ESYA centre

A LITERATURE REVIEW ON

GENDER BIAS IN GENERATIVE AI

IMPLICATIONS FOR INDIA AND RECOMMENDATIONS FOR THE WAY FORWARD

APRIL 2024 | ISSUE No. 041

A LITERATURE REVIEW ON

GENDER BIAS IN GENERATIVE AI

IMPLICATIONS FOR INDIA AND RECOMMENDATIONS FOR THE WAY FORWARD

APRIL 2024 I ISSUE No. 041

Attribution: Meghna Bal, Mohit Chawdhry and Noyanika Batta. *A Literature Review on Gender Bias in Generative AI: Implications for India and Recommendations for the Way Forward*. April 2024, Esya Centre.

Esya Centre
B-40 First Floor
Soami Nagar South,
New Delhi - 110017, India

The Esya Centre is a New Delhi based technology policy think tank. The Centre's mission is to generate empirical research and inform thought leadership to catalyse new policy constructs for the future. More details can be found at www.esyacentre.org.

About the Authors: Meghna Bal is Director of the Esya Centre. Mohit Chawdhry is a Fellow at the Esya Centre, and Noyanika Batta is a Junior Fellow at the Esya Centre.

Layout & Design: Khalid Jaleel

Cover Image: Generated by AI using Dream Studio by Stable Diffusion.

CONTENTS

LIST OF TABLES AND FIGURES

EXECUTIVE SUMMARY

Generative AI is known for creating new content and insights, and is transforming human-computer interaction and decision making. This transformation is raising challenges, of bias, and gender bias in particular. Our literature review describes how gender bias emerges throughout the value chain of generative AI, as reported in academic literature and studies by international organizations. While we recognize that bias disproportionately affects women, we also consider the impact of bias on other genders, including men and people that identify outside the gender binary.

The main aim of generative AI as defined in Strobel et al (2024) is to produce new, probabilistic information with varied results derived from identical inputs. This differentiates generative AI from traditional AI, which uses fixed rules to analyze data. Unlike traditional AI, which is used largely for pattern recognition and data classification, generative AI makes use of techniques like deep learning to analyze as well as generate data and ideas, and it excels at creating innovative and contextually relevant patterns.

Scholars highlight a paradox in large generative AI models, sometimes termed foundation models: while scale enhances their capabilities and performance, it also increases the unpredictability of their outputs and functions. This in turn escalates the policy challenge around such models – with gender bias, being a notable problem stemming from this paradox. As such systems come to be increasingly deployed across sectors, it becomes important to understand how such challenges multiply and manifest throughout the value chain.

The value chain of generative AI spans development, deployment, and application, and is vulnerable to bias at every stage. We also note that bias in the upstream stages of the value chain is likely to prompt, reinforce, and amplify bias downstream. The different stages of the generative AI value chain, and the manner in which bias arises in each phase, is summarized below:

Problem framing – Generative AI developers are required to translate high-level strategic goals into objectives and tasks for the algorithm to understand. At the stage of problem framing, there is a risk of oversimplification, and important contexts may be erased, or unaccounted for, leaving room for bias to creep in.

Design – Crucial decisions made at the design stage of development can significantly influence the emergence of gender bias. One such decision is the choice of language used in AI models. For instance, many Indic languages assign a ‘grammatical gender’ to nouns. Such embedded gendering can unintentionally reinforce social biases. This is a gap in current AI research, much of which has focused on ungendered, high-resource languages such as English, and often over-looks the complexities of gender ascription.

Data collection – Biases may arise due to improper sampling or skewed representation, with biased selection or measurement skewing the gender distribution of the sample.

Model selection – The wide use of foundational models for generative AI is another potential entry point. As these models are trained on datasets of information taken from the Internet, the prevailing prejudices and stereotypes may be perpetuated by the model.

Training and labeling – Models trained using unsupervised learning (where datasets are unlabeled) are prone to picking up biases contained in the underlying data, and thereby forming incorrect relations (between social groups, for instance) and producing skewed outputs. Supervised learning, where the datasets are labeled and annotated by humans, may also be vulnerable to bias. Labels may reflect the evaluator's own biases or may insufficiently account for the existence of non-binary genders.

Testing and validation – Before being deployed in the real world, AI systems are subjected to testing and validation using a pre-determined set of parameters. Bias may arise at this stage in two different ways. First, it is challenging to define the testing and validation parameters that will accurately capture gender bias. Second, even if gender bias is evaluated correctly, the existing debiasing techniques are inadequate to ensure gender neutral or equitable outcomes.

Deployment – Once a generative AI model has been deployed in the real world, the context and environment it operates in changes, prompting unforeseen instances of bias. In other words, the latent bias in a dataset or training method may become evident only when end-users use the model, as opposed to developers or deployers. And unlike earlier AI-ML models, generative AI relies on user feedback to improve its outputs. Thus, the bias contained in user inputs and feedback may also be reproduced by the AI model.

Corporate governance and workforce diversity – While not a part of the value chain per se, the corporate environment and team diversity in which AI models are developed can influence the tendency for gender bias in AI systems. Gender homogeneity amongst developers and decision makers can lead them to overlook gender bias, whereas diverse teams have been found more likely to identify and address gender biases in the AI development value chain.

Gender bias in AI systems poses a formidable challenge given its subtle and often unpredictable nature. A promising solution to identify and mitigate gender bias in generative AI is red teaming, which refines AI models by employing adversarial testing techniques. Researchers like Su et al (2023) have proposed red-teaming methods that can generate test cases automatically to reveal bias in large language models – in other words, by suggesting improvements using in-context learning rather than extensive retraining. Red teaming is not foolproof, however, as its effectiveness depends on clearly defined objectives, transparent access for external reviewers, and a combination of oversight mechanisms such as impact assessments and regulations.

Recognizing that it is not possible to eradicate bias, we recommend that policies aim to mitigate it instead. Developers and deployers could consider displaying disclosures/warning labels informing users that their systems may generate biased outputs, enabling users to tailor their interactions with AI to minimize bias, and incorporating user feedback on biased outputs. Governments may also consider working with each other, and with industry and civil society, to establish standardized procedures for red-teaming so as to enhance its consistency and effectiveness. Finally, promoting a diverse workforce and leadership in the AI value chain would help in recognizing and mitigating gender bias concerns, in tandem with policy incentives like those outlined in the US CHIPS and Science Act.

I. INTRODUCTION

Generative AI, known for its ability to generate new content and insights, is reshaping the landscape of human-computer interaction and decision making. Alongside its many benefits, however, generative AI presents unique challenges, particularly in terms of its differentiated impact on people of different genders. Our paper seeks to understand how gender bias manifests across the value chain of generative AI, through a comprehensive review of academic work and reports by international organizations. We analyze manifestations of gender bias across the value chain of generative AI models through a meta-analysis of this literature. While it is well-established that women face a disproportionate amount of bias, we examine the effects of bias on all genders, including men and those outside the gender binary, as bias in AI systems affects everyone.

In India, it is important to understand issues of bias in generative AI, as the technology is being rapidly adopted in the country. Sixty percent of IT professionals surveyed for the IBM Global AI Adoption Index reported active implementation of generative AI tools in their companies, while 34 percent indicated they were exploring AI adoption.[1] More important is the fact that problematic gendered considerations arise in all spheres of life here. The UNDP's 2023 Gender Social Norms Index indicates that 99.2 percent of people in India held at least one biased belief against women.[2] More girls die early than boys in India, though in the rest of the world, female children have a higher rate of survival at birth than boys and are better poised to be on track developmentally.[3] In India, girls are also more likely to drop out of school than boys. These early discrepancies advance into wider chasms as opportunities for higher studies and livelihood come along. Even if a woman enters the workforce in India, she is on average paid 64 percent less than her male counterpart.[4] Such bias can seep into the information generated by the public at large, which often serves as the training data for AI systems. Moreover, when generative AI is deployed in real-world scenarios, the context it is used in can introduce further new biases or exacerbate existing ones.

Workforce gender discrepancies in sectors like STEM result in generative AI research and development being dominated by men, which may preclude the identification and mitigation of bias in the technology. The lack of diversity at the board or governance level of generative AI companies may exacerbate gender bias as well. Women in India comprise only 17 percent of directors in the NIFTY 500 companies, far lower than the global average of 24. Many women directors are also members of the families who own the company, signaling a lack of affirmative gender diversity.[5] And for people outside the gender binary norm, research has shown that while their gender identities have historically existed in India and elsewhere, colonial influences, moral policing, religious and social norms play a key role in denying their acceptance and recognition.[6] To illustrate, most official surveys in India, as in many other nations, adhere to a binary norm of assigning a gender to the respondents, resulting in datasets that conceal non-binary sexes and genders.[7]

This paper studies how gender bias creeps in at different stages in the generative AI value chain and identifies points where interventions can be most effective. It makes recommendations for policymakers, developers, and deployers for mitigating gender bias and reducing the bias-related harms that emanate from generative AI.

II. TECHNOLOGICAL OVERVIEW

The term "generative AI" broadly refers to deep learning models that use raw data to create high-quality text, image, and other content that is a derivative of the data that they were trained on. Once trained, the model can typically create new works that are similar while not identical to the original data. For instance, a dataset containing images of cars can be used to build a model that can generate new images of cars that have never existed but still look real. This is because the AI model learnt rules that govern the appearance of a car. Generative models have long been used in statistics to analyse numerical data.[8] The earliest generative models, such as the Hidden Markov Models and the Gaussian Mixture, were devised in the 1950s primarily to produce data sequences such as speech.[9] The advent of deep learning significantly broadened the use of generative models, enabling them to create realistic images, videos, and write engaging text.[10]

According to Strobel et al (2024) there are four kinds of deep generative models:[11]

Variational Auto Encoders

Variational Auto Encoders (VAE) are neural network autoencoders comprised of two neural networks: an encoder and a decoder. The encoder compresses the input data into a smaller, dense representation (the latent code) while the decoder conducts the inverse operation, translating the compressed data back into the original form.[12] After repeatedly undergoing this training process, the encoder 'learns' an optimized latent representation that can capture the fundamental characteristics of the data, enabling precise reconstruction.[13]

VAEs do not simply reconstruct data: they churn out variations of the original.[14] To do so, they apply regularization to the latent code – that is, they add rules to the mix that will prevent the decoder from replicating the input data perfectly.[15] Instead, they develop a general understanding of what the data looks like, enabling them to generate diverse data samples.[16] Essentially, VAEs take the original data and turn it into a map of possibilities. Rather than offering a single point to describe each characteristic of the input data, they provide a whole range of possibilities. The ability to learn the underlying probability distribution of a dataset and generate new data samples that resemble the training data is what makes VAEs so useful for generative modelling.[17] Common applications of VAEs include generating synthetic data, realistic images and speech, and detecting anomalies.[18]

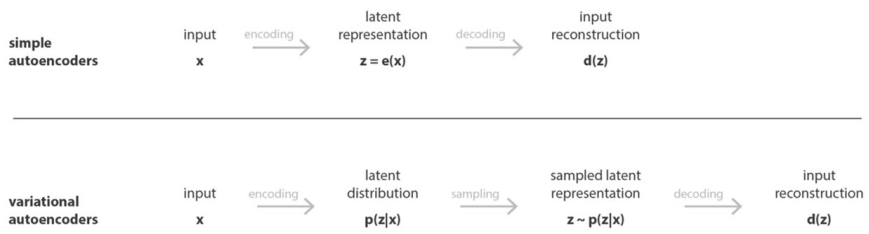

Figure 1: Difference between autoencoders (deterministic) and variational autoencoders (probabilistic) (Source: Towards Data Science)

VAEs' ability to generate novel data ignited a rapidfire succession of new technologies, from generative adversarial networks (GANs) to diffusion models, which are capable of producing even more realistic images. VAEs thus set the stage for today's generative AI.[19]

Generative Adversarial Networks (GANs)

GANs are an interplay between two competing neural networks (the generator and the discriminator) that are trained simultaneously.[20] The generator might be thought of as a counterfeiter, while the discriminator plays the role of the police.[21] The generator's aim is to learn the statistical distribution of real data, so as to generate fake data that closely resembles the original.[22] In training, it needs to deceive the discriminator into accepting the generated data as real or original. Conversely, the discriminator acts as a classifier distinguishing the generator's fake data from real-world data. This feedback loop helps refine the generator's capabilities. As the two networks continue training, both grow stronger, and learn to generate data as close to the original as possible.[23] GANs are widely used in text-to-image synthesis, image-to-image translation, and have many potential medical applications, in medical image analysis, classification and segmentation tasks, to help detect and diagnose disorders and disease.[24]

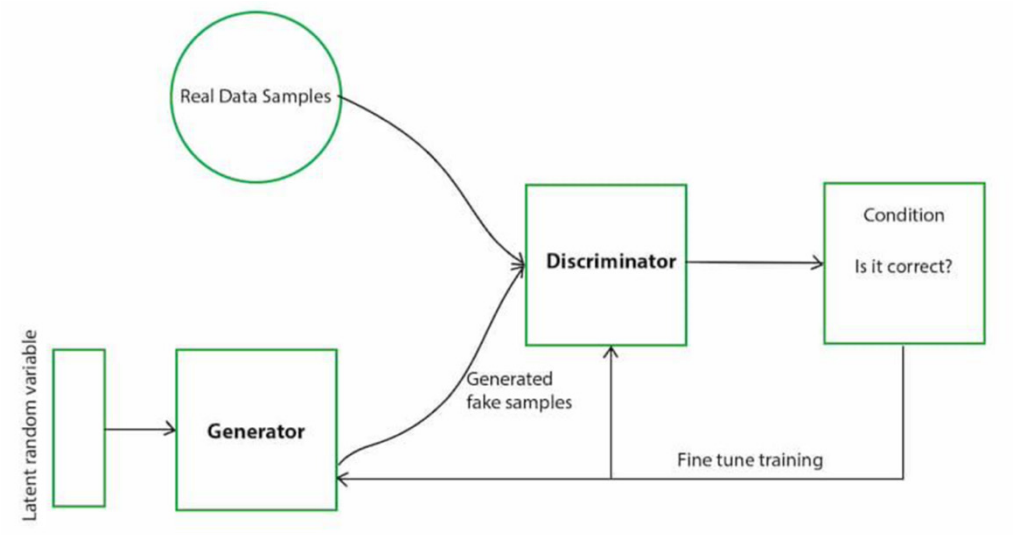

Figure 2: How a GAN works

(Source: Geeksforgeeks)

Transformer Architecture

Transformers are neural networks that track relationships between chunks of data to derive 'meaning' from them.[25] Unlike traditional models that process words step by step,[i] transformers concurrently process all parts of a sentence,[ii] making them efficient and GPU-friendly. Introduced by Google in 2017 in a landmark paper called 'Attention is All You Need', transformer models employ a mathematical technique known as attention or self-attention,[26] which lets the model evaluate how distant elements within the given data influence and depend on one another, over large spans of text.[27]

Earlier deep learning techniques, including recurrent neural networks (RNNs) and long short-term memory networks (LSTNs), would process each word separately, an ineffective method when there is a large gap between the relevant information and the point where it is needed.[28] The method was ineffective because the information had to pass through each step and the longer the chain, the higher the likelihood of losing relevant information along the way. To address this, researchers developed attention mechanisms to focus on specific words, as any word in a sentence may contain pertinent information. The attention mechanism ensures precise decoding by evaluating every word in the input data.[29] Transformers employ self-attention to weigh the significance of different parts of an input sentence. The mechanism enables the model to capture long-term dependencies and intricate relationships in the data, overcoming the limitations of earlier architectures, which often struggled with longer sentences.[30] With their parallel processing capabilities and attention mechanisms, transformers have mitigated these challenges.

Another reason for the popularity of transformers is the growing availability of computational power as well as large datasets. Prior to the advent of transformers, neural network training required large, labeled datasets that were expensive and time consuming to build.[31] By mathematically identifying the patterns between elements, transformers eliminate this requirement, letting researchers train increasingly large models without having to pre-label the data.[32] As a result, new models can now be trained on billions of pages of text, and generate answers perceived as having more depth.

i. Traditional models like RNN process words one by one, considering each word's context in turn. Like in the sentence, 'I love ice cream', a traditional model begins by processing the word 'I', then takes the context into account when processing the word 'love', and so on. It sustains a hidden state that carries information over from the previous words, allowing it to interpret the sentence's overall meaning. This severely limits the amount of information such a model can remember at once.

ii. Transformers don't rely on sequential processing, allowing them to do computation in parallel and process sentences more efficiently. For example, in the sentence, 'The cat sat on the mat', a transformer would break it down into smaller units called tokens ('The', 'cat', 'sat', 'on', 'the', 'mat' and punctuation marks). Each token is represented as a vector that encodes its meaning and context. The transformer then learns how these tokens relate to each other in order to 'understand' the sentence. The window of information they can process is virtually unlimited in transformers as they can access information from any element of the input sentence.

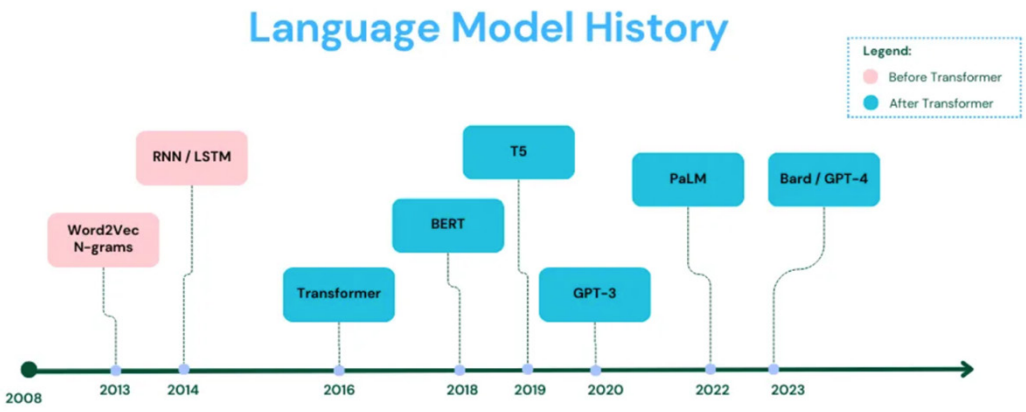

Figure 3: Timeline of language model development

(Source: Medium)

OpenAI's popular ChatGPT text generation tool makes use of transformers for prediction, summarizing, replying to questions, etc. The GPT in the tool stands for generative pre-trained transformer. Because transformers can translate text and speech in near real time, they have quickly become fundamental to natural language processing.[33] Transformers have also become a cornerstone of complex language modelling tasks because they are effective at understanding and representing contextual relationships within data.

Latent Diffusion Models

Latent Diffusion Models (LDMs) are deep learning models with powerful high-resolution image generation and manipulation capabilities.[34] Probabilistic in nature, they excel at generating high-quality images by starting with random noise and gradually transforming it into realistic images, through a process of diffusion.[35] The process is split into forward and reverse diffusion. Forward diffusion is the process of turning an image into noise (Figure 4 below), while reverse diffusion turns that noise back into the earlier image. The essential idea of diffusion is to slowly destroy the structure found in a data distribution through forward diffusion and then to restore the structure in the data using reverse diffusion.[36]

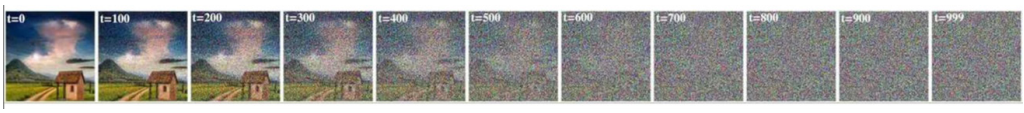

Figure 4: A visual representation of the forward diffusion process

(Source: erdem.pl)

A distinctive feature of LDMs is that they apply the diffusion process not directly to the raw pixels of an image, but instead to a compressed image representation.[37] The compressed representation captures the most important features and semantics of the image in condensed form.[38] The diffusion process is then applied to the latent code rather than the pixels themselves. This lets the model manipulate the image in a more controlled manner, by modifying the latent code alone.[39] Once the diffusion process has altered the latent code to generate the desired output image, a decoder converts the latent code back into the pixel space to reconstruct the final high-resolution image.[40] By operating in this compressed latent space instead of directly on pixels, LDMs achieve enhanced computational efficiency without losing quality. Diffusion models are the current go to for image generation, and are the foundational model for popular image generators such as Dall-E, Stable Diffusion, Midjourney, and Imagen.[41]

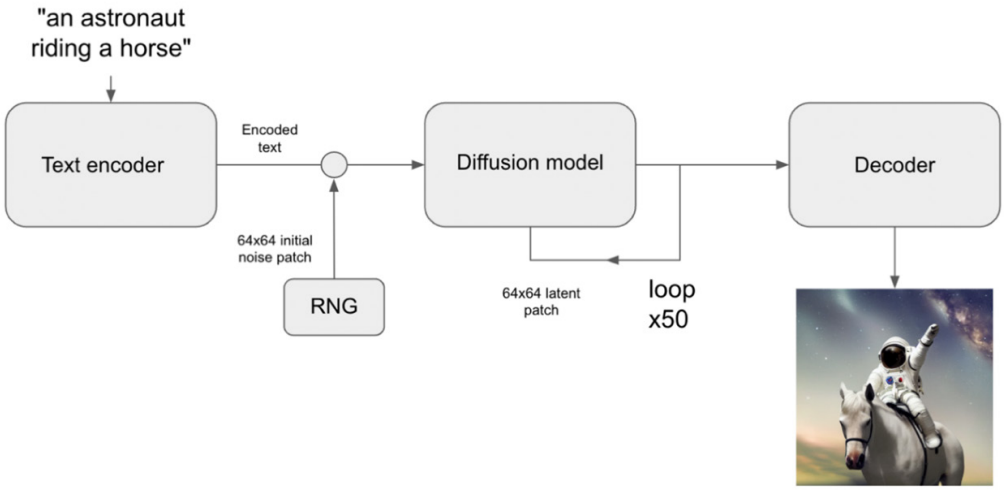

Figure 5: Architecture of latent diffusion models

(Source: Keras)

In this paper, we rely on Strobel et al's conceptualisation of the main objective of generative AI: creating new, probabilistic information with varied outputs based on the same input. This capability is what distinguishes generative AI from traditional AI, which largely analyzes data using pre-existing rules and information.[42] Unlike traditional AI, which excels at pattern recognition and data classification, generative AI uses advanced machine learning techniques, such as deep learning, to not only analyze but also produce data and ideas. These techniques enable generative AI to excel at pattern creation, producing content that is both creative and contextually relevant.[43]

Ganguli et al (2022) note that the large generative models referred to by Bommasani et al (2021) as foundation models[44] present a paradox. On the one hand, their capabilities and performance improve with a higher expenditure on development.[45] On the other hand, as the scale of development increases, so does the unpredictability of the models' capabilities, outputs, and inputs. This in turn presents a policy challenge around the deployment of such models. The gender and other biases they display are one such outcome of the paradox outlined above.[46] And with the increasing pervasiveness of these systems across spheres of commerce and society, it is important to understand how biases can arise and proliferate along different parts of their value chain, and what strategies can be deployed to address them.

III. UNDERSTANDING BIAS IN THE AI VALUE CHAIN

The generative AI value chain is comprised of the steps involved in developing, deploying and using the AI models described in Section II. While the terminologies may vary, the generative AI value chain typically involves the stages shown below.[47] Research suggests that bias can creep in at any of these stages. [48]

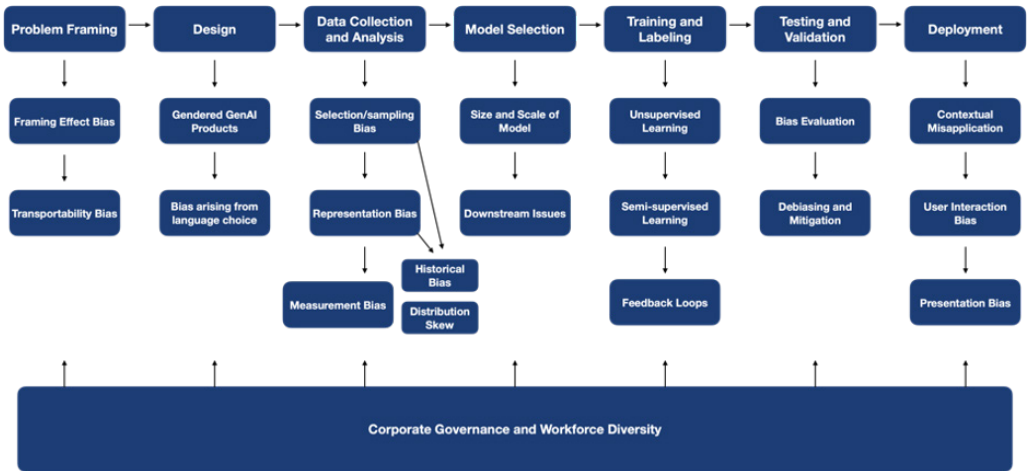

Figure 6: Diagram of bias across the AI value chain

(Source: Author compilation)

Stage 1: Problem Framing

Problem framing involves clearly defining the problems sought to be addressed by the AI model and identifying priority use cases and applications for it.[49] The problem-framing stage also establishes the scope and direction of later stages of model development. For example, certain decisions on data collection may be taken here, and the parameters to be used to assess a model's success or failure are established.[50] Failing to account for bias at this stage may amplify related challenges down the value chain.[51] Typically, the bias that most commonly arises at the problem framing stage is the:

***Framing effect bias**:* These biases arise from the way a problem has been formulated or framed.[52] Formulating problems in generative AI entails the translating of high-level strategic goals into objectives and tasks easily understood and performed by the algorithms. This in turn involves identifying the appropriate variables and proxies, and this can be a challenge to get right.[53] For instance, the problem may come to be oversimplified to match model capabilities, by ignoring cultural and historical contexts (Srinivasan and Uchino 2021). Oversimplification such as this may yield biases

relating to gender. For instance, Srinivasan and Uchino (2021) showed how framing-effect biases relating to style transfer in artworks result in gender bias when the cultural context of a particular artwork isn't adequately accounted for. They examined the performance of the Abacus.AI online tool, which converts user-uploaded images of people into someone of a different gender. They test the tool using paintings of young men made by Renaissance artists, including Raphael and Piero di Cosmo.[54] The tool mistakenly identified the young men as women, and created masculine versions of them by adding beards. The misclassification was attributed to the tool's failure to account for the artworks' cultural context, in which it was a norm for young men to have long hair.[55] The oversight led to a '**transportability bias**,' whereby the tool stereotypically associated men with short hair and failed to recognize the spatial and temporal variations in the ways we present or perform gender.

Stage 2: Design

At the design stage, decisions are made that can be the basis for all kinds of bias to creep in further downstream. Developers make choices here that can have a wider social impact, as shown in the example of largely 'female' chatbots below. However, the decisions may be a response to market demand. For instance, Borau et al (2021) find that female gendering increases the perception of chatbots as being human and improves their acceptance. In such instances, where does the responsibility lie for resolving biases of perception? Examples of bias at the design stage include:

***Gendered generative AI products**:* Businesses often decide to use female identifiers for chatbots, which may perpetuate stereotypes associating women with subservient or service-oriented roles. For example, Feine et al (2020) find that the names, avatars and descriptions of over 75 percent of 1,375 chatbots could be classed as female.[56] Specifically, 76.9 percent of chatbots with gender-identifiable names were found to have female names, 77.6 percent of avatars were classified as female, and 67.4 percent of chatbot descriptions used female pronouns to refer to the bots.[57] Data also indicate that a large proportion of chatbots are implicitly designed to appear female.[58] West et al (2019) study popular AI-based voice assistants like Siri and Cortana and find deep embeddings of characteristics traditionally associated with effeminate genders, such as submissiveness.[59]

***Bias arising from choice of language**:* Another issue can emerge at the design stage in the choice of language used. Many languages are grammatically gendered. For instance, languages such as Hindi and Marathi ascribe a gender to every noun, including nouns used for inanimate objects.[60] Such associations will affect the pronouns, participles and adjectives used in sentence construction. Evaluating gender stereotypes and bias in gendered languages is quite different from ungendered languages like English. Yet it is ungendered languages that are the focus of research evaluating gender bias in order to mitigate it. Indeed, Stanczak and Augenstein's (2021) survey of 204 papers on bias in natural language processing systems highlights that most of this research was conducted in monolingual contexts using English or other high-resource languages.[61]

Failing to account for the gendered nature of many languages can lead to bias in

problem formulation. Gupta et al (2021) show that asserting a strict binary of grammatical gender, or linking pronouns exclusively with gender, in Hindi–English machine translation can dismiss or make invisible those who do not identify with either of those genders.[62]

Stage 3: Data Collection and Analysis

The problem formulation or pre-design stage is normally followed by collecting and processing the data needed to train the AI model. Generative AI models rely on data to understand patterns and create text, image, audio and video outputs. The larger generative AI models are trained on vast amounts of data – for instance, OpenAI's ChatGPT is trained on numerous datasets, including the Adversarial Natural Language Inference Corpus, Quora question pairs, and Common Crawl.63 When such datasets contain biased sources or when the data are incomplete or missing critical information, the outputs of the AI model may be biased as well.[64]

Various types of bias can therefore enter the AI pipeline at the stage of data creation or collection. Broadly, the biases arising at the data collection stage can be classed as follows:

Selection/Sampling/Self-Selection bias:[65] Selection bias occurs when there is insufficient randomization in choosing individuals, groups or data tuples for analysis. Sampling bias arises from non-random sampling of a population, where certain subpopulations become likely to be included. Selection bias is a precursor to sampling bias, as it may result in samples that do not represent a random selection of the population or subpopulations. Self-selection bias occurs when only a subset of the target population chooses to participate in an experiment, creating inaccurate or skewed conditions because of the respondents' decision (not) to participate in the research.

Although selection, sampling and self-selection bias are sometimes used interchangeably, it is important to distinguish between the three. Imagine a researcher conducting a survey in Andhra Pradesh by mailing questionnaires to selected respondents. If the respondents are only chosen from certain areas of the state, the failure to ensure a random representation of the state's diverse populations is an example of selection bias. Independently, if the collected samples are not representative of the population of Andhra Pradesh, this would be an instance of sampling bias. Finally, if there is no bias in selecting the respondents, yet only a small portion decide to respond, it may lead to self-selection bias. An example of this is a survey asking, 'Do you like responding to surveys?' with the options 'Yes, I love responding to them' and 'No, I toss them in the trash.' If only 10 percent of respondents participate and 99 percent choose the first option, the results are likely invalid, as the 90 percent who chose not to respond are more likely to have chosen the second option.

Selection bias can translate into gender bias when the process of selecting data is influenced by gender considerations, leading to unrepresentative or skewed samples that do not accurately reflect the gender distribution of the population. Where data is being collected afresh, the personal gender biases, ignorance or prejudices of the data collectors and analysts may also carry over into the data.[66]

***Representation bias**:* Representation bias occurs when the training data inadequately represent and consequently struggle to effectively generalize certain segments of the target population.[67] Shahbazi et al (2023) argue that representation bias is almost always ensured where a 'systematic approach to data collection' is not followed.[68] Representation bias may also arise due to other biases informing the data collection process. These include:[69]

- **Historical bias**: These are preexisting biases resulting from socio-technical challenges in the world. An example of historical bias is evident in Google's image search results. When searching for the term 'CEO United States,' the results predominantly feature images of male CEOs, with fewer images of female CEOs. This is a reflection of the fact that only 8.1% of Fortune 500 CEOs are women, leading to a bias in the search results to favour male CEOs.[70]

- **Underlying distribution skew**: The foundational distribution from which data are gathered may not contain an equal or equitable ratio or adequate representation of all subgroups. In such an instance the base distribution is naturally skewed, with no intentional discrimination behind it. For instance, the use of data from official surveys in India, which adopt a binary gender classification, may yield training datasets that do not account for people of non-binary genders or sexes.[71]

  Representation bias can also arise because of sampling, selection or self-selection biases.

Measurement bias: Measurement bias arises due to human errors in capturing data, or defects in the devices used to capture data. The errors or defects can affect the completeness or accuracy of the data, leading to biased outputs.[72] In the context of generative AI, measurement bias can be significant when digitally capturing the artworks used to train models. Srinivasan and Uchino (2020) note that camera characteristics, such as lighting and angle of capture, can affect the transportability of the captured artwork, resulting in bias.[73] However, the connection between measurement and gender bias is at present unclear.

Stage 4: Model Selection

While developers may choose to create their own model from scratch, many existing generative AI applications are based on one of several foundation models.[74] Foundation models are trained on broad datasets and then adapted or fine-tuned for use in a wide range of downstream tasks.[75] Some notable examples of foundational models currently in use include Anthropic's Claude 2, Meta's Llama 2, and OpenAI's GPT-4. Using such pre-trained models to develop generative AI applications helps lower costs and improve efficiency, making them a popular choice among developers.

The growing trend of using foundation models to develop generative AI applications can also introduce bias into the value chain. To reiterate, these models are typically

trained on vast datasets that include information sourced from the Internet. These datasets may exhibit gender stereotypes and prejudices that are replicated by the model. Bender et al (2021) highlight that creating larger language models can increase the risk of disproportionately emphasizing dominant viewpoints and embedding biases harmful to marginalized groups, including women and people outside the gender binary norm.[76] They observe that such large language models can display prejudice in subtle ways, such as implying that doctors are always non-female, or in more obvious ways, like by recognizing only two sexes or genders (male and female) thereby discriminating against those outside the binary.[77]

Another problem with the growing reliance on foundation models is that while the original developers may be aware of the model's problems and limitations, developers and deployers downstream may not. The loss of context can lead to the replication of bias when the model is reused for a different downstream task or application.[78]

Stage 5: Training and Labeling

At the training and labeling stage, generative AI models are trained to recognize patterns in the data collected and to use their understanding of these patterns to output new, similar kinds of data:

Unsupervised learning: Unsupervised AI models work independently to understand the data's inherent structure and derive outputs without specific guidance or labeling from humans.[79] Such models are typically 'pre-trained' on large existing datasets before being fine-tuned for specific downstream applications. This helps developers save time and cost in creating cleaned and labeled datasets or text corpora, a reason for the popularity of unsupervised learning.[80]

Biases can arise in several ways in the course of unsupervised learning. Among the most prominent is the amplification of biases already present in the underlying dataset. For instance, a learning model that relies on incomplete or biased data will likely reflect these biases in its output as well. Bolukbasi et al (2016) show that 'word embeddings' derived from seemingly neutral datasets like Google News articles exhibit gender stereotypes to a large extent.[81] Word embeddings are mathematical representations of language that encode the semantic relationships between words.[82] As they are derived from vast text corpora, they may reflect and perpetuate existing gender, racial or ethnic biases. Bolukbasi et al (2016) found that when queried with the prompt, *'man is to computer programmer as woman is to x',* the algorithm replied with 'homemaker', indicating that the word embeddings for man and woman are deeply biased, although the underlying data primarily contained articles from professional journalists and authors.[83]

Another common kind of bias that may arise in unsupervised learning is *confounding bias*, which occurs when a model learns incorrect relations between different groups, or does not account for all the relevant factors in determining patterns.[84] Steed and Caliskan (2021) find that state-of-the-art unsupervised generative models trained using ImageNet, a dataset curated from internet images, automatically acquire gender biases.[85] They tasked the models with auto-completing cropped images of male and

female faces, and found that 52.5 percent of the completions featured a bikini or low-cut top for female faces, while only 7.5 percent of male faces wore low-cut tops or were shirtless.[86] Further, 42.5 percent of male completions wore a suit or other career-specific attire, reinforcing harmful stereotypes about the professional involvement of both women and men.[87]

***Semi-supervised learning**:* Supervised AI models are trained on labeled or annotated data to give them the necessary contextual information and guide them toward the output desired.[88] Semi-supervised learning is a combination of supervised and unsupervised learning that involves training models on both labeled and unlabeled data.[89] The main motivation behind semi-supervised learning is to ensure consistent and accurate outputs even with slight changes in the input, by training the model on both labeled and unlabeled data.[90] Semi-supervised learning is particularly useful in generative AI for it allows developers to combine a small set of labeled data with a large corpus of unlabeled data to create realistic and accurate outputs, such as images.[91]

Because semi-supervised learning makes use of unlabeled data, the bias-related concerns in the context of unsupervised learning are relevant here as well. Additionally, a distinct set of gender bias concerns arises due to the use of labeled or annotated data to train the AI models. First, the data labeling process may reflect the gender biases and arbitrary preferences of the evaluator, especially when the data being labeled is ambiguous or highly subjective.[92] For instance, Schwemmer et al (2020) analyzed images of powerful male and female politicians and found that images of women received three times the number of annotations about their physical appearance as men.[93] The disproportionate focus on the physical appearance of women in labeling practices may be encoded in the AI model's training dataset. Consequently, a model trained on this dataset may learn and perpetuate the bias by focusing unduly on physical appearance when tasked with generating content about female politicians.

Second, labeling practices may only account for the two conventional genders, leading to misrepresentation or the exclusion of people of other genders. Jaiswal et al (2023) find that gender analyzers, which predict a person's gender based on the text of a post or comment on sites like Reddit and Tumblr, misattribute content posted by non-binary people to females in over 50 percent of cases.[94] Finally, *confirmation bias,* which describes the tendency to search for, interpret and validate information in a way that confirms our existing beliefs, may also prejudice the process of labeling.[95]

Feedback loops: Taori and Hashimoto (2023) find that relying on model outputs as a source of training for an AI model may amplify biases in the underlying dataset.[96]

Stage 6: Testing and Validation

During testing and validation, a model's performance is assessed against a set of metrics chosen to evaluate the quality and accuracy of the output generated. For instance, an 'inception score' can be a consistent and objective measure of the quality of a generated image using a probability distribution.[97]

Testing and validation offer the opportunity to identify bias and mitigate it in generative

AI. But it is not straightforward to test for bias and mitigate it, for at least two reasons. First, evaluating bias in generative AI models can be a challenge as it involves turning concepts of bias into variables that can be measured. Vyas et al (2021) evaluate 146 'fairness metrics' for algorithms and discover significant variations in the way they quantify a model's behavior towards different social groups.[98] Such differences in metric parametrization may also lead to different interpretations of bias. Metrics meant to capture group fairness can compare false positive rates or true positive rates across groups. Metrics focused on counterfactual fairness can assess how changing a single identity term in a sentence (like changing a name from typically male to female) affects the output.[99] Similarly, Goldfarb-Tarrant et al (2021) argue that the metrics used to evaluate intrinsic bias, such as the biases present in word embeddings, do not correlate with evident bias in downstream applications.[100]

Second, even if the bias in a generative AI model is evaluated correctly, there are shortcomings in existing debiasing techniques. Gonen and Goldberg (2019) find that the methods used to lower gender bias in word embeddings only reduce it superficially, with the counterproductive effect of hiding bias rather than preventing it.[101] Word embeddings are a technique used to convert words to numbers so that computers can understand them. Each word is represented as a vector (a correlated list of numbers) that can capture the semantic sense of the word. Words similar in meaning are those with vectors close to each other in the vector space. The technique makes it easier for algorithms to process language and perform tasks like translation, search and text analysis, simply by comparing and computing vectors.[102] In our example, Gonen and Goldberg (2019) conclude that the current debiasing techniques are inadequate and cannot be trusted to yield gender-neutral modeling.[103] The debiasing techniques used in Google's Gemini large language model are known to result in inaccurate depictions of historical events and personalities in pictures generated by AI. For example, prompted to generate images of the 'founding fathers' of the United States, a group of White men, Gemini created an image showing a group comprised of Black and White individuals. Similarly, when prompted to generate images of German soldiers in 1943, it created images depicting women, who were not active combatants during the War.[104]

Stage 7: Deployment and Application

When generative AI models are deployed, the context of their application and interaction with users can also introduce bias into the value chain.[105] It is common for models to perform well during validation and testing, only to encounter difficulties when deployed in the real world.[106]

When an AI model is deployed in the real world, the context it is used in can introduce new biases or exacerbate existing ones. Sogancioglu and Kaya (2022) analyze gender bias in four different pre-trained word embeddings in the context of depression as a mental disorder. They found that the type of word embedding used influenced the direction of bias relating to depression towards different gender groups. In other words, the way that depression was associated with gender would vary based on the embedding used. Biases in these embeddings were found to transfer to downstream tasks, specifically in depression phenotype recognition.[107]

Biases prevalent in the consumers of generative AI products may also drive some less-than-desirable gendered outcomes, such as the preponderance of female-linked voice assistants. For instance, Mahmood and Huang (2023) find that men interact more sociably with feminine voiced assistants than masculine ones, and that apologetic or submissive voice assistants tend to be perceived as warm.[108]

Importantly, generative AI is different from earlier forms of AI-ML because user interaction plays a significant role in generating model outputs and reinforcing learning via feedback loops.[109] Generative AI chatbots like ChatGPT rely on user prompts to create multi-modal outputs such as text and images, where users can grade the chatbot's response, sharing feedback on which responses were accurate or desirable and which were not.[110] A large volume of user content, including prompts and feedback on outputs, is used to make these AI models more accurate and improve their capacities.[111] Such information exchange between users and AI models can give rise to bias in different ways. For instance, user prompts and inputs may contain gender bias that is replicated by the model.[112] Users may also provide positive feedback to biased outputs, reinforcing the existing biases in the model.[113]

Lastly, generative AI applications (especially chatbots) are also susceptible to bias in presentation and ranking. When tasked with summarizing and presenting information to users, chatbots need to determine which information is most relevant and to deliver it succinctly. In the process a vast amount of information remains unseen, which can lead to a distorted perception of what information is available or important.[114] Similarly, chatbots need to rank information on relevance, and this can cause ranking bias, or the belief that the highest-ranked results are the most relevant, rewarded with more user clicks. Ranking bias can significantly alter user interaction and perception of information.[115]

Evidence also suggests that derogatory content created by generative AI tends to target people of certain genders more frequently and consistently. For instance, one report found that 96 percent of deepfakes were non-consensual sexual depictions, and of them, 99 percent were created featuring women.[116]

Corporate Governance and Workforce Diversity

Though corporate governance and workforce diversity are not a part of the value chain for generative AI per se, they play an important role in shaping its output. Put simply, generative AI is a product of the environment it is built in. A lack of adequate representation in the workforce or boards of the organizations developing such models can lead to biases in them. West (2019) notes that gender bias in AI is not only a technical issue, it is deeply rooted in social and organizational structures, including the composition of the teams responsible for AI development.[117] They argue that the under-representation of women, trans people, and those outside the binary gender norm in the AI workforce is linked intrinsically to the degree of gender bias these models evince – because bias will more likely go overlooked or unrecognized when teams lack gender diversity.[118] Wajcman et al (2020) observe that AI models, like other critical technologies, are normally created for the many by the few. This poses a risk that the models will consolidate the existing power dynamics, creating negative

feedback loops that reinforce bias.[119] Such feedback loops are especially common in the AI sector, where women constitute just 22 percent of the workforce worldwide.[120] In India, women comprise just 14 percent of the workforce in STEM.[121]

There is also evidence of the positive impact of a representative workforce in addressing gender bias in AI. Quiros et al. (2018) note that more gender-diverse teams not only help identify and prevent gender bias but also improve performance and innovation, particularly in so-called knowledge-based industries, like computer science, as they are more likely to understand the needs and desires of a wide consumer base.[122] Diverse teams bring a variety of perspectives to the table that can yield more comprehensive and creative solutions.[123] Meanwhile, Verma et al (2020) show that individual developers who identify with the same gender are more likely to corroborate each other's biases and errors.[124] A demographically diverse team is less likely to 'double down' on such bias, improving team performance and outcomes.

IV. RECOMMENDATIONS

Gender bias in AI systems is a challenge to overcome as it can creep in even when there is adequate accounting for gender related concerns. It can be unforeseen or unforeseeable, exacerbating the difficulty of resolution.

A solution gaining ground for overcoming gender bias in generative AI is known as red-teaming. It entails using manual or automated technologies to adversarially test a language model for harmful outputs, then updating the model to prevent them. Red teaming techniques proposed by researchers have had some success in mitigating gender bias in large language models. For instance, Su et al (2023) propose a novel method to automatically generate test cases that can detect potential gender bias in large language models.[125] Their method was used to test three well-known LLMs, and the generated test cases were effective at revealing the presence of biases. To counteract the biases they identified, they suggest a mitigation strategy that uses the generated test cases as demonstrations for in-context learning, eliminating the need to fine-tune the model parameters. Their experimental results show that LLMs can produce more equitable responses using this approach.

Red teaming is not a silver bullet, however. Friedler et al (2023) outline the situations where red-teaming is effective and where it is not (summarised in Table 1 below).[126]

WHEN RED TEAMING IS EFFECTIVE	WHEN RED TEAMING IS NOT EFFECTIVE
The objectives and flaws targeted by the exercise are clearly defined. It is more successful when the criteria for success are explicit, so that all parties can recognize when the red-team uncovers new ways to compromise a system. For instance, clear outcomes might include unauthorized access to private data like credit card numbers or bypassing safeguards such as content filters.	**The desired outcomes or system behaviors are complex or disputed.** When the goals of an assessment or the system's actions are more nuanced than a simple 'yes' or 'no', it can be a challenge to evaluate the results of red-teaming. For example, assessing a system for 'fairness' without a clear and accepted definition of fairness may lead to disputes over whether an outcome is genuinely 'fair.'
It is combined with openness, disclosure, and access for external entities. Red-teaming can help external groups and the public understand, evaluate, and trust a system's testing. For external red-teaming to work, these groups need complete and transparent access to the system. Disclosing findings can also build trust and can help others learn from the identified problems.	**It is used as a stamp of approval.** Red-teaming can only evaluate a system based on the specific tests conducted: it cannot ensure that all interactions with the system will be 'safe' or 'fair.' Moreover, as red-teaming is of limited effectiveness in isolation, it should be combined with accountability measures like impact assessments, participatory governance, and government regulation.

WHEN RED TEAMING IS EFFECTIVE	WHEN RED TEAMING IS NOT EFFECTIVE
It is part of a comprehensive evaluation framework. Red-teaming is most effective when used together with other methods as it only evaluates specific safety indicators. When conducted in an inclusive process open to external participants, it can help uncover unforeseen failures or 'unknown unknowns.'	**The process and system are not open to external scrutiny.** Red-teaming conducted in a closed manner, such as by a company's internal engineers, misses the chance to foster public trust through transparency. It also demands extra effort to ensure the red team includes individuals distant enough from the system's development to identify unexpected failure modes or 'unknown unknowns.'
Stakeholders are committed to addressing the findings. When red-teaming exposes vulnerabilities, there should be a plan and commitment to mitigate the concerns. If the system is already in operation, there should be redressal mechanisms for those affected by the problems.	**There is a lack of resources, commitment, or plans to address the findings**. Identifying vulnerabilities is pointless without a plan or resources to tackle the identified concerns. It is crucial for organizations to empower individuals to take meaningful action by implementing appropriate mitigation measures.

Table 1: Factors determining the effectiveness of red-teaming AI models

(Source: Summarized from Friedler et al)

Given the limited potential of red-teaming to resolve bias in generative AI, we recommend the following:

Policy strategies to address bias must focus on mitigation rather than complete removal. Fairness evaluations and metrics can be expensive to implement, and may be inaccessible to most deployers and developers of generative AI. Businesses should attempt to resolve bias concerns within the generative AI value chain on a best-efforts basis. In the context of generative AI, specific stipulations for removing bias should be eschewed, as they may create a compliance-related entry barrier for startups. Implementing comprehensive bias removal typically increases the cost of computation and training and may require additional data collection or annotation,[127] and for smaller entities or startups, the costs associated with removal can be prohibitively high, deterring newer entrants and stifling innovation in the AI sector. Any stipulations for mitigating gender bias should be limited to entities that have reached a certain scale of business. Any such stipulations must be decided in concert with industry and remain mindful of the limitations of debiasing techniques.

Deploy disclosures/warning labels. While the capacity to remove bias may be limited, it may be useful for companies to warn users that their systems may yield outputs that reflect gender bias.

Empower consumers to mold their interactions with generative AI to mitigate bias. Besides business disclosures, consumers can be empowered to tailor their interactions with generative AI systems to mitigate bias. Generative AI products could be designed to ask questions that help them get the desired output right. For instance, if ChatGPT is asked to produce an image of an Indian executive, rather than responding to the prompt automatically and creating an opening to assume the gender of the person depicted in the image, it could ask follow-up questions such as what gender the person should be.

Incorporate consumer feedback. Consumers should be given the opportunity to share feedback on the biased outputs emanating from generative AI products. For instance, after an output, a prompt could appear asking the consumer if they were satisfied, and if not, the reason for dissatisfaction. The developers of such AI products must also guard against taking on board positive feedback for gender-biased outputs, as that would further entrench the existing bias in the algorithm. Some companies like OpenAI have already made such facilities available to users.

Develop standards for red-teaming. In 2023, the President of the United States issued an executive order on The Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, which mandates red-teaming for certain high-risk generative AI models. Burt (2024) points out that while the measure is welcome, there is little clarity on what constitutes a red team, how to standardize testing procedures, and how to codify and distribute the results once the testing ends.[128] Even if it is excessive to expect this much detail from an executive order, the critique is valid, as without the appropriate context and standards, a red-teaming mandate may be difficult to implement. It may also be implemented inconsistently, with varying levels of stringency, diluting its potential as a solution for mitigating bias or other challenges raised by AI. India, and indeed other nations looking to mitigate the problem of gender bias in generative AI systems, may consider working with stakeholders from industry and civil society to develop standards for red teaming where applicable.

Create incentives for workforce and leadership diversity. Workforce and leadership diversity can play a key role in identifying and mitigating gender bias in generative AI. Policymakers should seek to introduce policies that promote and incentivize gender diversity in the AI sector. For instance, the US CHIPS and Science Act emphasises improving diversity within the STEM workforce, including the AI sector, through comprehensive data collection and focused recruitment and retention strategies.[129] It provides for the collection of detailed demographic data on job applicants as well as faculty involved in federally funded STEM programs, covering aspects such as race, ethnicity, sex, and other socioeconomic indicators. The data collection process aims to provide a clearer understanding of the prevailing diversity landscape, to identify areas that need targeted interventions.[130]

Increase the Threshold of Compliance in Accordance with the Harm Context: In contexts where generative AI may lead to bodily or personal injury harm, such as injury or death, such as medical uses where it is used for diagnoses, requirements may be introduced for representativeness of datasets as well as corroboration of output to ensure safety, and limit the potential for malpractice.

ENDNOTES

1. *IBM Global AI Adoption Index,* IBM (2024). https://in.newsroom.ibm.com/2024-02-15-59-of-Indian-Enterprises-have-actively-deployed-AI,-highest-among-countries-surveyed-IBM-report#:~:text=The%20'IBM%20Global%20AI%20Adoption,like%20R%26D%20and%20workforce%20reskilling

2. *'2023 Gender Social Norms Index: Breaking Down Gender Biases'*, United Nations Development Programme (2023). https://hdr.undp.org/system/files/documents/hdp-document/gsni202303pdf.pdf

3. *Gender Equality,* UNICEF India, https://www.unicef.org/india/what-we-do/gender-equality

4. Ibid.

5. *Indian Boards: Structure and Breadth,* Institutional Investor Advisory Services, (2021) https://www.iiasadvisory.com/institutional-eye/indian-boards-structure-and-breadth

6. Sritama Dutta, Sneha Singh, Tanveer Rehman, Srikanta Kanungo, and Sanghamitra Pati. *'Why India's Nationally Representative Surveys Need to Look beyond Gender Normativity'*, BMJ Glob Health (2023) https://gh.bmj.com/content/8/8/e012920

7. Ibid.

8. G.M Harshwardhan, M. Gourisaria, M. Pandey & S. Rautaray, '*A comprehensive survey and analysis of generative models in machine learning*', Computer Science Review, Vol. 38 (November 2020)

9. M. White, '*A brief history of generative AI*', Medium (January 2023); '*History of generative AI'*, Toloka ai (August 2023)

10. *History of generative AI'*, Toloka ai (August 2023)

11. G. Strobel, L. Banh, F. Moller, & T. Schoorman. ' *Exploring Generative Artificial Intelligence: A Taxonomy and Types*', Hawaii International Conference on System Sciences (January 2024), https://www.researchgate.net/publication/373927156_Exploring_Generative_Artificial_Intelligence_A_Taxonomy_and_Types

12. J. Rocca, '*Understanding Variational Autoencoders (VAEs)*', Medium (September 2019), https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

13. A. S. Choudhary, '*An Overview of Variational Autoencoders (VAEs)*', Analytics Vidhya (October 2023), https://www.analyticsvidhya.com/blog/2023/07/an-overview-of-variational-autoencoders/

14. K. Martineau, '*What is generative AI'*, IBM (April 2023), https://research.ibm.com/blog/what-is-generative-AI

15. Ibid.

16. J. Rocca, 'Understanding Variational Autoencoders (VAEs)', Medium (September 2019)

17. K. Martineau, '*What is generative AI'*, IBM (April 2023)

18. K. Martineau, '*What is generative AI'*, IBM (April 2023)

19. K. Martineau, '*What is generative AI'*, IBM (April 2023)

20. P. Salehi, A. Chalechale & M. Taghizadeh, '*Generative Adversarial Networks (GANs): An Overview of Theoretical Model, Evaluation Metrics, and Recent Developments*', Arxiv (2020) https://arxiv.org/ftp/arxiv/papers/2005/2005.13178.pdf

21. I.J. Goodfellow, J.P. Abadie, M. Mirza, B. Xu, D. Farley, S. Ozair, A. Courville & Y. Bengio '*Generative Adversarial Nets*', Arxiv (June 2014), https://arxiv.org/abs/1406.2661

22. P. Salehi, A. Chalechale & M. Taghizadeh, 'Generative Adversarial Networks (GANs): An Overview of Theoretical Model, Evaluation Metrics, and Recent Developments', Arxiv (2020)

23. Ibid.

24. A. Makhlouf, M. Maayah, N. Abughanam & C. Catal, '*The use of generative adversarial networks in medical image augmentation'*, Neural Computing and Applications 35(6) (October 2023), https://www.researchgate.net/publication/374749013_The_use_of_generative_adversarial_networks_in_medical_image_augmentation

25. G. Giacaglia, '*How Transformers Work- The Neural Network used by Open AI and DeepMind'*, Medium (March 2019), https://towardsdatascience.com/transformers-141e32e69591

26. J. Uszkoreit, '*Transformer: A Novel Neural Network Architecture for Language Understanding'*, Google Research (August 2017), https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

27. SD. Jurafsky & J. H. Martin, '*Transformers and Large Language Models*', Speech and Language Processing (Feb 2024), https://web.stanford.edu/~jurafsky/slp3/10.pdf

28. G. Giacaglia, 'How Transformers Work- The Neural Network used by Open AI and DeepMind', Medium (March 2019).

29. Ibid.

30. E. Zhang, A. Cheok, A. Pan, Z. Cai & Y. Yan, 'From Turing to Transformers: A comprehensive overview and tutorial on the evolution of applications of generative transformer models', Computational Linguistics and Artificial Intelligence (December 2023), https://www.mdpi.com/2413-4155/5/4/46

31. Ibid.

32. R. Merritt, '*What is a Transformer Model'*, Nvidia (March 2022), https://blogs.nvidia.com/blog/what-is-a-transformer-model/

33. '*What is a transformer model'*, IBM, https://www.ibm.com/topics/transformer-model

34. A. Neto, '*What is Latent Diffusion in AI*', Medium (October 2023), https://medium.com/@aguimarneto/what-is-latent-diffusion-in-ai-43aa1ad4f71e

35. L. Bouchard, '*How Stable Diffusion Works? Latent Diffusion Models Explained*' (August 2022) How Stable Diffusion works? Latent Diffusion Models Explained (louisbouchard.ai)

36. K. Erdem, '*Step by step visual introduction to Diffusion Models'*, Medium (November 2023)

37. A. Neto, '*What is Latent Diffusion in AI*', Medium (October 2023), https://www.louisbouchard.ai/latent-diffusion-models/

38. P. Gholami & R. Xiao, '*Diffusion Brush: Region Targeted Editing of AI Generated Images'*, Arxiv (October 2023) https://arxiv.org/pdf/2306.00219.pdf

39. A. Neto, '*What is Latent Diffusion in AI*', Medium (October 2023)

40. Ibid.

41. G. Lawton, '*Generative models: VAEs, GANs, diffusion, transformers, NeRFs'*, TechTarget (April 2023), https://www.techtarget.com/searchenterpriseai/tip/Generative-models-VAEs-GANs-diffusion-transformers-NeRFs

42. B. Marr, '*The difference between Generative AI and Traditional AI: an easy explanation for anyone'*, Forbes (June 2023)

43. Anand Mahurkar, 'Why Discriminative AI Will Continue to Dominate Enterprise AI Adoption in a World Flooded with Discussions on Generative AI'. Fast Company (25 July 2023) Why discriminative AI will continue to dominate enterprise AI adoption in a world flooded with discussions on generative AI (fastcompany.com)

44. Rishi Bommasani et al, *On the Opportunities and Risks of Foundation Models*, ArXiv (Aug 2021) http://arxiv.org/abs/2108.07258 arXiv: 2108.07258

45. Ganguli, Deep, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 'Predictability and Surprise in Large Generative Models'. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1747–64. Seoul Republic of Korea: ACM, 2022. doi:10.1145/3531146.3533229

46. Ganguli, Deep, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 'Predictability and Surprise in Large Generative Models'

47.The stages and their descriptions are derived from previous research by Bandi et al: https://www.mdpi.com/1999-5903/15/8/260 and Lama H. Nazer et al, at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10287014/

48. Ramya Srinivasan and Ajay Chander, *Biases in AI Systems: A Survey for Practitioners,* ACM Queue (2021) https://dl.acm.org/doi/pdf/10.1145/3466132.3466134

49. Ajay Bandi et al., *The Power of Generative AI: A review of Requirements, Models, Input-Output Formats, Evaluation Metrics, and Challenges,* Future internet (July 2023) https://www.mdpi.com/1999-5903/15/8/260

50. Ibid.

51. Reva Schwartz et al., *Towards a Standard for Identifying and Mitigating Bias in Artificial Intelligence,* National Institute of Standards and Technology, (March 2022) https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

52. Emilio Ferrara, *Should ChatGPT Be Biased: Challenges and Risks of Bias in Large Language Models,* First Monday Vol. 28 No. 11, (November 2023) https://arxiv.org/pdf/2304.03738.pdf

53. Samir Passi and Solon Barocas, *Problem Formulation and Fairness,* Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (2019) https://arxiv.org/pdf/1901.02547.pdf

54. Ramya Srinivasan and Kanji Uchino. '*Biases in Generative Art – A Causal Look from the Lens of Art History'*, ACM Conference on Fairness, Transparency, and Accountability (March 2021) https://arxiv.org/abs/2010.13266

55. Ibid.

56. Jasper Feine, Ulrich Gnewuch, Stefan Morana, Alexander Maedche, *Gender Bias in Chatbot Design,* Institute of Information Systems and Marketing (2019) https://conversations2019.files.wordpress.com/2020/01/conversations_2019_paper_6-preprint.pdf

57. Ibid.

58. Ibid.

59. Mark West et al., '*I'd blush if I could – Closing Gender Divides in Digital Skills through Education',* EQUALS and UNESCO (2019) https://unesdoc.unesco.org/ark:/48223/pf0000367416/PDF/367416eng.pdf.multi

60. Neeraja Kirtane and Tanvi Anand, *Mitigating Gender Stereotypes in Hindi and Marathi,* Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing, (May 2022) https://arxiv.org/pdf/2205.05901.pdf

61. Karolina Stanczak and Isabelle Augenstein, *A Survey on Gender Bias in Natural Language Processing*, University of Copenhagen (December 2021) https://arxiv.org/pdf/2112.14168.pdf

62. Gauri Gupta, Krithika Ramesh, Sanjay Singh, *Evaluating Gender Bias in Hindi-English Machine Translation,* Manipal Institute of Technology (June 2021) https://arxiv.org/pdf/2106.08680.pdf

63. Tom B. Brown, Benjamin Mann, Nick Ryder, et al; *Language Models are Few Shot Learners,* Advances in Neural Information Processing Systems, (July 2020) https://arxiv.org/pdf/2005.14165.pdf

64. Emilio Ferrara, *Should ChatGPT Be Biased: Challenges and Risks of Bias in Large Language Models*, First Monday Vol. 28 No. 11 (November 2023) https://arxiv.org/pdf/2304.03738.pdf

65. Summarized from Nima Shahbazi et al., *Representation Bias in Data: A Survey on Identification and Resolution Techniques,* ACM Computing Surveys Vol. 55, (March 2023) https://arxiv.org/pdf/2203.11852.pdf

66. Ramya Srinivasan and Ajay Chander, *Biases in AI Systems: A Survey for Practitioners,* ACM Queue (2021) https://dl.acm.org/doi/pdf/10.1145/3466132.3466134

67. Nima Shahbazi et al., *Representation Bias in Data: A Survey on Identification and Resolution Techniques,* ACM Computing Surveys Vol. 55, (March 2023) https://arxiv.org/pdf/2203.11852.pdf

68. Nima Shahbazi et al., *Representation Bias in Data: A Survey on Identification and Resolution Techniques,* ACM Computing Surveys Vol. 55, (March 2023) https://arxiv.org/pdf/2203.11852.pdf

69. Summarized from Nima Shahbazi et al., Representation Bias in Data: A Survey on Identification and Resolution Techniques.

70. *Generative AI: Perspectives,* Stanford Human-Centered AI (March 2023)

71. Sritama Dutta, Sneha Singh, Tanveer Rehman, Srikanta Kanungo, and Sanghamitra Pati. *'Why India's Nationally Representative Surveys Need to Look beyond Gender Normativity'*, BMJ Glob Health (2023) https://gh.bmj.com/content/8/8/e012920

72. Ramya Srinivasan and Ajay Chander, *Biases in AI Systems: A Survey for Practitioners,* ACM Queue (2021) https://dl.acm.org/doi/pdf/10.1145/3466132.3466134

73. Ramya Srinivasan and Kanji Uchino. '*Biases in Generative Art – A Causal Look from the Lens of Art History'*, ACM Conference on Fairness, Transparency, and Accountability (March 2021) https://arxiv.org/abs/2010.13266

74. *Generative AI: Perspectives,* Stanford Human-Centered AI (March 2023) https://hai.stanford.edu/sites/default/files/2023-03/Generative_AI_HAI_Perspectives.pdf

75. *Generative AI: Perspectives,* Stanford Human-Centered AI (March 2023)

76. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, *'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *', Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, (2021), https://dl.acm.org/doi/10.1145/3442188.3445922

77. Ibid.

78. Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag, '*Quantifying Gender Bias in Different Corpora,'* Companion Proceedings of the Web Conference (2020), https://lin-web.clarkson.edu/~jmatthew/publications/GenderBias_FATES2020.pdf

79. R. Sathya, Annamma Abraham, *Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification*, IJARAI (2013) https://thesai.org/Downloads/IJARAI/Volume2No2/Paper_6-Comparison_of_Supervised_and_Unsupervised_Learning_Algorithms_for_Pattern_Classification.pdf

80. Xu Han, et al, *Pre-trained models: Past, present and future,* KeAI (2021) https://www.sciencedirect.com/science/article/pii/S2666651021000231

81. Ibid.

82. Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. '*Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings'* (2016) http://arxiv.org/abs/1607.06520

83. Ibid.

84. Ramya Srinivasan and Ajay Chander, *Biases in AI Systems: A Survey for Practitioners,* ACM Queue (2021) https://dl.acm.org/doi/pdf/10.1145/3466132.3466134

85. Ryan Steed, Aylin Caliskan, *Image Representation Learned with Unsupervised Pre-training contain human like biases,* FAccT, (January 2021) https://arxiv.org/pdf/2010.15052.pdf

86. Ibid.

87. Ibid.

88. R. Sathya, Annamma Abraham, *Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification*, IJARAI (2013) https://thesai.org/Downloads/IJARAI/Volume2No2/Paper_6-Comparison_of_Supervised_and_Unsupervised_Learning_Algorithms_for_Pattern_Classification.pdf

89. David Brown et al., *On the Opportunities of Generative AI Models,* (December 2023), https://www.researchgate.net/profile/Emily-Johnson-167/publication/376482883_On_the_Opportunities_of_Generative_AI_Large_Models/links/657a557a6610947889c5d09a/On-the-Opportunities-of-Generative-AI-Large-Models.pdf

90. Ibid.

91. Ibid.

92. Ramya Srinivasan, Ajay Chander, *Crowdsourcing in the Absence of Ground Truth – A case study,* Cornell University (2019) https://arxiv.org/abs/1906.07254

93. Carsten Schwemmer, Carly Knight, et al, *Diagnosing Gender Bias in Image Recognition Systems,* Socius (2020) https://journals.sagepub.com/doi/full/10.1177/2378023120967171

94. Siddharth D Jaiswal, Ankit Verma, et al, *Auditing Gender Analyzers on Text*, IEEE/ACM ASONAM (2023) https://arxiv.org/pdf/2310.06061.pdf

95. Ramya Srinivasan and Ajay Chander, *Biases in AI Systems: A Survey for Practitioners,* ACM Digital Queue (2021), https://dl.acm.org/doi/10.1145/3466132.3466134

96. Rohan Taori and Tatsunori B. Hashimoto. 'Data Feedback Loops: Model-Driven Amplification of Dataset Biases', 2022. doi:10.48550/ARXIV.2209.03942

97. Tim Salimans, Ian Goodfellow et al., *Improved Techniques for Training GANS,* ArXiv (2016) https://arxiv.org/abs/1606.03498

98. Yogarshi Vyas et al., *Quantifying Social Biases in NLP: A Generalization and Empirical Comparison,* Transactions of the Association for Computational Linguistics (2021) https://arxiv.org/pdf/2106.14574.pdf

99. Ibid.

100. Seraphina Tarrant et al, *Intrinsic Bias Metrics Do Not Correlate with Application Bias* (2020) https://arxiv.org/abs/2012.15859

101. Hila Gonen and Yoav Goldberg, *'Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But Do Not Remove Them,'* (September 2019) https://arxiv.org/pdf/1903.03862.pdf

102. Paraphrased from Barla, Nilesh. 'The Ultimate Guide to Word Embeddings'. *Neptune.Ai*, 21 July 2022. https://neptune.ai/blog/word-embeddings-guide

103. Ibid.

104. Adi Roberston, *Google apologizes for 'missing the mark after the Gemini generated racially diverse Nazis,* The Verge (February 2024) https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

105. Jintang Xue et al., *Bias and Fairness in Chatbots: An overview* (December 2023) https://arxiv.org/html/2309.08836v2

106. Lama H. Nazer, et al, *Bias in artificial intelligence algorithms and recommendations for mitigation,* PLOS Digital Health (June 2022) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10287014/

107.Gizem Sogancioglu and Heysem Kaya, *'The Effects of Gender Bias in Word Embeddings on Depression Prediction',* PAI4MH Workshop (December 2022), https://arxiv.org/abs/2212.07852v1

108. Amama Mahmood and Chien-Ming Huang, *Gender Biases in Error Mitigation by Voice Assistants* (October 2023) https://arxiv.org/pdf/2310.13074.pdf

109. Jintang Xue, Yun- Chen Wang, *Bias and Fairness in Chatbots* (December 2023) https://arxiv.org/html/2309.08836v2

110. Ibid.

111. *How Your Data is Used to Improve Model Performance,* OpenAI (2024), https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

112. Ibid.

113. Regina Bernhaupt et al., *Methods in Human-computer interaction research and practice: challenges and innovations,* Interacting with Computers, (July 2023) https://academic.oup.com/iwc

114. Ninareh Mohrabbi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, *'A Survey on Bias and Fairness in Machine Learning'* (January 2022), https://arxiv.org/abs/1908.09635

115. Ibid.

116. Sensity, *Deepfakes vs Biometric KYC Verification,* (2022) https://sensity.ai/reports/

117. Sarah Myers West, *Discriminating Systems: Gender, Race and Power in AI – A Report,* AI Now Institute (2019), https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2

118. Sarah Myers West, *Discriminating Systems: Gender, Race and Power in AI – A Report,* AI Now Institute (2019), https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2

119. Judy Wajcman et al., *The Digital Revolution: Implications for gender equality and women's rights 25 years after Beijing,* UN Women (August 2020), https://www.unwomen.org/sites/default/files/Headquarters/Attachments/Sections/Library/Publications/2020/The-digital-revolution-Implications-for-gender-equality-and-womens-rights-25-years-after-Beijing-en.pdf

120. Genevieve Smith and Ishita Rustagi, *When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equality,* Stanford Social Innovation Review (March 2021), https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity

121. Ankita Deshkar, *Bridging the gap: Women in AI and the challenges to inclusion,* Indian Express (March 2024), https://indianexpress.com/article/technology/artificial-intelligence/bridging-the-gap-women-artificial-intelligence-9202984/

122. Carlota Quiros et al., *Women in the Digital Age, European Commission,* Directorate-General for Communications Networks, Content and Technology (2018), https://op.europa.eu/en/publication-detail/-/publication/84bd6dea-2351-11e8-ac73-01aa75ed71a1

123. Ibid.

124. Samuel Deng, et al, *Biased Programmers? Or Biased Data? A field experiment in Operationalizing AI Ethics* (December 2020) https://arxiv.org/pdf/2310.11079.pdf

125. Su et al, *Learning from Red Teaming: Gender Bias Provocation and Mitigation in Large Language Models* (2023), https://arxiv.org/abs/2310.11079

126. Sorelle Friedler, et al, *AI Redteaming is not a one stop solution to AI Harms,* Data & Society (2023) https://datasociety.net/wp-content/uploads/2023/10/Recommendations-for-Using-Red-Teaming-for-AI-Accountability-PolicyBrief.pdf

127. T.D Jui, *Fairness issues, current approaches, and challenges in machine learning models,* International Journal of Machine Learning and Cybernetics, (January 2024), https://link.springer.com/article/10.1007/s13042-023-02083-2

128. Andrew Burt, *How to red team a Gen AI Model,* Harvard Business Review (January 2024) https://hbr.org/2024/01/how-to-red-team-a-gen-ai-model

129. Andrea Peterson, *Workforce Diversity Initiatives in the CHIPS and Science Act,* American Institute of Physics, (Sept 2022) https://ww2.aip.org/fyi/2022/workforce-diversity-initiatives-chips-and-science-act

130. S. 10312, The US Chips and Science Act, H.R. 4346, https://www.congress.gov/117/bills/hr4346/BILLS-117hr4346enr.pdf

B-40 First Floor
Soami Nagar South
New Delhi - 110017
contact@esyacentre.org
www.esyacentre.org