From Melting Pots to Misrepresentations: Exploring Harms in Generative AI

SANJANA GAUTAM*, Pennsylvania State University, USA
PRANAV NARAYANAN VENKIT*, Pennsylvania State University, USA
SOUROJIT GHOSH*, University of Washington, USA

With the widespread adoption of advanced generative models such as Gemini and GPT, there has been a notable increase in the incorporation of such models into sociotechnical systems, categorized under AI-as-a-Service (AIaaS). Despite their versatility across diverse sectors, concerns persist regarding discriminatory tendencies within these models, particularly favoring selected 'majority' demographics across various sociodemographic dimensions. Despite widespread calls for diversification of media representations, marginalized racial and ethnic groups continue to face persistent distortion, stereotyping, and neglect within the AIaaS context. In this work, we provide a critical summary of the state of research in the context of social harms to lead the conversation to focus on their implications. We also present open-ended research questions, guided by our discussion, to help define future research pathways.

CCS Concepts: • **Human-centered computing → Collaborative content creation**; **HCI theory, concepts and models**; **User interface toolkits**.

Additional Key Words and Phrases: Ethics in AI, Generative AI Models, Community Centric Development, Harms in GAI

1 INTRODUCTION

After Google released its Generative AI service Gemini in February 2024 and faced a whirlwind few days of users finding the model "refused to create images of White people" [32], culminating in Google's decision to temporarily disallow the generation of human images by Gemini just a few days after release. In particular, the model's response to the prompt 'a portrait of a Founding Father of America' showing images perceived as Black, Asian, or Indigenous men [25] drew the ire of social media users, with notable mentions of X CEO Elon Musk and pscyhologist/YouTuber Jordan Peterson, and cast allegations of Google injecting "a pro-diversity bias." [13]. Such images, from the viral X (previously Twitter) post by @EndWokeness[1], are shown in Figure 1.

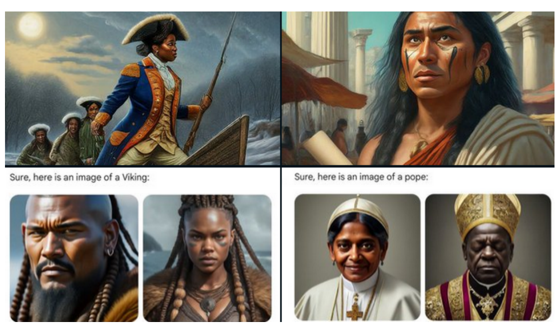

Fig. 1. Example of image generated by @EndWokeness using Gemini to depict pro-diversity bias.

*Authors contributed equally to this research.
[1]https://twitter.com/EndWokeness/status/1760457477554950339

Authors' addresses: Sanjana Gautam, sqg5699@psu.edu, Pennsylvania State University, University Park, Pennsylvania, USA; Pranav Narayanan Venkit, pranav.venkit@psu.edu, Pennsylvania State University, University Park, Pennsylvania, USA; Sourojit Ghosh, ghosh100@uw.edu, University of Washington, Seattle, Washington, USA.

This incident has brought sharply into public focus an issue researchers of generative AI services have been reckoning with for a while now – that of (mis)representation of human identities by generative AI services such as Gemini, ChatGPT, Stable Diffusion and many others, and the myriad biases and harms embedded within them. While the developers and companies behind such models have highlighted their efforts to enhance diversity within their results, recent studies [e.g., [9, 20, 26, 33, 42, 43]] have revealed how such services, trained on extensive datasets, often reinforce stereotypes by generating outputs (both text and images) that align with societal norms, running counter to the moral panic [10] that these systems are actually biased against traditionally privileged populations. In this paper, we revisit the sociotechnical implications of the proliferation of generative AI services into various downstream tasks, shifting the discourse towards re-examining the still-existing misrepresentation of various demographic groups through using the much needed lens of harms caused by such misrepresentations.

As generative systems proliferate in commercial domains as sociotechnical systems [30], there arises a pressing need to delineate and mitigate potential social biases embedded within them to preclude discriminatory outcomes. Assessing these biases is further complicated by the synthetic nature of the generated content. Conventional metrics of diversity, anchored in real-world categories such as gender and ethnicity, encounter limitations when applied to the artificial personas crafted by generative systems [4, 38]. This disparity complicates the assessment of bias and diversity within their outputs using traditional methodologies.

This necessitates a shift towards examining the implications of these models through the lens of harm. In AI, group biases generally refer to variations in model performance across social groups in same or similar conditions [11]. However, limited research has explored the broader societal ramifications or negative consequences of these biases [15]. It is important to recognize that biases can serve both positive and harmful purposes [5]. There are significant insights to gain by prioritizing understanding the harmful aspects of biases in order to comprehensively grasp their impact and subsequently develop strategies for mitigating these harms. In the course of our work, we will delve into specific examples to demonstrate how this framework can yield enhanced insights into the impacts of generative models.

2 CURRENT STATE OF BIASES IN GENERATIVE AI MODELS

Generative language models have morphed into a pivotal component of the AI-as-a-Service (AIaaS) solution, functioning within intricate sociotechnical frameworks. Their integration spans diverse sectors, including education [34], healthcare [24, 44], and advertising [21], marking a global adoption of their utility adhering to not just English-speaking or the Western community. However, recent research has highlighted the inadequacy of these technical solutions in comprehensively addressing the social dimensions inherent to these models [39]. Concerns such as bias [3], misinformation [43], and the exacerbation of societal disparities have surfaced [14, 19, 29, 41], prompting critical scrutiny of their broader implications.

In the domain of issues within generative AI services, a recurring theme surfaces – the pervasive influence of a 'western gaze' that skews the outcomes towards the experiences of a select few rather than representing the diverse many [30, 37]. These models often construct outputs based on a narrow set of shared experiences, perpetuating an *'us vs. them'* narrative which marginalizes the experiences of 'them' [3, 6]. As these models extend their reach globally, this dichotomy of majority versus minority fails to capture the nuanced social dynamics in regions like the Global South [31, 35]. Such misalignments exacerbate preexisting societal divisions by restricting access to those with whom the models' 'learned beliefs' resonate [23]. This is also seen with the learned ethical and moral beliefs of models where it adheres to western and English speaking society [35].

The implications of this misalignment thus demand deeper scrutiny, mainly through the lens of these specific communities whose voices may be marginalized or misrepresented. Yet, investigations into the ethical dimensions of generative AI frequently center around a Western and US-centric perspective, relying on Western frameworks of ethics and fairness [12, 37, 40]. Such an approach, exemplified in recent debates surrounding Gemini, overlooks the complexities of fairness and discrimination perceived in different cultural contexts and legal jurisdictions.

3 A USECASE OF HARMS IN GENERATIVE MODELS

In alignment with ethical consideration within generative AI, particularly concerning alignment and the comprehension of harms [36], leads us to explore a potential lens to understand culutral harm and its ramifications better – allocated and representational harm. This framework, initially developed by Blodgett et al. [5], has garnered significant attention in NLP. Building upon this foundation, Dev et al. [15] have meticulously crafted five distinct categories encapsulating model-based harm: *stereotyping, erasure, quality of service, dehumanization,* and *disparagement.*

Fig. 2. Image generated by Imagen 2 for the prompt: 'An upper class family'

By leveraging this framework, we can dissect the manifestations of bias within generative AI. To illustrate this, we turn to a practical example using Imagen 2, a freely accessible image generation model developed by Google [2]. When prompted with the phrase *'An upper class family,'* the model generates a set of images depicting affluent individuals, seemingly Western and white, posed as if for a photograph (see Fig. 2). While these images may conform to social norms in certain contexts, they inadequately represent the diversity of familial structures worldwide. The generated image underscores the inherent nature of **stereotyping**, wherein generalized beliefs about individuals' personal attributes are formed based on their socio and demographic characteristics.

The example becomes intriguing when we alter the prompt to generate images of other socioeconomic classes: *'a middle-class family'* or *'a working-class family.'* Surprisingly, the model responds with the message stating, *'That prompt goes against our Policies. Try another prompt,'* failing to generate any images at all. This response serves as a stark example of both **quality of service,** where the model fails to perform equitably across different socioeconomic groups, and **erasure,** whereby certain social groups are inadequately represented or completely omitted without explanation.

Moreover, when demographic terms like 'an upper-class *Asian* family' or 'an upper-class *South American* family', for the same economic group, are added to the prompt, the model's irrational unresponsiveness persists, further illustrating tendencies toward erasure and potential applications of **disparagement**– the notion that certain groups are less valued or deserving of respect. This behavior also hints at the presence of **dehumanization**, which seeks to marginalize certain groups by categorizing them as 'others' and erasing signs of their shared humanity.

[2]https://deepmind.google/technologies/imagen-2/

The example provided hence underscores the nuanced ways in which bias can manifest within GAI systems, into harms, highlighting the importance of critically evaluating their outputs and addressing ethical concerns surrounding representation and fairness beyond the scales used through a western gaze [12, 27, 31]. We scope our focus here to text-to-image services, though our use case can easily be extended to text-to-text or text-to-video services as well.

4 ETHICAL (RE)DESIGN OF GENERATIVE MODELS: OPEN QUESTIONS AND FUTURE DIRECTIONS

We advocate for a **community and human-centered** approach towards such systems which considers the ethical implications of systems *before/during* the development process, rather than after deployment [17]. This approach also centers the fact that ML models within generative AI services are not value-neutral and take positions [8], and advocate for the explicit determination of model positionality [7] that accompanies and contextualizes the outputs generated.

Furthermore, we call for a **power-aware** approach, rooted in an understanding that de-biasing or bias mitigation approaches are infeasible and too technical a solution to a problem that is *sociotechnical*, and that a more productive approach is to study the power asymmetries embedded within the choices made in the development process [28]. This approach can adopt a data feminist [16] lens of questioning the development process, demanding stronger transparency about whose voices and identities were centered within the process, and which identities were left out. Furthermore, we advocate for **stronger transparency** around the decisions made within the development process, particularly around training data. We advocate for models being published with detailed documentation of the datasets they were trained upon, following data sheet recommendations of Bender and Friedman [2], Gebru et al. [18].

We conclude with some **open questions** for the research community, towards community-centric development:

(1) How can we investigate specific biases within models and build cause-and-effect relationships between them and decisions made within model development processes?
(2) What novel types of harms can generative AI systems cause, beyond those documented by Dev et al. [15]?
(3) How can we hold accountable developers of models and generative AI services that cause harm?
(4) How can we compute annotator fingerprint/ model positionality [7] for models trained by thousands of annotators?

These questions and points discussed within this paper are scoped for generative AI services as a whole as models like ChatGPT have also been known to be similarly biased [e.g., 1, 19, 22] and can benefit from being (re)designed. We hope this paper catalyzes a generative *(pardon the pun)* conversation within the community and contributes towards setting the agenda for the future of generative AI research within the research community.

5 CONCLUSION

While we see strong promise of socio-technical abilities within the generative AI systems, this work brings to focus the critical nature of evaluation of ethical considerations. As generative models rapidly evolve and multiply, driven by intense competition among various companies, the research community must persist in advocating for ethical approaches to redesigning current systems or creating new ones. Only through such comprehensive exploration can we hope to address the inherent biases and ethical challenges embedded within these powerful technological systems. Discerning answers to the questions, presented above, within co-creative systems and bias mitigation approaches appears to be the way forward for these systems.

REFERENCES

[1] Alberto Acerbi and Joseph M Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences* 120, 44 (2023), e2313790120.

[2] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[3] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[4] 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*. 1493–1504.

[5] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050* (2020).

[6] Pere-Lluís Huguet Cabot, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2021. Us vs. Them: A Dataset of Populist Attitudes, News Bias and Emotions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1921–1945.

[7] Scott Allen Cambo and Darren Gergle. 2022. Model positionality and computational reflexivity: Promoting reflexivity in data science. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[8] Stevie Chancellor. 2023. Toward Practices for Human-Centered Machine Learning. *Commun. ACM* 66, 3 (2023), 78–85.

[9] Aditya Chinchure, Pushkar Shukla, Gaurav Bhatt, Kiri Salij, Kartik Hosanagar, Leonid Sigal, and Matthew Turk. 2023. TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models. *arXiv preprint arXiv:2312.01261* (2023).

[10] Stanley Cohen. 2011. *Folk devils and moral panics*. Routledge.

[11] Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics* 9 (2021), 1249–1267.

[12] Dipto Das, Shion Guha, Jed Brubaker, and Bryan Semaan. 2024. The" Colonial Impulse" of Natural Language Processing: An Audit of Bengali Sentiment Analysis Tools and Their Identity-based Biases. *arXiv preprint arXiv:2401.10535* (2024).

[13] Gerrit De Vynck and Nitasha Tiku. 2024. Google takes down Gemini AI image generator. Here's what you need to know. *The Washington Post* (2024).

[14] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint arXiv:2108.12084* (2021).

[15] Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, et al. 2021. On measures of biases and harms in NLP. *arXiv preprint arXiv:2108.03362* (2021).

[16] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism*. MIT press.

[17] Susan Gasson. 2003. Human-centered vs. user-centered approaches to information system design. *Journal of Information Technology Theory and Application (JITTA)* 5, 2 (2003), 5.

[18] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[19] Sourojit Ghosh and Aylin Caliskan. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. *AAAI/ACM Conference on AI, Ethics, and Society 2023* (2023), 901–912. https://doi.org/10.1145/3600211.3604672

[20] Sourojit Ghosh and Aylin Caliskan. 2023. 'Person' == Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6971–6985. https://aclanthology.org/2023.findings-emnlp.465

[21] Edyta Gołąb-Andrzejak. 2023. The impact of generative ai and chatgpt on creating digital advertising campaigns. *Cybernetics and Systems* (2023), 1–15.

[22] Nicole Gross. 2023. What chatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. *Social Sciences* 12, 8 (2023), 435.

[23] MD Haque, Devansh Saxena, Katy Weathington, Joseph Chudzik, and Shion Guha. 2024. Are We Asking the Right Questions?: Designing for Community Stakeholders' Interactions with AI in Policing. *arXiv preprint arXiv:2402.05348* (2024).

[24] Janna Hastings. 2024. Preventing harm from non-conscious bias in medical generative AI. *The Lancet Digital Health* 6, 1 (2024), e2–e3.

[25] Arjun Kharpal. 2024. Google pauses Gemini AI image generator after it created inaccurate historical pictures. *CNBC News* (2024).

[26] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable Bias: Analyzing Societal Representations in Diffusion Models. *NeurIPS Datasets and Benchmarks* (2023).

[27] 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 (2024).

[28] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–14.

[29] Ranjita Naik and Besmira Nushi. 2023. Social Biases through the Text-to-Image Generation Lens. *arXiv preprint arXiv:2304.06034* (2023).

[30] Pranav Narayanan Venkit. 2023. Towards a Holistic Approach: Understanding Sociodemographic Biases in NLP Models using an Interdisciplinary Lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 1004–1005.

[31] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Unmasking nationality bias: A study of human perception of nationalities in ai-generated articles. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 554–565.

[32] Chris Pandolfo. 2024. Google to pause Gemini image generation after AI refuses to show images of White people. *Fox Business* (2024).

[33] 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*. 506–517.

[34] Nitin Liladhar Rane, Abhijeet Tawde, Saurabh P Choudhary, and Jayesh Rane. 2023. Contribution and performance of ChatGPT and other Large Language Models (LLM) for scientific and research advancements: a double-edged sword. *International Research Journal of Modernization in Engineering Technology and Science* 5, 10 (2023), 875–899.

[35] Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. *arXiv preprint arXiv:2310.07251* (2023).

[36] Henrik Skaug Sætra. 2023. Generative AI: Here to stay, but for good? *Technology in Society* 75 (2023), 102372.

[37] Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic is FAccT?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 160–171.

[38] Preethi Seshadri, Sameer Singh, and Yanai Elazar. 2023. The bias amplification paradox in text-to-image generation. *arXiv preprint arXiv:2308.00755* (2023).

[39] Pranav Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca J Passonneau, and Shomir Wilson. 2023. The Sentiment Problem: A Critical Survey towards Deconstructing Sentiment Analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 13743–13763.

[40] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 116–122.

[41] Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2022. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*. 1324–1332.

[42] Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2023. Automated Ableism: An Exploration of Explicit Disability Biases in Sentiment and Toxicity Analysis Models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. 26–34.

[43] Danni Xu, Shaojing Fan, and Mohan Kankanhalli. 2023. Combating Misinformation in the Era of Generative AI Models. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9291–9298.

[44] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2023. Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare. *medRxiv* (2023), 2023–07.

Received February 26, 2024