Bias in Language Models: A Survey

Arif Ahmad, Pushpak Bhattacharyya

CFILT, Indian Institute of Technology Bombay
190110010@iitb.ac.in, pb@cse.iitb.ac.in

Abstract

Language models (LMs), trained on extensive text corpora, exhibit impressive capabilities across diverse natural language processing (NLP) tasks but also risk propagating entrenched societal biases. This paper examines the manifestation and amplification of such biases, particularly within the context of India—a region rich in linguistic, religious, and cultural diversity. Given the predominance of Western-centric research and benchmarks, there is a significant need for frameworks that effectively address and mitigate biases in multilingual and culturally diverse settings. Through a comprehensive survey, we explore various dimensions of bias in LMs, focusing on both discriminatory tendencies and the mechanisms through which these biases are embedded and perpetuated in model outputs. Additionally, evaluating models' cultural competence and the ethical implications of their use in global settings is of great importance. By integrating insights from a wide range of studies and benchmark datasets, this survey highlights the critical need for more inclusive and equitable AI practices, proposing a refined approach to developing and evaluating models that are truly representative of and responsive to the diverse fabric of global societies.

1 Introduction

Language models (LMs), which are trained on expansive textual datasets, have shown impressive performance across a spectrum of natural language processing (NLP) tasks. Despite their effectiveness, there is growing concern about their propensity to propagate existing societal biases and stereotypes, which are often embedded within their training data (Blodgett et al., 2020a; Bender et al., 2021a; Sahoo et al., 2022). Such biases, when manifested in NLP applications, can have deleterious effects on various demographic groups, making it imperative to develop robust benchmarks that can accurately gauge these models' biases in diverse social contexts (Savoldi et al., 2021; Ziems et al., 2022; Mozafari et al., 2020).

In a multicultural nation like India, where the societal fabric is intricately woven with varied languages, religions, castes, and regional identities, the urgency for effective bias mitigation frameworks becomes even more pronounced. The challenge is compounded by the predominance of research and benchmark datasets like Nangia et al. (2020a) and Nadeem et al. (2021), which largely cater to English and Western cultural norms. This oversight leads to significant gaps in our understanding and capabilities to counteract biases in contexts relevant to the Indian milieu (Blodgett et al., 2021a). The complexity of social identities in India further necessitates the examination of intersectional biases, which remain largely unaddressed.

Bias within LMs manifests through discriminatory tendencies towards specific demographic groups or sensitive issues (Hammersley and Gomm, 1997; Singh et al., 2022). Extensive studies confirm that these biases are not merely reflections but also amplifications of societal prejudices inherent in the data used for training these models (Bolukbasi et al., 2016; Jia et al., 2020; Zhao et al., 2017; Sheng et al., 2021). While initial mitigation efforts were predominantly focused on Western contexts, there is a burgeoning body of work exploring biases in data representing diverse languages and cultural backgrounds, such as Arabic (Lauscher et al., 2020), French (Kurpicz-Briki, 2020), and Italian (Sanguinetti et al., 2020). Nevertheless, studies specifically addressing the Indian context remain sparse, though some recent initiatives are bridging

this gap (Sahoo et al., 2024; Pujari et al., 2020; Malik et al., 2022; Sambasivan et al., 2021; Bhatt et al., 2022; Jha et al., 2023).

Existing methodologies for creating bias benchmarks often utilize predefined word sets or template-based sentences to assess biases concerning particular demographics and sensitive attributes (Caliskan et al., 2017; May et al., 2019; Manzini et al., 2019). Despite these efforts, the lack of focus on non-Western cultures and the intricacies of multilingual contexts in India is a critical shortfall that this research aims to address. By extending existing benchmarks to include the Hindi language and focusing on regional and religious stereotypes, this study endeavors to construct a more comprehensive understanding of social biases in LMs.

The exploration of bias extends beyond academia into real-world platforms like Twitter, which, despite its recent rebranding to 𝕏, continues to be a significant venue for public discourse (Malik et al., 2019). It is not only important to examine the factors contributing to the virality of tweets but also investigates how embedded social biases may influence this virality, potentially exacerbating social divisions and perpetuating stereotypes (Amon et al., 2020; Hasan et al., 2021; Guo et al., 2022; Elmas, 2023).

In tandem with these social studies, this paper also discusses the ethical implications of text-to-image (T2I) generative models like Stable Diffusion-XL, Imagen 2, and DALLE-3 (Podell et al., 2023; Saharia et al., 2022; Betker et al., 2023). These models, while transforming creative industries, must be scrutinized for their cultural competence—particularly their ability to represent and respect the diversity of global cultures accurately (Bird et al., 2023; Weidinger et al., 2023). By focusing on cultural awareness and diversity, one can evaluate these models' performances across a spectrum of cultural contexts, thereby addressing crucial gaps in global technological equality (Prabhakaran et al., 2022; Jha et al., 2024; Basu et al., 2023).

Recent advancements in text-to-image (T2I) generative models, exemplified by Stable Diffusion-XL (Podell et al., 2023), Imagen 2 (Saharia et al., 2022), and DALLE-3 (Betker et al., 2023), have transformed creative industries like digital arts, advertising, and education. These models offer unprecedented capabilities in creative expression and communication, suggesting a potential revolution across various sectors. However, their global proliferation has also brought forth significant ethical and social considerations (Bird et al., 2023; Weidinger et al., 2023), particularly in ensuring equitable and inclusive functionality across diverse cultures (Qadri et al., 2023; Mim et al., 2024).

Historically, T2I model evaluations have concentrated on photo-realism and accuracy (Saharia et al., 2022; Hu et al., 2023; Cho et al., 2024; Huang et al., 2023), but recent findings highlight significant gaps in how these models handle cultural content (Cho et al., 2023; Bianchi et al., 2023; Luccioni et al., 2024). These gaps stem largely from the models' development within mono-cultural ecosystems, which may not adequately represent the diversity of global cultures, thereby risking the perpetuation of cultural biases and stereotypes (Prabhakaran et al., 2022). In response, this paper focuses on geo-cultural differences, defined here as the cultures formed within specific national boundaries, to explore how T2I models perform across varied cultural settings (Rapport and Overing, 2002; Li et al., 2024).

These disparities can lead to the suppression of sub- and co-cultures and limit the models' applicability across different geo-cultural contexts (Qadri et al., 2023; Mim et al., 2024). Despite the efforts to build resources that detect biases and stereotypes, there remains a lack of comprehensive evaluation tools that assess the richness and diversity of cultural representations in T2I models (Jha et al., 2024; Basu et al., 2023).

In summary, this paper presents a comprehensive survey on bias in models, focusing on the multifaceted challenges posed by discriminatory tendencies within language models and text-to-image generative capabilities. Through this endeavor, we not only enhance our understanding of biases inherent in models trained on multilingual and multicultural data but also pave the way for more equitable AI practices. Our approach stands to influence future research directions, encouraging a broader application of fairness and inclusivity standards in AI technologies globally.

Thus, this survey contributes to the ongoing discourse on mitigating bias in AI, promoting a technology ecosystem that is truly reflective of, and responsive to, the rich diversity of human society.

2 Definition of Bias

This section delineates the concepts of "bias" and "fairness" within the realm of large language models (LLMs), highlighting the nuances of social bias as it manifests in natural language processing (NLP) tasks and throughout the lifecycle of LLM development and deployment.

2.1 Social Bias and Fairness

The imperative to mitigate social bias and ensure fairness in NLP systems is a significant theme in recent research. Efforts typically involve technical solutions such as enhancing datasets to balance representation of social groups or adjusting model objectives to promote fairness. Yet, despite these efforts, there is often a lack of clarity about the specific harms caused by biased model behaviors: identifying *who* is harmed, understanding *why* such behavior is detrimental, and discerning *how* these behaviors reflect and perpetuate existing social hierarchies (Blodgett et al., 2020b, 2021b).

In this context, many strategies propose an ideal criterion—typically that model outputs should not vary based on social group characteristics within the inputs. However, these frameworks frequently neglect to articulate the underlying normative social values that justify such criteria. This section aims to clarify these concepts, drawing from foundational works in machine learning and sociolinguistics (Barocas et al., 2019; Bender et al., 2021b; Crawford, 2017; Mehrabi et al., 2021; Suresh and Guttag, 2021; Weidinger et al., 2022; Beukeboom and Burgers, 2019; Craft et al., 2020; Loudermilk, 2015; Maass, 1999).

Gallegos et al. (2024) propose refined definitions of "bias" and "fairness," with a focus on detaching these definitions from specific technical mechanisms, recognizing that language itself is a medium that inherently carries cultural and social values. This perspective aligns with the understanding that social groups, although often legally defined, are fundamentally social constructs that can reinforce existing power dynamics and perpetuate discrimination.

2.1.1 Definitions and Taxonomy

We begin by establishing clear definitions for terms critical to our discussion:

Social Group A *social group* represents a segment of the population that shares certain identity traits, which can be inherent, contextual, or socially constructed. Such groups often include those recognized by anti-discrimination laws—referred to as "protected groups" or "protected classes"—which can include characteristics like age, ethnicity, disability, gender identity, national origin, race, religion, sex, and sexual orientation.

Protected Attribute A *protected attribute* is an identity trait that defines the membership of individuals within a social group.

Recognizing the fluid and often contested nature of these group delineations is essential, as they can legitimize disparities, reinforce societal hierarchies, and have tangible, adverse impacts on marginalized communities (Hanna et al., 2020; Beukeboom and Burgers, 2019).

Social Bias *Social bias* refers to the uneven treatment or outcomes across different social groups, arising from deep-seated structural and power asymmetries. This encompasses both representational harms (e.g., stereotypes, derogatory language) and allocational harms (e.g., discrimination) that are often linked to societal norms and the distribution of power.

3 Characterization of Social Biases in Indian Context

In India, the societal fabric is intricately woven with deep-seated biases linked to disparities such as *Caste, Religion, and Region*, which significantly impact the social dynamics and interactions within the country. The prevalence of caste-based discrimination remains a critical issue, highlighted in historical and contemporary academic analyses, such as those by Ambedkar (2014) in "Annihilation of Caste". These studies emphasize the persistent inequalities that affect marginalized communities, including Dalits, Adivasis, and Denotified

Tribes, despite legal and social reforms aimed at eradicating such discrimination.

Regional biases in India are also profound, as certain stereotypes are commonly associated with people from specific regions. de Souza (1977) in "Regional Stereotypes and Identities in India" was among the first to document the association of particular character traits with regional identities. This work has been expanded by recent studies such as those by Bhatt et al. (2022) in "Contextualizing Stereotypes and Bias in Indian Language Models", which demonstrate these stereotypes' persistence in modern datasets and language models like MuRIL and mBERT. Additional research by Sahoo et al. (2023) in "Prejudice in Indian Language Models", Rajadesingan et al. (2019) in "Smart, Responsible, and Upper Caste Only: Measuring Caste Attitudes through Large-Scale Analysis of Matrimonial Profiles", and Haokip (2021) in "Chinky, Tribals, and Terrorists: Understanding Racial Epithets in the Indian Context", further detail the specific biases faced by various subgroups within Indian society.

Religious disparities are similarly pervasive, with biases deeply rooted in the inter-religious dynamics of the country. Research such as Sabharwal and Sonalkar (2015) in "Dalits and Religious Conversions: Subjectivity and the Socio-Political Context" and McDuie-Ra (2012) in "Northeast India: Addressing Stereotypes and Fostering Understanding" highlights the interplay between religion and caste, and the regional nuances that influence these biases.

Global social disparities such as *Gender, Age, and Physical Appearance* manifest with unique characteristics in the Indian context. While stereotypes such as "Women can't do math" are globally prevalent, local narratives provide a richer texture. For example, women in traditional attire might be perceived differently across Indian states, illustrating how regional and cultural settings influence the reception and perpetuation of stereotypes:

S1: *Women wearing traditional attire in* ***Rajasthan*** *are seen as* ***conservative.***

S2: *Women wearing traditional attire in* ***West Bengal*** *are seen as* ***cultural ambassadors.***

These examples highlight the complex interplay between global stereotypes and localized cultural narratives, showing how the same attribute can have varying interpretations based on the regional context.

The integration of NLP technologies in sectors like legal, medical, education, and media in India necessitates a critical examination of the biases these technologies may carry. It is imperative that the research community develops and utilizes diverse, reliable, and context-specific benchmark datasets designed to measure and mitigate model biases. Such efforts are crucial for advancing the fairness of NLP applications and ensuring that they serve all sections of Indian society equitably, as underscored in "Advancing AI Fairness in India" by Pujari et al. (2020).

4 Bias Datasets

The exploration of bias in language models (LMs) necessitates a comprehensive examination of the datasets used to detect, quantify, and mitigate biases. This section reviews bias datasets, categorizing them based on their linguistic and cultural scopes as well as the types of biases they address.

4.1 English and Western-Centric Datasets

Much of the initial work in bias datasets has focused on English and is oriented toward Western societal norms. Datasets such as Stereoset (Nadeem et al., 2021) and Crows-pairs (Nangia et al., 2020a) have been instrumental in identifying and quantifying biases. Stereoset evaluates a model's propensity to choose stereotypical over non-stereotypical responses, offering insights into ingrained biases. Crows-pairs, by contrast, employs pairs of minimally differing sentences to highlight discriminatory behaviors in LMs across various demographics. These datasets play a crucial role in uncovering the extent of biases that emerge from predominantly Western data sources.

4.2 Datasets Focusing on Non-Western Contexts

As global awareness of AI's impact increases, the scope of bias research has expanded to include diverse cultural and linguistic contexts. For instance, the French extension of Crows-pairs (Névéol et al., 2022) adapts the

methodology to better fit French cultural nuances, marking a significant step toward inclusivity. Sahoo et al. (2024) introduce "Indibias" dataset which includes Hindi extension of Crows-pairs as well as Intersectional Bias measurement benchmark. However, such efforts remain limited, and comprehensive datasets for many non-Western languages and regions, especially those with complex socio-cultural dynamics like India, are still lacking.

4.3 Indian Context-Specific Datasets

Recognizing India's unique diversity, recent initiatives have developed datasets that specifically address biases pertinent to its sociocultural context. These efforts include frameworks and datasets such as those proposed by Sambasivan et al. (2021) and Bhatt et al. (2022), which consider multiple axes of identity like caste, religion, and region. However, there is a notable scarcity of datasets in regional Indian languages, which indicates a gap in the resources available to study and mitigate biases in India's multilingual landscape.

4.4 Intersectional Bias Datasets

The complexity of human identities demands datasets that can capture intersectional biases, where multiple axes of identity intersect. Tan and Celis (2019) have begun to address this need by developing methodologies to study biases across multiple demographics simultaneously, such as race and gender. These datasets are particularly important for understanding the nuanced ways in which biases manifest in AI systems but are still relatively underdeveloped for regions with intricate social structures like India.

4.5 Emerging Trends and Challenges

The development of bias datasets is fraught with challenges. Ensuring that these datasets are representative of diverse populations and remain relevant over time as societal norms evolve is a daunting task. Additionally, as AI technologies advance, the methods to detect and mitigate biases must also evolve.

In summary, bias datasets are crucial tools for understanding the biases embedded in LMs. The progression of the field towards more inclusive and comprehensive datasets will play a key role in ensuring that AI technologies are equitable and fair across all user demographics. Efforts to broaden the linguistic and cultural inclusivity of these datasets are essential for creating AI systems that are truly beneficial to global societies.

5 Bias in NLP Tasks

Language models, by their very nature, are deeply intertwined with social identity, power, and the dynamics of societal structures. They not only reflect but can also reinforce the categorizations and stereotypes embedded in language, which can manifest through both overt and subtle biases in various NLP tasks.

- **Text Generation:** Bias may manifest as skewed associations within localized contexts or through broader narrative arcs, affecting the representation of different social groups.
- **Machine Translation:** Translation models might inadvertently favor certain gender pronouns over others in gender-neutral contexts, reflecting and potentially reinforcing gender biases.
- **Information Retrieval:** Search algorithms could prioritize content that aligns with majority group perspectives, marginalizing minority viewpoints.
- **Question-Answering:** Models might rely on societal stereotypes when generating responses to ambiguous queries, potentially perpetuating harmful biases.
- **Natural Language Inference:** Inferential models could draw inappropriate conclusions based on biased premises, further entrenching stereotypes.
- **Classification:** Classification tasks, such as sentiment analysis or toxicity detection, might show differential treatment based on dialects, language variations, or demographic indicators embedded within the text.

These examples underscore the necessity of a nuanced approach to fairness, one that comprehensively addresses the multifaceted ways in which biases manifest in NLP. By refining our understanding and definitions of bias and

fairness, we better equip the research community to develop more equitable language technologies that respect and reflect the diversity of human experience.

6 Bias Benchmarking

Here we discuss the types of datasets utilized in the literature to evaluate bias and unfairness in large language models (LLMs). Gallegos et al. (2024) provides a structured classification based on the nature and structure of these datasets, aiming to guide the selection of appropriate metrics for bias evaluation.

6.1 Counterfactual Inputs

Counterfactual datasets typically consist of pairs or tuples of sentences designed to highlight discrepancies in model predictions across different social groups. These datasets employ a counterfactual approach where one variable (usually the social group) is altered in a sentence while keeping all other elements constant to observe changes in the model's outputs. This alteration could affect the probabilities of predicted tokens or the content of generated text, revealing potential biases.

6.1.1 Masked Tokens

Datasets in this category include sentences with a placeholder that the model needs to fill, often with options like gender-specific pronouns or terms that reflect stereotypical or counter-stereotypical attributes. These are particularly suited for evaluating models using masked token probability-based metrics or pseudo-log-likelihood metrics to assess the likelihood of specific fill-in-the-blank responses. They can also be used with accuracy metrics when multiple-choice answers are provided.

For instance, coreference resolution tasks often use such datasets. The Winograd Schema Challenge, proposed by (Levesque et al., 2012), and its derivatives like **Winogender** (Rudinger et al., 2018) and **WinoBias** (Zhao et al., 2018), are classic examples. These schemas challenge a model to resolve pronouns accurately in sentences that only differ by gender or other social terms, thus providing a direct measure of bias:

> "The doctor informed the patient that [MASK: **she**/**he**/**they**] would need to adjust their diet."

6.1.2 Unmasked Sentences

Unlike masked tokens datasets, unmasked sentences datasets do not involve fill-in-the-blank tasks but rather present complete sentences. The model is tasked with evaluating which sentence in a pair is more likely, which can be especially revealing when the sentences differ only in terms of demographic terms. This setup allows for the application of the same metrics used for masked tokens, and also enables comparisons using generated text-based metrics.

A notable example is the **Crowdsourced Stereotype Pairs (CrowS-Pairs)** dataset by (Nangia et al., 2020b), which includes pairs of sentences reflecting stereotypes versus a neutral or counter-stereotypical counterpart, covering various dimensions such as race, gender, and age.

6.2 Bias benchmarking Dataset Uses

Drawing from extensive discussions and analyses in the literature, especially the critiques by (Blodgett et al., 2021b), the following recommendations for using bias evaluation datasets are important depending on the use-case:

- Ensure that the datasets clearly define and articulate the specific forms of bias they aim to measure. It is crucial that the datasets not only capture stereotypical expressions but also accurately reflect the underlying power dynamics and societal contexts they are meant to represent.
- When selecting datasets, consider the cultural and demographic contexts they are designed for. Datasets developed within specific national or cultural settings might not be generalizable to other contexts, thus limiting their applicability.
- Given the potential for datasets to exhibit limitations in scope and depth, it is advisable to use multiple datasets to cross-validate findings and ensure a comprehensive evaluation of bias across different dimensions and scenarios.

These recommendations are aimed at fostering more accurate, reliable, and contextually appropriate evaluations of bias in LLMs, facilitating the development of more fair AI systems.

7 Metrics for Bias Evaluation

Gallegos et al. (2024) proposes a structured taxonomy for evaluating fairness in large language models (LLMs). While recent surveys, such as the one by Chang et al. (2023), have reviewed evaluation techniques for LLMs, they have not specifically addressed metrics for assessing fairness and bias. Here, we explore various metrics, formalize them mathematically, provide examples, and discuss the challenges each faces. This categorization of fairness evaluation metrics helps us understand and critique their effectiveness and limitations.

The evaluation of biases in LLMs requires consideration of multiple facets, each contributing uniquely to the understanding and measurement of bias:

- **Task-Specific Metrics:** Metrics and the datasets used for bias measurement are often tailored to specific NLP tasks like text generation, classification, or question-answering. These metrics are designed to capture biases that manifest uniquely across these varied tasks.
- **Bias Type:** The type of bias a metric can measure largely depends on the dataset employed.
- **Data Structure (Input to Model):** Metrics also vary by the type of data structure they assume. For example, several metrics apply to datasets consisting of sentence pairs, where one sentence is biased and the other is not, or is considered less biased.
- **Metric Input (Output from Model):** The input required by the metric—whether it be embeddings, model-generated probabilities, or the text output—also defines its applicability and effectiveness in measuring bias.

7.1 Taxonomy of Metrics Based on *What They Use*

Bias evaluation metrics can be classified based on the type of data they utilize from the LLMs, such as embeddings, probabilities, or generated text. This classification helps in understanding which metrics are best suited for particular types of model outputs.

- **Embedding-based metrics:** These metrics use the dense vector representations (embeddings) from models to measure bias. They are effective in capturing biases that are encoded in the geometric space of the embeddings.
- **Probability-based metrics:** These metrics utilize the probabilities output by models to estimate bias. This includes comparing probabilities assigned to different sociodemographic groups or assessing changes in probabilities when the input is perturbed.
- **Generated text-based metrics:** These metrics analyze the text generated by models to detect biases. They are useful in models where the direct outputs are texts, such as in dialogue systems or text generators.

Each category of metrics has its strengths and limitations, and their effectiveness can vary based on the specific characteristics of the bias being measured and the model being evaluated.

7.2 Embedding-Based Metrics

Embedding-based metrics typically measure distances or angles between embeddings to quantify bias. For instance, if embeddings of words related to certain demographics are closer to negative sentiment words, this could indicate bias. These metrics are potent for exploring how representational biases are embedded within the vector space of a model's outputs.

7.3 Probability-Based Metrics

Probability-based metrics assess how likely a model is to output certain responses based on the input's demographic characteristics. For example, if changing a name in a sentence from a typically male name to a female one changes the model's output probabilities significantly, this might indicate gender bias.

7.4 Generated Text-Based Metrics

Metrics that analyze generated text look at the content and structure of text outputs from models to identify biases. These metrics

are particularly relevant for generative models like GPT-3, where the nuances of the generated text—such as the themes or entities mentioned—can reveal underlying biases.

7.5 Challenges in Bias Evaluation

While the metrics described provide valuable tools for bias evaluation, they also come with limitations:

- **Context Dependency**: The effectiveness of bias metrics can vary greatly depending on the context in which they are used, including the specific tasks, datasets, and model architectures.
- **Interpretability**: Some metrics, especially those involving complex mathematical formulations, can be challenging to interpret, making it difficult to translate metric outcomes into actionable insights.
- **Coverage**: No single metric can capture all forms of bias. It is often necessary to use a combination of different metrics to get a comprehensive view of biases in a model.

The development and refinement of metrics for evaluating bias in LLMs are critical for advancing fairness in AI. By understanding the strengths and limitations of different metrics and applying them thoughtfully, we can better identify and mitigate biases in AI models, leading to more equitable and trustworthy systems. The proposed taxonomy provides a structured way to navigate the landscape of bias metrics, helping researchers and practitioners select the most appropriate tools for their specific needs.

8 Summary and Conclusions

This paper has systematically explored the pervasive and multifaceted issue of bias in language models (LMs) generative models, with a significant focus on the cultural and linguistic complexities within the Indian context. Through a comprehensive review of existing methodologies, benchmark datasets, and mitigation strategies, we have illuminated the critical need for inclusive and equitable AI practices that accommodate the rich diversity of global societies.

Key takeaways include:

- **Bias in LMs is deeply ingrained**, often reflecting and amplifying societal stereotypes and prejudices that exist in the training data. These biases are not confined to any single region or language, although there is a noticeable lack of research and resources addressing non-Western contexts.
- **Existing benchmarks and metrics for assessing bias**, while useful, predominantly cater to English and Western norms and fail to capture the unique socio-cultural dynamics of other regions, notably India. This oversight complicates the task of effectively identifying and mitigating biases in such diverse settings.
- **The impact of biased AI technologies** is profound, influencing a wide range of applications from automated text generation to dynamic image creation, and extending to critical domains such as healthcare, legal, and education sectors where the stakes of perpetuating biases are particularly high.

In conclusion, there is an urgent need for several key actions to improve bias mitigation in AI. First, the development of robust, context-aware benchmarks that are sensitive to the linguistic and cultural intricacies of all regions, including the Indian subcontinent, is crucial. This requires crafting datasets and metrics that respect and reflect the diversity within these contexts. Second, it is essential to advance mitigation strategies that address not only the symptoms but also the root causes of biases in AI models. This may involve revising model training practices, diversifying data sources, and incorporating ethical considerations into the development lifecycle of AI technologies. Third, promoting transparency and accountability in AI development is vital to ensure that AI systems are not only technically proficient but also socially and ethically responsible.

As AI technologies continue to evolve, the dialogue on bias mitigation must also progress, expanding to include voices from diverse cultural and professional backgrounds. By doing so, we can harness the full potential of AI to benefit society universally, ensuring that it

acts as a tool for social good, enhancing rather than compromising fairness and inclusivity.

In moving forward, it is imperative for the research community, industry stakeholders, and policymakers to collaborate in fostering an AI ecosystem that is as diverse as the human experience it seeks to augment.

References

B.R. Ambedkar. 2014. *Annihilation of Caste: The Annotated Critical Edition*. Verso Books.

Mary Jean Amon, Rakibul Hasan, Kurt Hugenberg, Bennett I. Bertenthal, and Apu Kapadia. 2020. Influencing photo sharing decisions on social media: A case of paradoxical findings. *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1350–1366.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the geographical representativeness of images from text-to-image models. *Preprint*, arXiv:2305.11080.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021a. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021b. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Camiel J Beukeboom and Christian Burgers. 2019. How stereotypes are shared through language: a review and introduction of the social categories and stereotypes communication (SCSC) framework. *Review of Communication Research*, 7:1–37.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504.

Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 396–410.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020a. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020b. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021a. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021b. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016.

Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Preprint*, arXiv:1607.06520.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2024. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *Preprint*, arXiv:2310.18235.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054.

Justin T Craft, Kelly E Wright, Rachel Elizabeth Weissler, and Robin M Queen. 2020. Language and discrimination: Generating meaning, perceiving identities, and discriminating outcomes. *Annual Review of Linguistics*, 6:389–407.

Kate Crawford. 2017. The trouble with bias. Keynote at NeurIPS.

Thomas A. de Souza. 1977. Regional and communal stereotypes of bombay university students. *Indian Journal of Social Work*, 38(1):37–44.

Tugrulcan Elmas. 2023. The impact of data persistence bias on social media studies. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23. ACM.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Preprint*, arXiv:2309.00770.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Preprint*, arXiv:2108.11896.

M. Hammersley and R. Gomm. 1997. Bias in social research. *Sociological Research Online*, 2(1):7–19.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 501–512, New York, NY, USA. Association for Computing Machinery.

Thongkholal Haokip. 2021. From 'chinky' to 'coronavirus': racism against northeast indians during the covid-19 pandemic. *Asian Ethnicity*, 22(2):353–373.

Rakibul Hasan, Bennett I. Bertenthal, Kurt Hugenberg, and Apu Kapadia. 2021. Your photo is so funny that i don't mind violating your privacy by sharing it: Effects of individual humor styles on online photo-sharing behaviors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *Preprint*, arXiv:2303.11897.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Preprint*, arXiv:2307.06350.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. Seegull: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. 2024. Visage: A global-scale analysis of visual stereotypes in text-to-image generation. *Preprint*, arXiv:2401.06310.

Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. 2020. Mitigating gender bias amplification in distribution by posterior regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2936–2942, Online. Association for Computational Linguistics.

Mascha Kurpicz-Briki. 2020. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, pages 552–561.

Huihan Li, Liwei Jiang, Jena D. Huang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *Preprint*, arXiv:2404.10199.

Brandon C Loudermilk. 2015. Implicit attitudes and the perception of sociolinguistic variation. *Responses to Language Varieties: Variability, Processes and Outcomes*, pages 137–156.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36.

Anne Maass. 1999. Linguistic intergroup bias: Stereotype perpetuation through language. In *Advances in experimental social psychology*, volume 31, pages 79–121. Elsevier.

Aqdas Malik, Cassie Heyman-Schrum, and Aditya Johri. 2019. Use of twitter across educational settings: a review of the literature. *International Journal of Educational Technology in Higher Education*, 16.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially aware bias measurements for hindi language representations. *Preprint*, arXiv:2110.07871.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Duncan McDuie-Ra. 2012. *Northeast migrants in Delhi: Race, refuge and retail*. Amsterdam University Press.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35.

Nusrat Jahan Mim, Dipannita Nandi, Sadaf Sumyia Khan, Arundhuti Dey, and Syed Ishtiaque Ahmed. 2024. In-between visuals and visible: The impacts of text-to-image generative ai tools on digital image-making practices in the global south. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PLOS ONE*, 15:1–26.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020a. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020b. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *Preprint*, arXiv:2307.01952.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural incongruencies in artificial intelligence. *arXiv preprint arXiv:2211.13069*.

Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar.

2020. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI '19, page 450–456, New York, NY, USA. Association for Computing Machinery.

Rida Qadri, Renee Shelby, Cynthia L Bennett, and Emily Denton. 2023. Ai's regimes of representation: A community-centered study of text-to-image models in south asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 506–517.

Ashwin Rajadesingan, Ramaswami Mahalingam, and David Jurgens. 2019. Smart, responsible, and upper caste only: measuring caste attitudes through large-scale analysis of matrimonial profiles. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 393–404.

Nigel Rapport and Joanna Overing. 2002. *Social and cultural anthropology: The key concepts*. Routledge.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Nidhi Sabharwal and Wandana Sonalkar. 2015. Dalit women in india: At the crossroads of gender, class, and caste. *Global justice: Theory, Practice, Rhetoric*, 8.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Preprint*, arXiv:2205.11487.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting unintended social bias in toxic language datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 132–143, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806, Mexico City, Mexico. Association for Computational Linguistics.

Nihar Sahoo, Niteesh Mallela, and Pushpak Bhattacharyya. 2023. With prejudice to none: A few-shot, multilingual transfer learning approach to detect social bias in low resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13316–13330, Toronto, Canada. Association for Computational Linguistics.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. *Preprint*, arXiv:2101.09995.

Manuela Sanguinetti, Gloria Comandini, Elisa Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Haspeede 2 @ evalita2020: Overview of the evalita 2020 hate speech detection task.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Sandhya Singh, Prapti Roy, Nihar Sahoo, Niteesh Mallela, Himanshu Gupta, Pushpak Bhattacharyya, Milind Savagaonkar, Nidhi Sultan, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. 2022. Hollywood identity bias dataset: A context oriented bias analysis of movie dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5274–5285, Marseille, France. European Language Resources Association.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in con-

textualized word representations. *Preprint*, arXiv:1911.01485.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv e-prints*, pages arXiv–2310.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.