / No**
In addition to thinking about gender and GenAI, are you developing kits and research on identifying and addressing intersectional social fault lines in GenAI technologies?	**Yes / No**

Question Set 2: Content and Usage Policies

For: Business and Trifecta

QUESTIONS	RESPONSE
Have you developed **service policies and terms** that define how the service is to be used, its performance limitations, and the forms of use that are prohibited?	**Yes / No**
Do you have a monitoring process/system setup that tracks uses of the service that go against **service policies and terms** and is this information made publicly available to the GenAI ecosystem?	**Yes / No**
Do the service policies and terms have a dedicated gender component, outlining different kinds of harmful or unfair uses that have been flagged and prohibited?	**Yes / No**
Have you prepared a workforce, either as contractors or an in-house division of some kind, to monitor, moderate, and protect the use of your GenAI service or system?	**Yes / No**
Do your service policies and terms have the potential to be weaponized against research and journalistic efforts meant to test and explore GenAI services and systems in the public interest?	**Yes / No**

Question Set 3.1: Prompts

For: Business and Trifecta

QUESTIONS	RESPONSE
Do you have measures like privacy filters that eliminate any **personally identifiable information (PII)** that may have gotten included in task requests submitted to the GenAI service?	**Yes / No**
Are countermeasures for harmful prompts continually being developed, integrated into live services, and tested for endurance and effectiveness?	**Yes / No**

QUESTIONS	RESPONSE
Have you prepared a basic guide containing well-tested prompts that can help improve gender-friendliness and inclusion?	**Yes / No**
Do you monitor patterns in what kind of tasks and work your service is used for?	**Yes / No**

Question Set 3.2: Prompts

For: The State/ State Actor

QUESTIONS	RESPONSE
Are you developing requirements and guidelines for protecting users' privacy?	**Yes / No**

Question Set 4.1: Outputs

For: Business and Trifecta

QUESTIONS	RESPONSE
Have you arranged ways and sites where user bases can directly present problems and grievances to personnel and teams involved in development and deployment?	**Yes / No**
Are you able to note and act on user reports of risks and harmful outcomes your service produces?	**Yes / No**
Do you rely on your user base to flag risks, harms and malicious uses your service is involved with?	**Yes / No**
Have you considered adding some kind of tag, tracer, or mark that can be used to identify and separate human-produced material from GenAI outputs?	**Yes / No**

QUESTIONS	RESPONSE
Do businesses have some form of **dissection and dissemination protocols** in place to explore safety incidents' possible causes and risk factors, and to share the details surrounding GenAI failings with the wider ecosystem for ideation and solution-making reasons?	**Yes / No**
Have you installed additional layers of safety measures that mediate and treat model outputs for gender safety and fairness before reaching the end-users as a finished output?	**Yes / No**
Have you integrated mechanisms that allow the system's outputs to be contested by people?	**Yes / No**

Question Set 4.2: Outputs

For: Society and CSOs

QUESTIONS	RESPONSE
Are there well-updated, publicly available records and case stockpiles that track safety incidents and broad areas of weak gender safety?	**Yes / No**

Introduction/Problem Statement

The need for state intervention in AI emerges from its immense potential in enabling high-level processes such as problem-solving and decision-making through machines.[232] This potential has effects that can diffuse not just across the AI landscape but also sectors that have been migrating to AI for enhancing effectiveness. In this context, the importance of well-crafted policies cannot be overstated.[233] Without policy guardrails and clear regulatory conditions, vendors and users may hesitate to venture into the realm of GenAI. In the context of the research conducted by Aapti Institute, such migration affects women and gender minorities who are often under-represented or unrepresented in the making of these services in several ways, as already described: through exclusion, non-representation, lack of training, and the absence of strong policy guardrails. The Report on Preventing Discrimination Caused by the Use of AI by the Committee on Equality and Non-Discrimination, operating under the Council of Europe, has explicitly stated that women and gender minorities often face a higher degree of discrimination through the use of AI.[234] The historical exclusion of women from healthcare data, to give an example, has led to a disproportionately clearer picture of exclusively male health and bodily responses.[235]

This creates a need to understand the various ways in which the state can be involved in developing AI in a manner that takes gender sensitivities into consideration. This study addresses this issue by leveraging a stakeholder approach that finds its origin in the United Nations Guiding Principles on Business and Human Rights (UNGPs). This brief anchors its approach to state regulation through the UNGPs, and by referencing international and national instruments and documents to highlight the contemporary landscape.

[232] NITI Aayog (2018), "National Strategy for Artificial Intelligence", last retrieved April 10, 2024

[233] Lee, P., L. Lucchini and M. Seng Ah Lee (2024, January 18), "Walking the tightrope: As generative AI meets EU regulation, pragmatism is likely", Deloitte, last retrieved April 10, 2024

[234] Council of Europe (2020), "Preventing discrimination caused by the use of artificial intelligence", Council of Europe report, last retrieved April 10, 2024

[235] Kundu, A. (2024, March 11), "The AI Act's gender gap: When algorithms get it wrong, who rights the wrongs?", Internet Policy Review, last retrieved April 10, 2024

Existing Policy Initiatives

Regulation in AI can take many forms, but an analytical model that takes into account all the factors affecting the AI service is essential. There are three interacting layers in this model: social and legal, ethical, and technical.[236] The social and legal layer focuses on norms, regulations, and legislation, and is more long-term. The ethical layer addresses frameworks and principles for fairness, and is a mid-term layer. The technical layer discusses data governance, algorithm accountability, and standards, and is deemed to be a short-term element. The first two layers can be addressed by states through the development of standards and principles for AI algorithms.[237] This blended approach would allow states to create and curate global principles for emerging technologies like AI systems. An efficient manifestation of this is when there are clear lines of responsibility for every layer to an entity or individual being held responsible.[238]

While most policy instruments addressed in this section discuss regulation in AI, there is seldom specific mention of GenAI. The reason is that all ethical AI questions that emerge in GenAI within the scope of this study surround human agency and oversight, which applies to most AI systems. Additionally, while most of these instruments do not address gender bias in AI systems and services, they encompass any ensuing gender human rights abuse that may occur at any stage of the value chain.

International covenants and documents

International covenants like the UDHR, ICCPR, ICESCR, and CEDAW have all included gender in their non-discrimination clauses, thereby potentially allowing non-binary persons to also avail of the benefits under these principles. Article 7[239] of the UDHR enshrines the right to non-discrimination[240] and equality before law for all persons as does Article 26 of the ICCPR. The ICESCR also lays down the right to self-determination[241] for every person. However, the overall language of the document makes references to only 'men' and 'women', leaving the rights of non-binary genders ambiguous. The use of CEDAW[242] to expand the definition of 'gender' to differentiate between a person's sexual orientation and gender identity and derive our understanding of gender-based

[236] Gasser, U. and V.a.F. Almeida (2017), "A Layered Model for AI Governance", IEEE Internet Computing, 21(6), 58–62
[237] Ibid
[238] Sanyal, S., P. Sharma and C. Dudani (2024), "A Complex Adaptive System Framework to Regulate Artificial Intelligence", Economic Advisory Council to the PM (EAC-PM/WP/26/2024)
[239] Article 7, The Universal Declaration of Human Rights (1948)
[240] Article 26, The International Covenant on Civil and Political Rights (1966)
[241] Article 1, The International Covenant on Economic, Social and Cultural Rights (1966)
[242] The UN Committee on the Elimination of Discrimination against Women (1979), Convention on the Elimination of All Forms of Discrimination against Women

discrimination through it has also been used (See Part III, Asset 4). India is a signatory to all the above covenants, and has ratified the ICCPR and ICESCR, making these international frameworks relevant to the understanding of GenAI in the Indian context.

These human rights instruments are authorities on our understanding of human rights impacts in international law. When we discuss the under-representation, exclusion, and human rights abuses of women and gender minorities across the three stages of the value chain, these international instruments act as foundational principles for mitigation of AI bias and consideration of fairness.

Multilateral initiatives

The Recommendation of the Council on Artificial Intelligence, a legal instrument devised by the Organisation for Economic Cooperation and Development (OECD), sets out the principles of responsible stewardship for trustworthy AI services.[243] This responsibility is towards stakeholders who are at risk due to the disparate impacts of AI in labour markets, inequalities, and the implications for democracy and human rights, privacy and data protection, and digital security.

Beyond this, the recommendations also outline how national policies can incorporate trustworthy AI in their language. These recommendations span a broad array of investing in AI research and development along with fostering a digital ecosystem for AI through international cooperation, shaping an enabling policy environment that supports innovation in a controlled environment. It also emphasises building human capacity and a fair transition for a workforce affected by AI. The recommendations consist of a well-rounded set of principles that states can introduce in their AI and gender bias policies.

In June 2023, the Council of Europe (CoE) introduced the Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law to recognise that the design, development, use, and decommissioning of AI systems comes at a cost to human dignity and individual autonomy.[244] The Convention is the first legally binding international treaty on AI.[245] It directs special focus on the protection of human rights and the integrity of

[243] OECD (2019, May 22), "Recommendation of the Council on Artificial Intelligence", OECD Legal Instruments, last retrieved April 10, 2024

[244] Council of Europe (2023), "Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law", Committee on Artificial Intelligence (CAI), last retrieved April 10, 2024

[245] Van Kolfschooten, H. and C. Shachar (2023), "The Council of Europe's AI Convention (2023–2024): Promises and pitfalls for health protection", Health Policy, 138, 104935, hosted on ScienceDirect

democratic processes when managing risks and adverse impacts of AI. The Convention opens doors in its chapter on implementation, by mandating the protection of the persons with disabilities, children, and whistleblowers, and urges that responsible usage go hand in hand with public consultation, digital literacy, and skilling.

UNESCO's Recommendations on the Ethics of Artificial Intelligence also build on similar principles for ethical and trustworthy AI, along with proffering participatory approaches as one of the principles on which the recommendations should be built.[246] It also has a section on gender, unlike most similar instruments. Under this section, it mandates ethical impact assessments to include a transversal gender perspective.[247] It places importance on dedicating funds to 'gender-responsive schemes', and 'gender-sensitive tech' that does not exacerbate existing gaps. It also encourages female entrepreneurship and participation in the life cycle of AI, and representation in AI research in academia, top management positions, and research teams.[248] These aspects of centring gender in conversations around AI lead to improved training, representation, and design.

The European Union's AI Act

The EU's Artificial Intelligence Act (AIA) was passed in the European Parliament on March 13, 2024, and is a landmark AI Act that addresses the risks posed by fast-moving technology.[249] The Act requires developers and deployers to have clear requirements and obligations for specific uses of AI. It is the first step forward in development of trustworthy AI and the first comprehensive legal framework that addresses the safeguarding of fundamental rights, safety, and ethical principles affected by AI services.[250] The AIA has become a precedent for subsequent AI legislation such as the OECD recommendations discussed earlier as well as regional AI policies, including India's.

The AIA's approach to regulation is similar to Europe's General Data Protection Regulation (GDPR), but it concerns itself exclusively with AI. It imposes certain fairness requirements for businesses, and stipulates stringent penalties for non-compliance. These fairness requirements urge businesses to focus on

[246] UNESCO (2021), "Recommendation on the Ethics of Artificial Intelligence", last retrieved April 10, 2024
[247] Ibid
[248] Ibid
[249] Krasodomski, A. and M. Buchser (2024, March 14), "The EU's new AI Act could have global impact", Chatham House, last retrieved April 10, 2024
[250] European Commission (2024, April 4), "Shaping Europe's digital future – AI Act", last retrieved April 10, 2024

performing a gap analysis to determine existing governance structures and policies, processes, risk categories, and metrics that will need to be augmented so that statutory compliance can be achieved.[251] Businesses can do this through various risk assessment frameworks that target responsible AI through an ethics lens. Frameworks like the National Institute of Standards and Technology (NIST)'s AI Risk Management Framework (AI RMF),[252] and UNESCO's Recommendations on the Ethical Use of AI[253] can guide businesses towards useful ethics indicators.

The AIA follows a risk-based approach that is split into four different bands of risks, based on the intended use of a system.[254] Article 69 of the Act discusses 'minimal risks', which can be regulated through codes of conduct, for instance. This is followed by 'limited risk', covered under Article 52, which can be mitigated through transparency filters. Above this is the 'high risk' segment that is addressed under Article 6, which is extensively regulated through compliance and conformity assessments. Finally, Article 5 lays down 'unacceptable risks', which are prohibited.

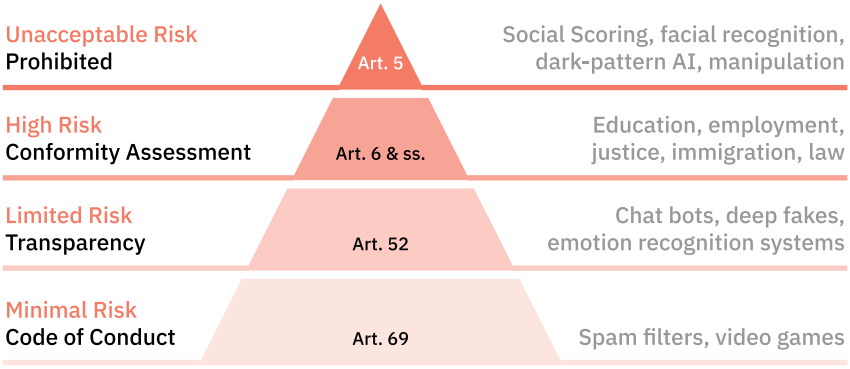


Figure 6. Source: The EU AIA's risk-based approach by Ada Lovelace Institute

The Act has paved the way for many nations and coalitions to adopt legislation that discusses a risk-based approach to AI regulation. A key consideration in the AIA is its focus on businesses taking responsibility for improving their AI models and services. While the state cracks down on risky behaviour by AI models, it plays only a punitive role in protecting human rights and leaves the rest to businesses to tackle through frameworks and extensive compliance management.[255] This framework distributes the onus of trustworthy AI across various stakeholders.

[251] Blackman, R. and I. Vasiliu-Feltes (2024, February 22), "The EU's AI Act and how companies can achieve compliance", *Harvard Business Review*, last retrieved April 10, 2024

[252] U.S. Department of Commerce and National Institute of Standards and Technology (2023), "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" [Report]

[253] UNESCO (2021), "Recommendation on the Ethics of Artificial Intelligence", last retrieved April 10, 2024

[254] Edwards, L. (2022), "The EU AI Act: a summary of its significance and scope", Ada Lovelace Institute, last retrieved April 10, 2024

[255] Fernhout, F. and T. Duquin (2024, February 13), "The EU Artificial Intelligence Act: our 16 key takeaways", Stibbe, last retrieved April 10, 2024

This study benefits from this particular risk-mapping because it allows for a degree of risk to determine mitigation pathways. Businesses and the state can work together in ensuring that gender-based risks at each stage invite a range of measures — from codes of conduct to punitive action.

India on AI regulation

The potential for GenAI to create monumental change, especially in India, has been highlighted through the country's National Strategy for Artificial Intelligence, created by NITI Aayog.[256] In terms of comprehending the AI landscape in India, the space is nascent, with government bodies feeling their way around regulatory frameworks and the potential of AI. NITI Aayog's National Strategy identifies certain focus areas where AI intervention can foster development.

Analysis of a sector-specific approach to understanding AI's potential in India throws up certain challenges: a lack of enabling data ecosystems, the research space being in its infancy, a dearth of AI expertise and manpower, along with low awareness around adopting AI in business procedures. Privacy, security, and ethical regulations still require further exploration, and intellectual property regimes remain too effete to support the research and adoption of AI.[257]

In its effort to further the AI regulation wave, the proposed Digital India Act (DIA)[258] advocates a legal and institutional quality testing framework to specifically examine regulatory AI models, algorithmic accountability, and vulnerability assessment through content moderation. The Act, which hopes to replace the Information Technology Act, 2000 (IT Act), proposes an 'accountable internet', wherein AI-based tools have an ethical responsibility to uphold constitutional rights. The Act is based on three principles — openness, accountability, and safety.[259] The Ministry of Electronics and Information Technology (MeitY) is considering reviewing the 'safe harbour' principles, which protect online platforms like X and Facebook from being accountable for the content posted on their platforms by their users.[260] The DIA hopes to create a larger framework that will comprise the DPDP Act, the DIA Rules, the National Data Governance Policy, and

[256] NITI Aayog (2018), "National Strategy for Artificial Intelligence", last retrieved April 10, 2024
[257] Ibid
[258] The Proposed Digital India Act, 2023
[259] Chauriha, S. (2023, August 8), "Explained: The Digital India Act 2023", Vidhi Centre for Legal Policy, last retrieved April 10, 2024
[260] Ibid

Indian Penal Code amendments addressing cybercrime[261] that will spotlight gender under its application.

The DIA focuses on building a regulatory set-up for the internet of our times, with vast public participation, along with multiple types of intermediaries including e-Commerce, digital media, social media, AI, OTT, gaming, and so on. User harms have also evolved from simple cybercrime to cyberstalking, doxxing, and several other nuanced forms of violence and harassment in the online space.[262] The proposed legislation adopts a principle-based approach, distinct from the EU approach, wherein it seeks to establish certain tenets of Digital India: an open internet, online safety and trust, accountability and quality of service, adjudicatory mechanisms, and the emergence of new technologies.[263] While the DIA currently does not discuss gender dimensions in its context, it is crucial that it address online gender-based violence and harassment due to their differential manifestation in the lives of women and gender minorities.

Several other Indian initiatives around building the AI ecosystem address various fairness principles. Australia, India, Japan, and the US have affirmed that their designing, developing, governing, and use of technology will be shaped through shared democratic values and respect for human rights.[264] India has also set up an AI Standardisation Committee to make AI uniform in its application across sectors.[265] One of its pillars is 'security', which cuts across all other layers of computing, data exchange, and infrastructure, to ensure that AI services are safe and secure.[266] Most of these initiatives do not adopt a gender-sensitive approach in their outlook on trustworthy AI, but international precedent indicates that gender-based harms and risks emerging out of GenAI models and applications can be addressed through these various instruments. Our expert interviews have also indicated that the DIA will likely address the adverse human rights impacts of AI on women and gender minorities.[267]

Privacy concerns emerge in AI as well owing to the manner in which data is accessed and processed. For instance, the DPDP Act[268] states that personal data can be processed only for lawful purposes after consent is obtained from the individual. However, the Act also states that it shall not apply to personal data made or

[261] Ibid
[262] The Proposed Digital India Act, 2023
[263] Ibid
[264] The United States Government, (2021, September 24), Quad Principles on Technology Design, Development, Governance, and Use, The White House
[265] INDIAai (2020, September 8), "DoT's AI Standardisation Committee releases Indian AI Stack discussion paper", last retrieved April 10, 2024
[266] Ibid
[267] Key informant interview with legal expert working in responsible AI and AI regulation.
[268] The Digital Personal Data Protection Act, Sec. 4 (2023)

caused to be made publicly available by the user. This raises concerns about how much data an AI model is entitled to scrape off the internet without prior consent from the individual who owns it.[269]

The IT rules and guidelines have incorporated principles of responsible usage through a data privacy and security lens. The IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011,[270] and the IT (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021[271] trace clear liability when it comes to ethical and safe processing and collection of sensitive and personal data. While the former recognises sexual orientation as sensitive personal data to be safeguarded, the latter mandates due diligence and grievance redressal mechanisms to ensure that intermediaries take responsible measures to safeguard against the infringement of privacy in any form that impacts gender discriminatorily.

Related to this, the human rights regulatory framework in India is dominantly derived from the Constitution.[272] Part III of the document enshrines the fundamental rights guaranteed to citizens, **a collection of civil, political, socio-religious, and cultural rights.** Article 21 of the Constitution includes the right to privacy as a fundamental right under the right to life and personal liberty of each individual. The landmark judgment in *Justice K S Puttaswamy versus Union of India* upholds this right, along with protection from interference by state and non-state actors in making autonomous life choices.[273] This argument holds ground amid emerging concerns about privacy in AI and GenAI systems. Fundamental rights also guarantee the right to equality and equal protection of the law, prohibition of discrimination based on gender and sex, as well as the right to equal opportunity. The application of these rights extends to tackling bias and exclusion as a way to redress human rights abuse caused to women and gender minorities.

[269] Pahwa, N. (2023, September 1), "Video: How Will India's Digital Personal Data Protection Law Impact Artificial Intelligence", MediaNama, last retrieved, April 10, 2024

[270] The Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules (2011)

[271] The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules (2021)

[272] The Constitution of India (1949)

[273] Justice K S Puttaswamy (Retd.) and Anr. versus Union of India and Ors., (2017), 10 SCC 1

Pathways to regulation

In considering what future governance models in AI could look like, it is important to address certain associated structural challenges.[274] The first and foremost of these is bridging the chasm of inaccessible and asymmetrical AI-based systems that policymakers and consumers cannot understand. Second, a balance must be struck between the potential benefits and risks of AI. Currently, there is a focus on risk and, while relevant, opening up spaces for a cost-benefit analysis and possibilities for trade-offs can be incorporated in designing AI models. This leverages the potential for AI to foster innovation and growth in a manner that is sensitive and representative of a diverse demographic. Lastly, aspirations for AI regulations are heavily influenced by governmental and societal consensus on what is and isn't desirable in the design of effective AI.

Together, all of these structural challenges sculpt the design requirements of future governance models for AI. These governance models must be rooted in the current issues surrounding emerging technologies.

The understanding of state regulation and intervention is substantiated by the following recommendations that help address specific human rights risks highlighted in this research. They range from developing uniform frameworks and assessments for inclusivity to gender-responsive budget allocation, promoting AI literacy, advocating AI for social good, and a host of regulatory and oversight mechanisms that can act as checks and balances in the development of AI and GenAI to include women and gender minorities within its scope.

[274] Gasser, U. and V.a.F. Almeida (2017), "A Layered Model for AI Governance", IEEE Internet Computing, 21(6), 58–62

RECOMMENDATION	RELEVANT UNGP
Creation of an inclusivity/fairness framework for gender	1, 3
Gender-responsive budget allocation	3
Cultivating AI literacy	2, 5, 8
Socially Responsible AI (SRAI)	8, 10
Regulatory oversight mechanisms	1, 3, 5, 8, 9
Public watchdogs and whistleblower protection	1, 5

Table 18. Consolidated recommendations for policy makers

1. Creation of an inclusivity/fairness framework for gender

The NIST AI RMF is considered a comprehensive risk assessment framework for AI.[275] Unfortunately, this framework has no gender dimensions to it through any direct references. It discusses diversity, equity, and inclusion in various forms throughout the AI life cycle. However, Aapti's research on the effects of gender-agnostic or gender-blind GenAI services reveals that if left unchecked for gender discrimination, GenAI services can lead to widespread harassment and violence through various use cases that often target women and gender minorities owing to their very gender identity.

On the technological side, researchers at Anthropic AI pioneered an approach, already referred to earlier, called Constitutional AI (CAI). This allows the state to create a set of principles to be created in the modelling or deployment of the AI models or services and is a way to standardise the AI-based risk mitigation processes in a systematic manner.[276] This set of principles or instructions can be used by models to supervise other AI models as well.[277]

Creating a gender inclusivity framework for businesses can enable the state to fulfil its responsibility of protecting against any human rights abuse that emerges or might potentially emerge from AI

[275] U.S. Department of Commerce and National Institute of Standards and Technology (2023), "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" [Report]

[276] Burt, A. (2023, November 1), "3 Obstacles to Regulating Generative AI", *Harvard Business Review*, last retrieved April 10, 2024.

[277] Bai, Y., S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, ... and J. Kaplan (2022), "Constitutional AI: Harmlessness from AI feedback", arXiv preprint, arXiv:2212.08073

services. This ties in with Principle 1 of the UNGPs, which holds the state responsible for protecting against such infractions. Principle 3 also urges states to enforce laws and governing policies to aid businesses and provide them with effective guidance on upholding human rights in their operations.

2. Gender-responsive budget allocation

Devoting public funding to support more women and gender minorities in the AI workforce and in AI through investment in research and development, scholarships, and training, is critical to building accountability. An 'Accountability Fund' can also be created to support research on the impacts of AI and ML on women.[278] Being responsible about fund allocation is a practice that the state can be involved in more meaningfully by creating policy measures or providing guidance to businesses, as laid down under Principle 3 of the UNGPs.

3. Cultivating AI literacy

Literacy on AI and its harms, along with its gender implications, is crucial at this juncture of AI development. Many experts in this space have flagged concerns about developers and engineers being unaware of questions surrounding bias and gender equality during their training and careers.[279] This creates a cohort of developers who do not have the capacity or sensitivity to address gender bias in the models they create. Further, with AI increasingly affecting more aspects of our lives, literacy becomes a critical requirement for every person. Experts have proposed that AI literacy must comprise the 4 Cs — **concepts, context, capability,** and **creativity.**[280] These should be covered at a fundamental level to define how AI works and its ability to be responsible and safe.[281]

At a more skill-based level, nurturing AI literacy has further impact. A report suggests that 47 percent of US jobs are at high risk of automation, putting women at risk of losing their jobs due to the number of women employed in clerical roles.[282] People with knowledge of AI systems will have a head start. To pre-empt and better plan for the future, it is important for AI literacy to cater to and be availed of by women.

[278] UNESCO (2020), "Artificial Intelligence and Gender Equality: Key findings of UNESCO's Global Dialogue", last retrieved April 10, 2024
[279] Ibid
[280] Talagala, N. (2023, December 29), "The AI Literacy Act – what is it and why should you care?", *Forbes*, last retrieved April 10, 2024
[281] Ibid
[282] Manyika, J., S. Lund, M. Chui, J. Bughin, J. Woetzel, P. Batra, R. Ko and S. Sanghvi (2017), "Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation", McKinsey & Company, last retrieved April 10, 2024

Under Principles 2 and 8 of the UNGPs, the state can set out expectations for businesses to engage in meaningful AI literacy that addresses the lived experiences of women and gender minorities — in line with the state's human rights obligations. States can also put in place adequate oversight mechanisms under Principle 5 to ensure this.

4. Socially Responsible AI (SRAI)

Instances of AI being used to screen job applicants, in education pricing, and in the criminal justice system, have led to the conclusion that AI-enabled tasks without human intervention and RAI standards for measuring their readiness can demonstrate corporate irresponsibility. This can lead to severe consequences for the reputation of the business as well as the welfare of society at large.[283] Cumulatively, this indicates the need for the state to establish guidelines for SRAI to ensure that AI is leveraged in a way that aligns with human values. SRAI must include and demonstrate representation of gender equality and diversity at its core.

Principle 8 of the UNGPs urges the state to ensure that with relevant training and support through guidelines and frameworks, businesses can be made aware of the state's human rights obligations. The state can attempt to do this through capacity and awareness-building activities.

5. Regulatory oversight mechanisms

As in the European Union, it is important that high-risk use cases of AI have an accompanying oversight mechanism that can manage this technology. Effective oversight is facilitated through a clear vision of gender equality, with consultations and deliberations with key stakeholders and groups.[284] Independent oversight and audit institutions can play a pre-emptive as well as remedial role in ensuring gender mainstreaming in the AI value chain. Specific auditors focusing on gender-friendliness can also strengthen the evidence base and systematically measure progress towards gender equality performance, based on gender impact indicators and measurable outcomes.[285]

[283] Chang, Y. and J. Ke (2023), "Socially Responsible Artificial Intelligence Empowered People Analytics: A Novel Framework Towards Sustainability", Human Resource Development Review, *Sage Journals*, 23(1)
[284] OECD (2020), "Towards a Gender-sensitive Framework for Sound Public Governance", GOV/PGC/GMG(2020)1, last retrieved April 10, 2024
[285] Ibid

Guided by Principles 1, 3 and 5 of the UNGPs, the state can take appropriate steps, including regulations and legislation, to prevent, punish, and redress all forms of gender-based discrimination, harassment, and violence. Regulatory oversight also ensures that governmental departments, multilateral institutions, and other agencies are aware of and observe the state's human rights obligations, under Principles 8 and 10.

6. Public watchdogs and whistleblower protection

Public watchdogs, or a dedicated group of people that can keep a check on the misuse of AI to perpetuate misinformation and violence, are essential to check for any threats to democracy and human rights.[286] The UN Secretary-General Antonio Guterres has expressed support for setting up a high-level AI advisory body to regularly review AI governance arrangements and provide pathways for these arrangements to align with human rights and the rule of law.[287] This body would be similar to what the International Atomic Energy Agency (IAEA) does for nuclear non-proliferation.[288] Public watchdogs like the European Data Protection Board (EDPB) have set up task forces to curb ChatGPT's potential data breaches.[289]

A case in point was Stanford University researchers' study on AI models that could screen thousands of photographs from dating sites to determine people's sexual orientation.[290] The AI application was created to highlight concerns over the threat to privacy and safety of the LGBTQIA+ community, but the project was received with condemnation and criticism by advocacy groups and academic circles. This controversy illuminates a bigger problem — social scientists do not have clear ethical guidelines to prevent them from accidentally harming people. Stanford put an institutional review board (IRB) in place, which required researchers and scientists to get its approval if they wished to research humans for any study.[291] An intersecting problem also emerges when 'big data' research does not fall under the purview of federal scrutiny through watchdogs.[292] Because of this vacuum in regulatory guidelines, sensitive personal information of women and, especially, persons of diverse SOGIESCs can be jeopardised, thereby impacting key

[286] ANI (2023, June 13), "UN Secretary-General calls for establishment of watchdog to monitor AI", *ThePrint*, last retrieved April 10, 2024
[287] Ibid
[288] Ibid
[289] Reuters (2023, April 14), "European privacy watchdog creates ChatGPT task force", *The Economic Times*, last retrieved April 10, 2024
[290] Chen, S. (2017, September 18), "AI Research Is in Desperate Need of an Ethical Watchdog", *WIRED*, last retrieved April 10, 2024
[291] Ibid
[292] Ibid

fundamental and human rights guaranteed to all persons under international law and Indian laws wherever applicable.

This branches into users acting as watchdog/ombudsman as well. Whistleblower protection as part of regulation encourages individuals to report wrongdoing within organisations with legal and financial safeguards.[293] Whistleblower protection can promote a culture of transparency and accountability, and can also be coupled with rewards that incentivise whistleblowers to come forward. This could include award of a portion of any financial penalty that is levied.[294] Newer laws like the EU Directive on Whistleblowing require member states to provide workers in the public and private sectors with effective channels to report breaches in ethical AI.[295] Creating guides for employees on their whistleblowing rights can be empowering and create awareness about how they can report wrongdoing fearlessly.

Under Principle 5 of the UNGPs, states can devise and exercise adequate oversight mechanisms to meet their international human rights obligations to aid businesses in providing more gender-inclusive AI services. These oversight mechanisms can involve external watchdogs and whistleblowers to cover more stakeholders.

State regulation is a necessary pillar to ensure that businesses are mindful of representing and upholding the human rights of women and gender minorities. Regulation looks different in differing set-ups: the EU allows for AI development but with punitive guardrails, whereas in China no company can produce AI services without proper approvals. The US is developing its own approach and is currently focused on limiting the use of AI in law enforcement and hiring.[296] All this underlines that the state is an essential pillar in the furtherance of responsible usage and user safety in AI.

[293] Jones, E. (2023), "Keeping an eye on AI", Ada Lovelace Institute, last retrieved April 10, 2024
[294] Ibid
[295] Gibson, J. (2024, January 29), "A.I. regulations doomed to fail without whistleblower protections", The Signals Network, last retrieved April 10, 2024
[296] Rozen, C. and J. Deutsch (2024, March 13), "Regulate AI? How US, EU and China Are Going About It", Bloomberg

ASSET 3 | A Value Chain for Mapping GenAI Development and Deployment: Unpacking the sites of Change in GenAI Value Chains

Chapters 4 through 7 explored the stages value chain, while highlighting sub-components and key elements in GenAI development and deployment that could host and address gender risks. Following the stage-wise exploration, it is now possible to assemble the articulation and visualisation of the previous chapters into a more comprehensive portrait, within the value chain introduced as part of this study. Figure 7 shows the entirety of this value chain formed as a result of this study.

While this study introduced the sub-components within the stages value chain, its mapping emphasised that greater detail could be more useful to approaches like gender un-risking, especially for businesses. The specific nature of sub-components thinking means that different technical elements can be mapped without restriction or boxing into particular stages.

Though more complex, this approach helps provoke the search for specific sites of change that both host risks and can potentially undergo un-risking in some form. For instance, the dataset preparation stage can be unbundled into smaller flows and chains of labour producing data, and different kinds of data for various uses, that can contribute to risks and un-risking.

For example, data in the GenAI context can also refer to adversarial data used to train models to safeguard against particular kinds of human usage, or user-provided feedback data from services like ChatGPT that use models like GPT-4. Thinking about data should not be restricted to the “beginning” of the GenAI value chain. Through the sub-components approach, models are mapped alongside AI red teaming and evaluations, which are processes that help improve models’ safety and behaviour.

Businesses often provide tailored services for the preparation of data for AI purposes. Here, two considerations behind data production for AI can be: the nature of the data labour force, in terms of gender representation, including the nature of contracts

and specifications; and data production. The inclusion of women and gender minorities and requiring non-binary genders' representation in preparing data can potentially help GenAI's underlying datasets acquire better representation and reduced bias.

The sub-components route also helps think about external access to the GenAI value chain. As the size and scope of the GenAI ecosystem grows, the question of scrutiny from people and groups outside those who administer GenAI value chains or their sub-components becomes important. Thus, the sub-components value chain also attempts to demarcate its elements into those that are either open to external stakeholders, or are capable of being opened up to beyond a sub-component's ownership or administration.

However, this demarcation of external access should not be seen as fixed or concrete. This iteration of the value chain unpacks the three broad "stages" and reimagines itself as a more continuous process. The study posits that the three stages are involved in more ways than a simple "beginning-middle-end" trajectory. Different businesses adopt different approaches regarding what they make available to external stakeholders. For example, Meta's Llama 2 is "open source", with the model and model weights statedly available for download.[297] Google's Gemma models, on the other hand, practise what they call open models, where "terms of use, redistribution, and variant ownership vary according to a model's specific terms of use, which may not be based on an open-source licence".[298]

The following diagram (Figure 7) shows the various sub-components within the stages value chain created for the purpose of this study. The team believes that this detailed view of the GenAI value chain is useful for the larger ecosystem working with such digital technologies to improve transparency and understanding.

[297] Meta Llama (n.d.), Meta Llama 2
[298] Google Open Source (2024, February 21), "Building Open Models Responsibly in the Gemini Era", Google Open Source blog

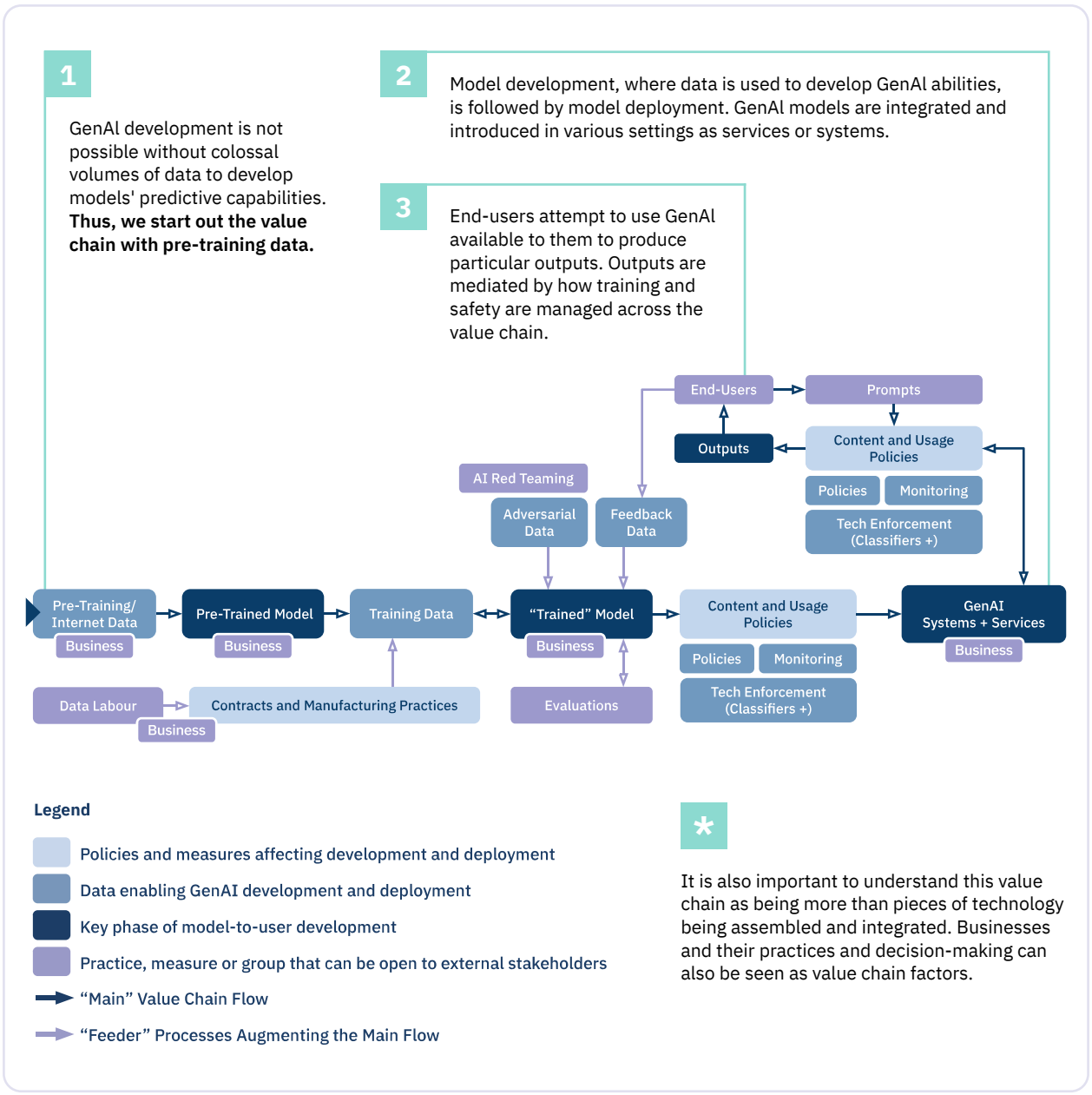


Figure 7. The Consolidated Sub-Components within the Value Chain

Alongside the new elements and departure from a panoramic overview of the GenAI development and deployment process, this magnification hosts a colour scheme that offers basic commentary on the characteristics of its components. Two colourings are of immediate note, while all of them are explained in the legend provided. The darkest blue is used to mark key phases in the pipeline between model developers and end-users, starting with the pre-trained model and "concluding" with end-users and the outputs GenAI furnishes.

Next, this iteration uses purple to highlight components where external stakeholders are involved or can potentially be involved in the current GenAI landscape. Prompts are one site where administrators or owners of businesses are not the sole drivers of GenAI outcomes, with end-users and outside parties being able to make dynamic kinds of requests to GenAI services.

The data stage(s) and types: The first noticeable major difference in the magnified value chain is the addition of a **"data stage"** focusing on processes like preparing training data. Existing research indicates that multiple forms of data can feed the GenAI value chain at different points in the GenAI life cycle. Pre-training data, or "internet data", involves colossal volumes of data which are used to develop a model's initial suite of broad capabilities. This corpus is often assembled through data sourced from publicly available sources on the internet or various third parties, ranging anywhere from the BBC to Reddit. Such data enables the development of "pre-trained models".

To improve pre-trained models for particular tasks and behaviours, models are then exposed to specifically curated data. This process is called **"fine-tuning"** and the data involved is **"training data".** Differing from the first value chain, pre-training data and training data have different purposes and preparation behind them and need to be approached differently in terms of addressing gender risks and harms. Pre-training data needs to be procured in large quantities, and can involve automated mechanisms in its gathering, creating the need for filtering and other ways to separate and purge the harmful, unfair, and hateful content that might seep into pre-training data from the internet.

Training data is prepared more directly by humans, emphasising considerations of data preparation by parties like data workers, the prudent site of examination and addressal. For example, firms specialising in preparing datasets may receive specifications for gender representation that seek half the dataset to be cis-male populated and the other half cis-female populated.[299] Such arrangements exclude gender identities outside the cis-binary, and the contracts become a site of change and potential gender-based improvement. Thus, data labour forces and their contract-related manufacturing practices are a key addition to the GenAI value chain.

[299] Key informant interview with ecosystem expert with expertise in data preparation.

Modelling: This magnified value chain **breaks "modelling" down into the pre-trained model and the trained model,** with the latter having numerous, affiliated elements "feeding" its development and growth. Models typically undergo several rounds of training in the pursuit of tuning and are exposed to training data alongside **feedback data** and **adversarial data.**

Feedback data: Data that is supposed to create GenAI feedback loops by collecting end-user inputs through various formats ("thumbs up and thumbs down" ratings).[300] The research indicates that feedback data could potentially become a source of gender-related improvement. **Adversarial data** complements the training data by bolstering models' robustness against external jailbreaking and misuse through methods like adding treated samples of adversarial uses to the training data processes.[301]

AI Red Teaming: AI red teaming works along similar logic, and involves **stress-testing** to understand a model's ability and tendency to provide prohibited, dangerous or harmful outputs.[302] AI red teams can also allow stakeholders outside the organisation that owns or administers a value chain component to contribute in making GenAI less harmful, with OpenAI's red teaming network being one example of external stakeholders contributing to testing models.

Evaluations: In addition to AI red teaming seeking to locate openings and areas where the model produces harmful outcomes, evaluations are another space for testing model capabilities. Evaluations are tests meant to measure models' performance regarding specific tasks or capabilities. Evaluations can be a part of the public-facing reporting businesses engage in for their models and GenAI services.

Content and Usage Infrastructure and Policies: Trained models are also affected by content and usage policies (CUPs) that define harmful and prohibited usage of GenAI models and services. CUPs can be technologically operationalised to help prohibit specific forms of GenAI use, as in the case of OpenAI's GPT-4, making them another site where gender concerns can be introduced and addressed. CUPs can often help make GenAI safer and less prone to harm and abuse by: (i) informing evaluations, (ii) aiding the

[300] "AI Foundation Models: initial review" (2023, May 3), Government of the United Kingdom
[301] Shekhar, R. (n.d.), "Tools for Responsible AI", INDIAai, last retrieved April 1, 2024
[302] Burt, A. (2024, January 4), "How to Red Team a Gen AI Model", *Harvard Business Review*

development of classifiers to find and track harmful content, (iii) strengthening efforts towards monitoring model use, and (iv) developing a model's capacity for refusing prompts for particular kinds of tasks.

Monitoring: Monitoring is another aspect of both 'models' and 'deployment' generative AI and services that focus on observing GenAI use. Both humans and automated mechanisms can be involved in monitoring and responding to misuses detected.[303] Usage monitoring efforts and responses also involve a model or service's CUPs, which outline and list the prohibited applications of GenAI. GenAI-oriented models like GPT-3.5 and GPT-4 are integrated into GenAI systems and services like ChatGPT, which are available to larger markets, groups, and user-bases. The firm's decision-making in matters like the service design, the way the model is deployed, the firm's CUPs, associated processes and infrastructures, and the way the user-base works with the services shapes the gender risks and harms at the deployment stage of GenAI. GenAI technologies' scale and resource-intensiveness make it difficult for actors beyond a select few to alter the entire model. A considerable amount of focus goes in mediating and working on models' output between the point of where models return responses and the point where end-users receive them.[304]

End-users: End-users are important as more than a source of feedback data or the prompts that elicit outputs from GenAI — they provide potential areas for positive gender transformation of the GenAI value chain in terms of spreading awareness and sensitisation regarding the harms GenAI can have and cause, thus fostering cultures of responsible use and involvement.

Outputs: Formats like images, videos and text are the product of GenAI's interactions with end-users and their prompts.[305] From book summaries to pornographic deepfakes, GenAI outputs can inhabit and see circulation across the internet and the public, resulting in various consequences. The AI domain is also working on **"synthetic data",** which could potentially see application in various scenarios, like serving as a substitute in situations of data shortage or scarcity.[306] Thus, outputs are an essential part of any GenAI value chain.

[303] OpenAI (2023, March 23), GPT-4 System Card.
[304] Key informant interview with expert connected to AI and ML.
[305] Deloitte AI Institute (n.d.), "The Generative AI Dossier: A selection of high-impact use cases across six major industries"
[306] De Wilde, P., P. Arora, F. Buarque, Y. Chan Chin, M. Thinyane, S. Stinckwich, E. Fournier-Tombs and T. Marwala (2024, April 9), "Recommendations on the Use of Synthetic Data to Train AI Models", United Nations University

Understanding 'gender' to strengthen inclusionary thinking

The scope of this study entails analysing how jurisprudence and thinking on gender have evolved over the years. While landscape analyses have indicated momentum towards more gender-inclusive policies, women still face implications based on their socio-normative contexts. Ecosystem initiatives aimed at enabling and empowering women solidify the need to strengthen efforts that are still required to provide women meaningful access to basic rights and services. The lead-up to the Gender Dimensions of the UN Guiding Principles on Business and Human Rights highlights the disproportionate impact that women and girls continue to face and provides guidance and recommendations on this basis.

Additionally, international human rights law (IHRL) recognises "non-binary" as a term in its official records through reports and research, and through special rapporteurs and other independent experts. However, for the term to be more effectively incorporated into the IHRL framework, strategies need to be devised[307] for *categorical enlargement, conceptual expansion, and group-conscious universal application.*

Globally, understanding gender and its nuances has been very complex. Global regulations, principles and guidelines are being leveraged to create equitable rights and policies for all gender groups. However, these approaches still require a nuanced understanding of gender-inclusionary thinking and a more substantiated approach to embed these nuances within the zeitgeist and global ecosystem.

The Sustainable Development Goals (SDGs)[308] constitute an instrumental tool in defining and promoting gender equality for women and girls globally, yet might be considered limited when understanding rights and liberties. The term 'gender' remains rooted in the binary, overlooking legal recognition for those identities that fall outside its order, such as intersex and non-binary.

[307] Smith, R. A. (2022, August 23), "How can international human rights law protect those who identify as non-binary?", OpenGlobalRights, last retrieved May 14, 2024

[308] UN Women (n.d.), "SDG 5: Achieve gender equality and empower all women and girls", last retrieved May 14, 2024

The UN Women definition of the term refers to the roles and behaviours considered appropriate for men and women at a given time in a given society. These roles and behaviours are socially constructed and changeable. Gender is part of a broader socio-cultural context that includes class, race, poverty level, ethnic group, sexual orientation, age, etc.[309] The definition does acknowledge that this is changeable and time-specific.

Similarly, the UN Committee on the Elimination of Discrimination Against Women (CEDAW)[310] expanded the definition of 'gender' to differentiate between a person's sexual orientation and gender identity. It elaborated on sex as the biological differences between males and females, and gender as socially constructed differences resulting in hierarchical relationships[311] between men and women. This definition has largely prevailed in regional mechanisms discussing gender; however, it has been unable to include non-binary gender groups.

International covenants have included gender in their non-discrimination clauses, thereby potentially allowing non-binary persons to also avail of benefits under these principles. The UDHR, ICCPR, ICESCR, and UN-CEDAW have all stated the right of non-discrimination and self-determination to every person.

India's Supreme Court, in its landmark 2014 judgment,[312] held that the **self-determination** of gender is an integral part of personal autonomy. Gender identity includes a personal sense of the body, which may be freely chosen and modified according to an **individual's experience of gender.** The Court held that a person's internal and individual experience of gender may not correspond with the sex assigned at birth.[313] Visual representations like the 'genderbread person'[314] also work to strengthen ecosystem understanding around sexual orientation and gender identities.

This study uses the rationale behind the evolution of such principles. It articulates why a more nuanced understanding could benefit businesses in their positioning on gender equality as a part of their organisational structures and processes, particularly when understanding the complexities that women and girls face whilst interacting with such technology, and for gender groups beyond the binary.

[309] UN Women (n.d.), Gender Equality Glossary, last retrieved May 14, 2024
[310] The UN Committee on the Elimination of Discrimination against Women (1979), Convention on the Elimination of All Forms of Discrimination against Women
[311] OutRight Action International (n.d.), "Gender Parity: Beyond The Binary", OHCHR, last retrieved May 14, 2024 (downloadable only).
[312] National Legal Services Authority Versus Union of India and others, 2014 INSC 275
[313] Ibid
[314] Killermann, B.S. (n.d.), "The Genderbread Person version 4", last retrieved May 14, 2024

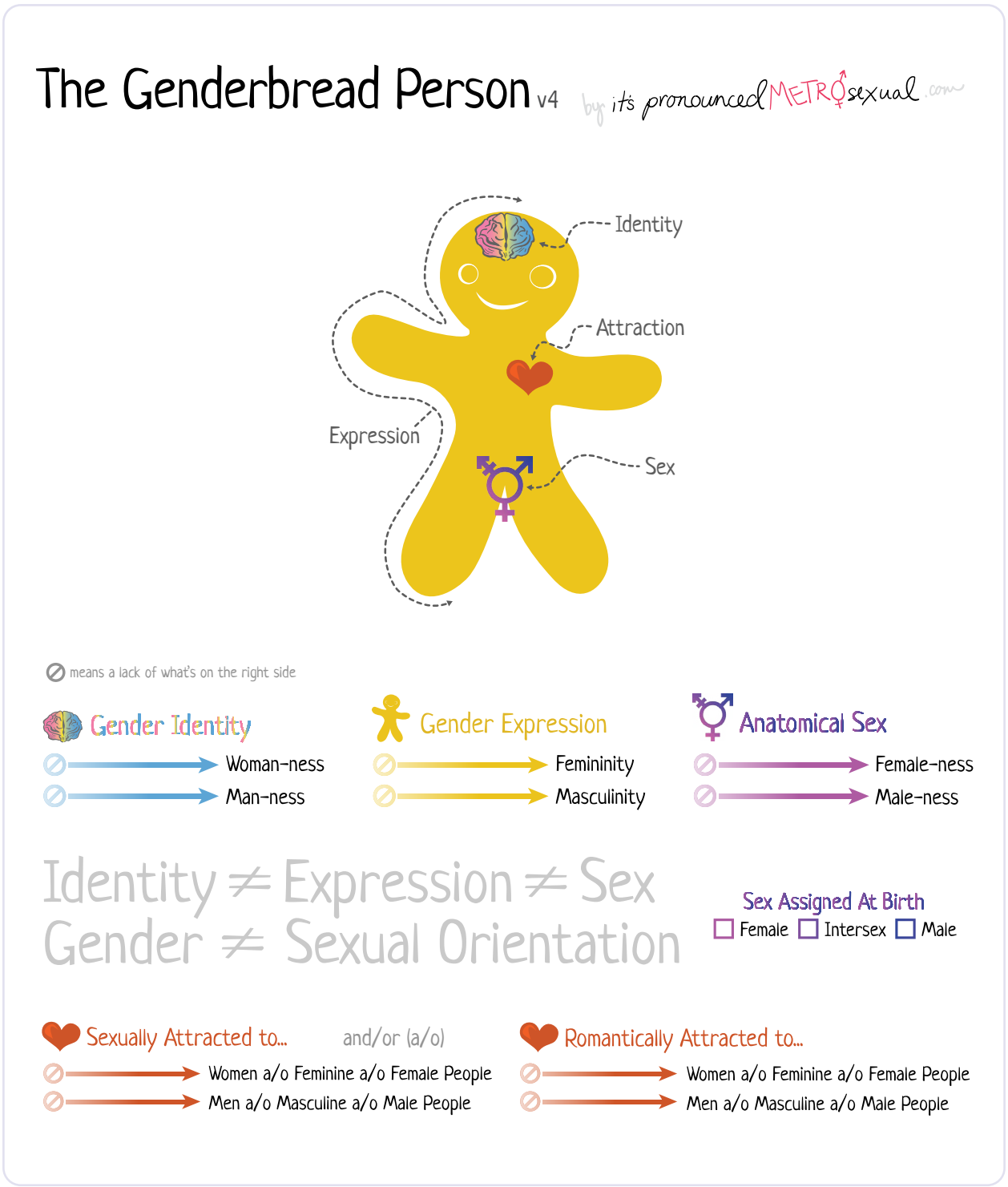


Figure 8. The Genderbread Person

ASSET 5 | Notes on Our Scope and Cross-Cutting Themes

During the course of this research, the team came across several cross-sections of the GenAI ecosystem that are impacted by factors starkly removed from gender and inclusivity. Though the research was focussed on limiting its scope to gender sensitivity in a stricter sense, the roles these cross-cutting themes play in understanding the GenAI value chain and supply chain are significant.

Intersectionality

The study examines the role women and gender minorities can play in the GenAI ecosystem, but does not capture the intersectionality in this identity. This study has not focused on the differential impact of gender bias in AI on women of colour, women with disabilities, and women from diverse caste and class backgrounds. The need for such a lens is extremely important, as several instances of race playing a major role in ML bias have been observed. In India, especially, GenAI-based technologies like facial recognition for law enforcement can often target disadvantaged or ostracised communities.[315]

Environmental costs

"Compute" can be defined both as part of the hardware (for instance, chips) that go into these models, or the number of computations needed to perform a particular task.[316] The computing power of GenAI is considerable and should be explored further through varied analyses. While the demand for compute has gone up exponentially in the past couple of years, this technology is considered environmentally and generally unsustainable. Chips are considered hazardous to produce and require an immense amount of energy.[317] Running data centres is also environmentally costly. Estimates believe that for every prompt run on ChatGPT, an entire bottle of water is used.[318] These considerations can also tie into observing the gendered impact of environmental unsustainability on persons of diverse gender as well as class and caste backgrounds.

[315] Chandran, R. and Thomson Reuters Foundation (2023, September 14), "India's scaling up of AI could reproduce casteist bias, discrimination against women and minorities", *Scroll.in*, last retrieved April 1, 2024

[316] Vipra, J. and S. Myers West (2023, October 11), "Computational power and AI", AI Now Institute, last retrieved April 1, 2024

[317] Ibid

[318] Ibid

Ancillary stakeholders

This study has also not holistically focused on the stakeholders that play a more tertiary role in this ecosystem: regulators, researchers, investors, and other CSOs. Considerations around these participants would bring in larger questions of participation, which the current study may not have been able to address due to the limitations of its scope.

ASSET 6 | Glossaries

Applicable UNGPs

The UNGPs have been utilised as a primary anchor point for this study, in order to map human rights risks and impacts to stakeholders for their mitigation. The recommendations in this report have derived relevance through the UNGPs' 'Protect, Respect, and Remedy' framework. The report has utilised the following UNGPs to frame the human rights impact in its recommendations.

The State's Duty to Protect

Guiding Principle 1
States must protect against human rights abuse within their territory and/or jurisdiction by third parties, including business enterprises. This requires taking appropriate steps to prevent, investigate, punish and redress such abuse through effective policies, legislation, regulations and adjudication.

Guiding Principle 2
States should set out clearly the expectation that all business enterprises domiciled in their territory and/or jurisdiction respect human rights throughout their operations.

Guiding Principle 3
In meeting their duty to protect, States should:

a. Enforce laws that are aimed at, or have the effect of, requiring business enterprises to respect human rights, and periodically to assess the adequacy of such laws and address any gaps;
b. Ensure that other laws and policies governing the creation and ongoing operation of business enterprises, such as corporate law, do not constrain but enable business respect for human rights;
c. Provide effective guidance to business enterprises on how to respect human rights throughout their operations;
d. Encourage, and where appropriate require, business enterprises to communicate how they address their human rights impacts.

Guiding Principle 4
States should take additional steps to protect against human rights abuses by business enterprises that are owned or controlled by the State, or that receive substantial support and services from State agencies such as export credit agencies and official investment insurance or guarantee agencies, including, where appropriate, by requiring human rights due diligence.

Guiding Principle 5
States should exercise adequate oversight in order to meet their international human rights obligations when they contract with, or legislate for, business enterprises to provide services that may impact upon the enjoyment of human rights.

Guiding Principle 8
States should ensure that governmental departments, agencies and other State-based institutions that shape business practices are aware of and observe the State's human rights obligations when fulfilling their respective mandates, including by providing them with relevant information, training and support.

Guiding Principle 10
States, when acting as members of multilateral institutions that deal with business-related issues, should:

a. Seek to ensure that those institutions neither restrain the ability of their member States to meet their duty to protect nor hinder business enterprises from respecting human rights;
b. Encourage those institutions, within their respective mandates and capacities, to promote business respect for human rights and, where requested, to help States meet their duty to protect against human rights abuse by business enterprises, including through technical assistance, capacity-building and awareness-raising;
c. Draw on these Guiding Principles to promote shared understanding and advance international cooperation in the management of business and human rights challenges.

The Corporate Responsibility to Respect

Guiding Principle 11
Business enterprises should respect human rights. This means that they should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved.

Guiding Principle 12
The responsibility of business enterprises to respect human rights refers to internationally recognized human rights – understood, at a minimum, as those expressed in the International Bill of Human Rights and the principles concerning fundamental rights set out in the International Labour Organization's Declaration on Fundamental Principles and Rights at Work.

Guiding Principle 13
The responsibility to respect human rights requires that business enterprises:

a. Avoid causing or contributing to adverse human rights impacts through their own activities, and address such impacts when they occur;

b. Seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts.

Guiding Principle 14

The responsibility of business enterprises to respect human rights applies to all enterprises regardless of their size, sector, operational context, ownership and structure. Nevertheless, the scale and complexity of the means through which enterprises meet that responsibility may vary according to these factors and with the severity of the enterprise's adverse human rights impacts.

Guiding Principle 15

In order to meet their responsibility to respect human rights, business enterprises should have in place policies and processes appropriate to their size and circumstances, including:

a. A policy commitment to meet their responsibility to respect human rights;
b. A human rights due diligence process to identify, prevent, mitigate and account for how they address their impacts on human rights;
c. Processes to enable the remediation of any adverse human rights impacts they cause or to which they contribute.

Guiding Principle 16

As the basis for embedding their responsibility to respect human rights, business enterprises should express their commitment to meet this responsibility through a statement of policy that:

a. Is approved at the most senior level of the business enterprise;
b. Is informed by relevant internal and/or external expertise;
c. Stipulates the enterprise's human rights expectations of personnel, business partners and other parties directly linked to its operations, products or services;
d. Is publicly available and communicated internally and externally to all personnel, business partners and other relevant parties;
e. Is reflected in operational policies and procedures necessary to embed it throughout the business enterprise.

Guiding Principle 17

In order to identify, prevent, mitigate and account for how they address their adverse human rights impacts, business enterprises should carry out human rights due diligence. The process should include assessing actual and potential human rights impacts, integrating and acting upon the findings, tracking responses, and communicating how impacts are addressed. Human rights due diligence:

a. Should cover adverse human rights impacts that the business enterprise may cause or contribute to through its own activities, or which may be directly linked to its operations, products or services by its business relationships;
b. Will vary in complexity with the size of the business enterprise, the risk of severe human rights impacts, and the nature and context of its operations;
c. Should be ongoing, recognizing that the human rights risks may change over time as the business enterprise's operations and operating context evolve.

Guiding Principle 18

In order to gauge human rights risks, business enterprises should identify and assess any actual or potential adverse human rights impacts with which they may be involved either through their own activities or as a result of their business relationships. This process should:

a. Draw on internal and/or independent external human rights expertise;
b. Involve meaningful consultation with potentially affected groups and other relevant stakeholders, as appropriate to the size of the business enterprise and the nature and context of the operation.

Guiding Principle 19

In order to prevent and mitigate adverse human rights impacts, business enterprises should integrate the findings from their impact assessments across relevant internal functions and processes, and take appropriate action.

a. Effective integration requires that:
    i. Responsibility for addressing such impacts is assigned to the appropriate level and function within the business enterprise;
    ii. Internal decision-making, budget allocations and oversight processes enable effective responses to such impacts.
b. Appropriate action will vary according to:
    i. Whether the business enterprise causes or contributes to an adverse impact, or whether it is involved solely because the impact is directly linked to its operations, products or services by a business relationship;
    ii. The extent of its leverage in addressing the adverse impact.

Guiding Principle 20
In order to verify whether adverse human rights impacts are being addressed, business enterprises should track the effectiveness of their response. Tracking should:

a. Be based on appropriate qualitative and quantitative indicators;
b. Draw on feedback from both internal and external sources, including affected stakeholders.

Guiding Principle 21
In order to account for how they address their human rights impacts, business enterprises should be prepared to communicate this externally, particularly when concerns are raised by or on behalf of affected stakeholders. Business enterprises whose operations or operating contexts pose risks of severe human rights impacts should report formally on how they address them. In all instances, communications should:

a. Be of a form and frequency that reflect an enterprise's human rights impacts and that are accessible to its intended audiences;
b. Provide information that is sufficient to evaluate the adequacy of an enterprise's response to the particular human rights impact involved;
c. In turn not pose risks to affected stakeholders, personnel or to legitimate requirements of commercial confidentiality.

Components of AI and the GenAI Value Chain

Artificial Intelligence: Intelligence exhibited by machines including both "machine learning" (an approach to achieve Artificial Intelligence or AI), which uses algorithms to parse data, learn from it, and then make a determination or prediction, and "deep learning" (a technique for implementing machine learning), which is inspired by understanding the biology of our brains (Allison-Hope and Hodge, 2018)

Adversarial Data: One way to improve models' ability to handle adversarial attempts against the model or system is to train the AI model on examples of adversarial data and add corrected labels from the developers' side (INDIAAI).

AI Value Chains: "The organisational process through which an individual AI system is developed and then put into use (or deployed) (Engler and Renda, 2022)

Evaluations: Different kinds of tests for understanding models' performance for specific tasks and capabilities (OpenAI on GitHub). "Evaluation is the process of validating and testing the outputs that your LLM applications are producing" (OpenAI Cookbook).

Fine-Tuning: This is the process of improving an AI model's capabilities at certain kinds of tasks and applications by providing "training data" (OpenAI Platform).

Generative Artificial Intelligence (GenAI): Models built using deep learning that can use the data they have been trained on to produce statistically likely content, possibly in different modes like text and images (IBM Research).

GenAI Systems and Services: These systems represent the model as well as environment and assets that make the model's effective use, implementation and further development possible (Information Technology Industry Council). *For example, Open AI ChatGPT is a system that runs on the GPT-3.5 and GPT-4 models.*

Pre-trained model: The AI model that has been trained on a very large dataset, from sources like the internet, is called “pre-trained.” It is ready to be fine-tuned (Nvidia Blog).

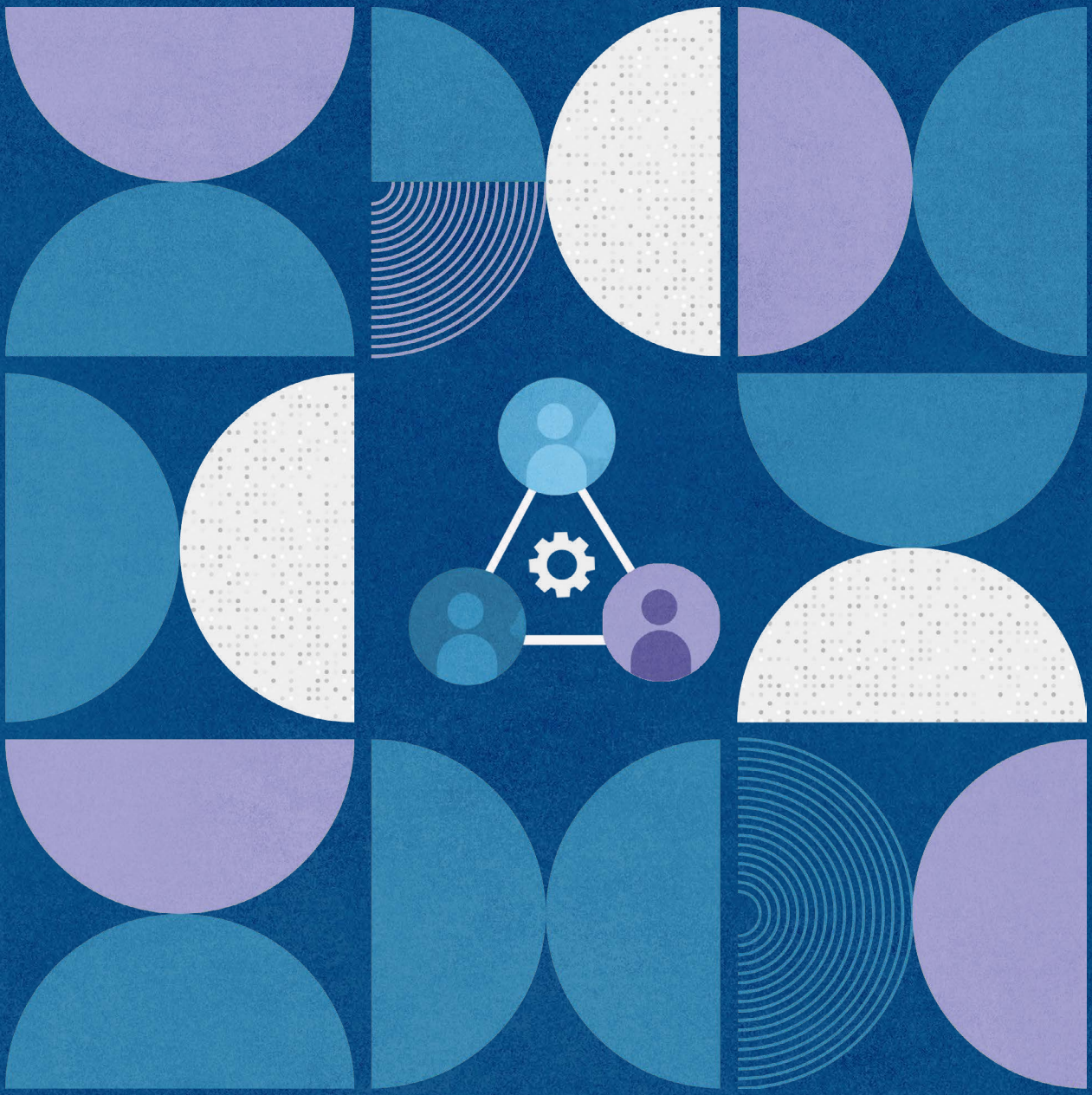
Prompts: The information and instructions that GenAI tools take from users to create outputs for humans’ use (Harvard University Information Technology).

Red Teaming: In the context of AI, red teaming can refer to a form of stress testing to understand a model’s ability and tendency to provide prohibited, dangerous or harmful outputs (Andrew Burt in Harvard Business Review).

Training Data: This is the data that is used to “demonstrate” tasks and performance to a pre-trained model to cultivate its ability with the desired tasks or applications, creating a **trained model** in the process (OpenAI Platform).

APPENDICES

Methods and Engagement



Methods and Engagement

Relevant accounts and stockpiles from our research work on the gender risks within GenAI.

APPENDIX 1 | Methodology

The research has adopted a multi-phase and multi-methodology approach to identify existing and potential harms, and solutions around GenAI systems. The study has employed **problem-centred interviews** with experts, conducted **doctrinal and desk** (non-doctrinal) research, and leveraged life history analyses through a consolidated approach to understand the lived experiences of gender minorities. This dynamic approach ensures that the research insights and findings consider real-time changes in the dialogue at the global level.

Problem-centred expert interview and analysis

This method provides an effective approach to gain specific nuances. Insights using this methodology are solution-oriented in nature, given that they are in control of strategies and are integral to the decision-making process.

Doctrinal analysis and desk research (non-doctrinal research)

a. **Doctrinal analysis:** Doctrinal analysis allowed the research to understand the legislation and regulatory frameworks around AI and human rights. International human rights instruments and national human rights legislation were scanned to ascertain the various human rights at stake due to deployment of AI. Labour codes were analysed for any internal inconsistencies that result in human rights abuses of the labour force.

b. **Non-doctrinal analysis and research:** Non-doctrinal research was also leveraged to understand the implications of the legislation from the perspective of other disciplines, and to integrate those perspectives into the current framework. This method enables holistic exploration of regulatory changes and movements.

Details of the methodologies:

RESEARCH STAGE	AIM AND ACTION	ANALYSIS/OUTCOME
Desk research (Stage 1): • Value chain identification • Human rights risk unpacking	**Aim:** Identification of GenAI value chain stages through exhaustive horizontal scans through various subject matter formats. **Aim:** Identification and long-listing of human rights risks through extensive secondary research sources to create research tools and strengthen the study narrative.	• Detailed research of the GenAI value chain disambiguating the various steps and stakeholders involved in GenAI from development to content dissemination. • Selection of steps of the value chain based on metrics of criticality and severity of impact of human rights risks.
Desk research (Stage 2): • Value chain articulation • Human rights risks mapping **Primary research (Stage 2):** • Expert interviews and Focussed Group Discussions	**Aim** (overall): Identification of (at least) **5** human rights risks and building the architecture of the toolkit. **Aim** (primary research): Review, identify, and analyse secondary data to conduct initial mapping of human rights risks and stakeholders by conducting: • **Interviews** with at least **25** ecosystem experts. • **1 - 2 FGDs** with ecosystem experts (interdisciplinary) consisting of **3 - 5** stakeholders each.	• Identify thematic areas in terms of gendered risks from GenAI in various stages of the value chain. • Note implications of the existing legal framework on human rights. • Collate qualitative primary data from expert interviews. • Carry out thematic mapping of risk areas through qualitative coding of data. • Arrive at key risk areas based on immediacy, adversity, and feasibility for redressal. • Identify key stakeholder groups to engage with while developing the toolkit.

RESEARCH STAGE	AIM AND ACTION	ANALYSIS/OUTCOME
Validations and consultations (Stage 3): • Articulation of key findings from research • Validation of approach and findings	**Aim** (overall): Development of a toolkit and research report for businesses and government as relevant through **2 consultation sessions** with experts. • **2 group consultations** consisting of 4 - 5 stakeholders each. - **Consultation outcomes:** Validation of the framework, key findings, and toolkits.	• Iteratively develop framework(s) and toolkit(s). The toolkit developed will be a guide on how gendered risks associated with GenAI can be mitigated. • Surface emerging solution strategies for each risk.
Report launch and dissemination (Stage 4): • Finalisation of research outputs • Launch of research outputs • Dissemination of research	**Launch event:** In collaboration with UNDP, the research team will host an event to launch the report, share key findings, and embed the research. The launch event is scheduled to be held on the 25th of April 2024 in New Delhi. UNDP is requested to support the event venue and logistics.	• Strengthen ecosystem narratives through strategic dissemination and embedding of research with key stakeholders.

Table 19. Consolidated methodological approach for this study

To engage meaningfully with the ecosystem, the research team spoke to **30 unique and cross-sectional stakeholders** to unpack the research questions and identify critical areas of harmful impact, calibrate the specific risks for focus, and discuss consequent pathways for mitigating policy intervention.

This multi-modal exhaustive approach presented an opportunity to cross-pollinate insights from various critical sectors including academia, business and entrepreneurship, philanthropy, technologists, governance, and civil society to provide a holistic picture of current realities and potential opportunities.

Expert selection rationale:

INTERVIEW LENS	MANNER OF SELECTION
Gender and GenAI	Experts from research, civil society, academia, businesses, and technology with knowledge of the gender and AI sectors
AI Value Chain Development and Training	Experts from research, civil society, businesses, and technology with knowledge of AI value chain development and training
Policy and Regulatory Landscapes for GenAI	Expert interviews of experts from research, civil society, legal policy, and government with knowledge of the policy and regulatory landscape in the AI ecosystem

Table 20. Types of experts selected for primary research interactions
**Interviewees may overlap across sectors and horizontals.*
***Cross-stage interviews may also be conducted based on the expertise of interviewees.*

Limitations and boundaries of the research

This study is limited to understanding the role of gender diversity in GenAI inclusivity, and does not address other relevant cross-cutting themes in this ecosystem at great length. These themes have been highlighted briefly.

Stakeholder identification

The study has identified the following key stakeholders, based on consolidated research and key principles collated from this process. It posits that these stakeholder groups identify the expertise to leverage ecosystem insights and embed the research.

STAKEHOLDER	ROLE	EXAMPLES
State	The State has an inherent interest in protecting the human rights of its citizens who either exist in the market as consumers (demand side) or labour force (supply side).	Ministry of Electronics and Information Technology, Department of Science and Technology, Ministry of Health, state and quasi-state authorities, Ministry of Finance, and Ministry of Labour and Employment (to gain a holistic understanding of the impact of the GenAI value chain on various sectors).

Table 21. Types of stakeholders relevant for this study

STAKEHOLDER	ROLE	EXAMPLES
Businesses	Increased trade and investment are direct benefits for businesses when they adhere to the human rights frameworks.	• Development: OpenAI, Meta, Google, NVIDIA, TMC, CoreWeave, QuantumBlack • Modelling: Sarvam AI, AWS, OpenAI, Oracle, Stability AI • Deployment: OpenAI, Google, Meta, Microsoft, AWS, QuantumBlack
Consumers/Users/ Experts	Usage of AI in businesses directly impacts the human rights of consumers. Various human rights such as the rights to equality, health, and privacy are at stake.	Represented by technology experts, legal experts, civil society, community representatives, and the GenAI labour force. • Philanthropies: Omidyar Network India, Gates Foundation, Nilekani Philanthropies • Research institutions: NITI Aayog, IT for Change, NORAD, Fairwork Foundation, Ada Lovelace Institute, ORF, Oxford Internet Institute • Journalists and practitioners investigating AI: Karen Hao from *The Atlantic,* and Hilke Schellmann, who contributes to the *Wall Street Journal* and *the Guardian*

Table 21. Types of stakeholders relevant for this study

Anticipated outcomes from the study

This research lays the ground for future inquiry on the harms of GenAI for women and gender minorities. The examination has been carried out through a human rights lens, to assess how businesses and the state can play a role in nurturing and developing mitigation measures and better practices in AI spaces and their ranges of use. The successful completion of this study should lead to the following outcomes:

1. An articulation of the evolution pathways for and dynamics of a non-binary gendered lens to GenAI.
2. An articulation of a value chain approach that is conducive to thinking about the gender-related risks and harms of GenAI technologies.
3. Generation of crucial research that identifies and strengthens the narrative for how regulation can contribute to transforming AI across the value chains for gender safety and inclusivity.

APPENDIX 2 | List of Experts Engaged

EXPERT NAME	AFFILIATION	DESIGNATION
Abhishek Singh	National e-Governance Division (NeGD), MeiTY	President and CEO
Aditya Gopalan	Department of Electrical Communication Engineering (ECE), IISc Bangalore	Associate Professor
Aditya Prasad	Department of Computational and Data Sciences, IISc Bangalore	Ph.D. candidate
Ameen Jauhar	Associate Research Fellow	Centre for Responsible AI (CeRAI)
Amrita Sengupta	CIS India	Research and Programme Lead
Antara Vats	International Innovation Corps (IIC)	Project Associate - Data & AI
Arpita Paul	IWWAGE	Senior Research Associate
Blair Attard-Frost	Faculty of Information, University of Toronto	Ph.D. candidate
Dhanya Lakshmi	Peloton Interactive	Machine Learning Infrastructure Engineer
Eunsong Kim	UNESCO	Gender Equality and Social Inclusion Specialist
J.Y.Hoh	BSR	Manager, Technology Sectors
Janaki Srinivasan	IIIT Bangalore	Associate Professor
Kavya Karthik	Ada Lovelace Institute	Visiting Senior Researcher
Malavika Rajkumar	IT for Change	Project Associate - Digital Justice
Nidhi Sudhan	Citizen Digital Foundation	Co-founder
Nitya Kuthiala	Gamoteca	Digital Project Manager

EXPERT NAME	AFFILIATION	DESIGNATION
Sachin Malhan	Agami	Co-founder
Safiya Husain	Karya	Co-founder, Chief Impact Officer
Samone Nigam	BSR	Manager, Technology and Human Rights
Shimona Mohan	Gender & Disarmament and Security & Technology, UN Institute for Disarmament Research	Associate Researcher
Shivaram Kalyanakrishnan	IIT Bombay	Associate Professor
Shweta Gupta	Microsoft	Principal Software Engineer Manager
Siddharth Garg	NYU Tandon School of Engineering	Institute Associate Professor
Smita Gupta	Agami, OpenNyAI	Curator
Sneha Das	Department of Applied Mathematics and Computer Science, Technical University of Denmark	Assistant Professor
Sonakshi Chaudhary	The Quantum Hub	Gender, Strategic Partnerships & Communications
Sonali David	Equilo	Senior GESI Expert
Suhani Pandey	The Quantum Hub	Analyst
Varun Hemachandran	Agami, OpenNyAI	Senior Curator and Lead
Vedika Pareek	The Museum of Imagined Futures	Previously Centre for Internet and Society
Vivek Seshadri	Karya	Co-founder

APPENDIX 3 | The Base Question Bank for Expert Interactions

1. Can you walk us through how you perceive the GenAI/AI value chain?
2. How do we tackle the lack of adequate data collection and labelling in the development of GenAI? How can we look at this problem through an AI development lens?
3. Are there any standards for gender-friendly design and construction of GenAI or similar technologies?
4. What are the sectors or broad use cases for GenAI where gender and gender-related data will have a measurable impact on individuals?
5. How would you identify gender risks in GenAI, in terms of both development and deployment?
6. Do you know about any state, private or CSO initiatives for more gender-inclusive design, preventative measures and management/mitigation of harms?
7. What would you consider to be the failings in data collection and curation that produce gender risks and harms?
8. How do design and modelling practices contribute to GenAI's gender risks and harms?
9. As far as you know, are gender identities beyond cis-men and cis-women being researched or included in design?
10. How would you approach assigning accountability to the players across the value chain? Between the different groups and people developing models, the people picking up models or creating services using models?
11. How can the human factors behind GenAI development and deployment be bettered tobe more gender-friendly? For example, can more diverse hiring of developers and designers be a contributing force to better GenAI?
12. What does the current landscape of GenAI look like in terms of seeking recourse, reparation or change AFTER harms occur? (can be made more sector-specific based on who we speak to)

13. What could be some ways to curtail the generation of gender harms using GenAI? For example, GenAI has been used to create sexual content, often of living people, and circulated.
14. How do you define gender harms? What are the criteria for it being a harm or a risk?
15. How do we map out regulatory frameworks to protect non-binary genders in India when their legal status is still in question?
16. If we were to create a toolkit/model on gender-inclusive AI, what would be the most essential elements?
17. What are some authorities we can look to, to define 'gender' in more broader terms to include the non-binary identities we seek to include?

aapti institute

Aapti is a public research institute that works at the intersection of technology and society. Aapti examines the ways in which people interact and negotiate with technology both offline and online.

contact@aapti.in | www.aapti.in