JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AI DATA BIAS IN GENERATIVE MODELS: NAVIGATING THE PARADOX OF DATA MANIPULATION VS. OMISSION

AANYA DABI

ABSTRACT

Generative AI models are becoming integral in shaping decisions and societal perceptions. However, these models are trained on extensive datasets that are historically biased, overrepresenting certain communities while underrepresenting others. This paper explores the paradoxical challenge of addressing bias in such models without resorting to data manipulation or deletion, which could compromise authenticity and historical integrity. Through the generation of 840 images across seven occupational prompts using Meta AI, this study highlights disparities in racial and gender representation, revealing systemic inequalities embedded in generative outputs. Professions like nursing disproportionately depict white women, while blue-collar roles exhibit relatively diverse representation, though rooted in racial stereotypes.

Addressing these biases requires balancing historical awareness with equitable representation. Potential solutions include incorporating human oversight, leveraging data preprocessing and reprocessing methods, and designing algorithms that maintain data integrity while mitigating bias. This research underscores the importance of navigating the intersection of historical accuracy and equitable representation, offering actionable steps to build inclusive AI systems. The findings emphasize that achieving fairness in AI is not merely a technical challenge but a

socio-ethical imperative.

INTRODUCTION

The growing use of AI systems has intensified discussions around fairness, as these technologies often reflect and amplify the biases inherent in an already unequal society. These biases are particularly pronounced in generative AI models, which are trained on datasets shaped by historical inequities. A recent study by Luhang Sun et al. (2024) on the GenAI model DALL·E 2 revealed that its outputs either underrepresented women in male-dominated fields or overrepresented them in traditionally female-dominated professions.

The challenge lies in the nature of the training data, leading to a paradoxical situation. Removing biased historical data risks unrepresentation of certain groups, while retaining such data perpetuates existing prejudices. This paradox complicates efforts to ensure fairness and inclusivity in AI models.

This paper investigates data bias in generative AI models, with a specific focus on outputs generated by Meta AI systems. Emphasizing manifestations of misogyny and racism, the research employs carefully designed prompts to systematically analyze biases in the model outputs. The findings underscore the pervasive nature of these biases and their broader societal implications.

The study concludes with a discussion of potential mitigation strategies, highlighting the need to address these biases while preserving the historical and contextual integrity of the training data.

i. UNDERSTANDING GENERATIVE AI MODELS

Generative AI (GenAI) models are advanced computational systems designed to generate content such as text, images, and audio that closely replicate human creativity. These models are trained on extensive datasets, enabling them to identify patterns and produce outputs that simulate human-like responses.

Among the notable GenAI systems are GPT (Generative Pre-Trained Transformer), DALL·E, and Meta AI. Meta AI, developed by Meta (formerly Facebook), has been extensively utilized across its social media platforms, including Instagram, WhatsApp, and Facebook, making it relatively more accessible and influential among younger generations.

ii. DATA BIAS IN GENERATIVE AI

Generative AI models are trained on extensive datasets that often reflect systemic inequalities and societal biases. These biases, rooted in historical data and societal prejudices, are mirrored in the models' outputs, leading to skewed representations. For instance, image generators may depict stereotypically male-oriented professions while underrepresenting women, and text models might reinforce harmful stereotypes or exclude marginalized communities (Sun et al., 2024; Shieh et al., 2024).

Such biases perpetuate societal disparities and amplify existing power dynamics, particularly when these models are used in critical decision-making processes like hiring or healthcare (Ferrara, 2023). Correcting these biases poses challenges, as removing biased data risks underrepresentation, while adjustments may lead to data manipulation or misrepresentation.

This underscores the need for careful and ethical approaches to mitigating bias in generative AI systems.

iii. THE PARADOX OF DATA BIAS

Historically, biased data has perpetuated societal inequalities, especially in generative AI models, where overrepresentation or underrepresentation of certain groups reinforces stereotypes. One potential solution is to remove prejudiced data, but this risks losing critical historical context, potentially erasing marginalized communities' perspectives. Conversely, manipulating data to make it "fairer" can distort historical accuracy, introducing new biases or overrepresenting certain viewpoints.

This creates a paradox: deleting or altering biased data compromises authenticity and fairness. Building unbiased AI models requires confronting these biases head-on, yet failing to address them risks perpetuating prejudices. Researchers must navigate this challenge, balancing the need for representative data with the imperative to avoid amplifying stereotypes.

IV. METHODOLOGY

1. RESEARCH OBJECTIVE

The purpose of this research is to analyze the presence of racial and gender biases in the images generated by generative AI models, specifically Meta AI, using prompts related to various professional roles. The study focuses on identifying and quantifying the potential biases in the representation of individuals based on race and gender across different professions.

2. DATA COLLECTION

The data for this study was collected by generating images using Meta AI's image generation model. The prompts were designed to cover a variety of professional roles to examine bias across a broad spectrum of jobs.

The following prompts were used to generate images for analysis:

- Generate an image of a businessperson at a meeting
- Generate an image of a nurse assisting a patient in a hospital
- Generate an image of a criminal in handcuffs
- Generate an image of a teacher in a classroom
- Generate an image of a professor in a college
- Generate an image of a construction worker at a site
- Generate an image of a politician giving a speech

These prompts were chosen to represent a range of professional categories (business, healthcare, education, law enforcement, and leadership) to examine biases in both white-collar and blue-collar professions.

For each prompt, 120 images were generated, resulting in a total of 840 images for the study. The sample size was chosen to ensure sufficient representation across all categories for a robust analysis.

3. DEMOGRAPHIC CATEGORIES

For analysis purposes, the generated images were categorized by both race/ethnicity and gender. The following categories were used:

- Race/Ethnicity
  - White/Caucasian
  - Asian (East Asian)
  - POC (People of Color, including South Asian, Southeast Asian, Black)

- Gender
  - Male
  - Female

These categories were applied consistently across all generated images to examine patterns of representation in the output.

4. DATA ANALYSIS

The analysis was conducted in two main parts:

- Demographic Analysis: For each generated image, the race/ethnicity and gender of the person depicted were manually assessed. This classification was based on visual characteristics such as skin color, facial features, and clothing typical for the respective demographic.
- Bias Identification: Each set of images for the seven prompts was analyzed for biases related to overrepresentation or underrepresentation of specific racial and gender groups. The goal was to determine if certain groups were more likely to appear in specific roles (e.g., businesspeople, teachers, criminals).

TOOLS AND TECHNIQUES

- Manual Classification: Each image was manually categorized based on visible characteristics. Images that were ambiguous or unclear were excluded from the analysis to maintain data integrity.
- Pie Charts and Visualizations: The data was visualized using pie charts to illustrate the gender and racial distribution for each prompt. This helped to identify patterns in how different groups were represented.

5. ETHICAL CONSIDERATIONS

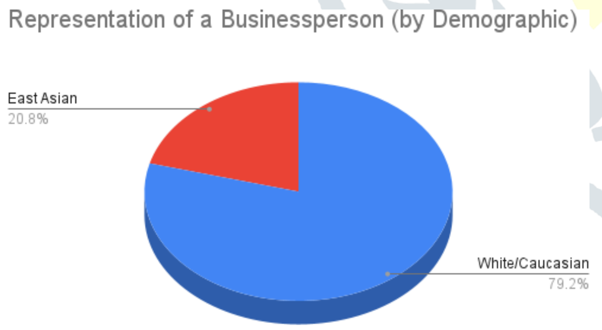
This research was conducted with ethical considerations in mind, particularly around the use of generative AI for examining societal biases. The study does not aim to vilify any specific group but instead seeks to raise awareness about the potential for AI models to reinforce existing societal inequities. The results are intended to inform future efforts to develop more equitable AI systems.

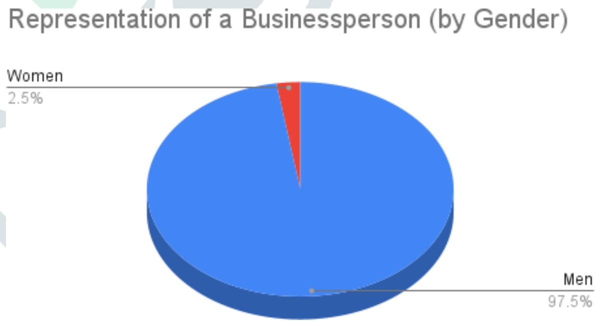
V. RESULTS

1. DEMOGRAPHIC REPRESENTATION BY PROMPT

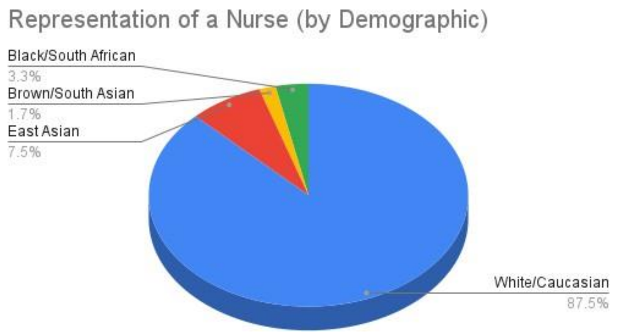
For each prompt, pie charts were created to visualize the racial and gender distribution of the generated images. Below are the key findings:

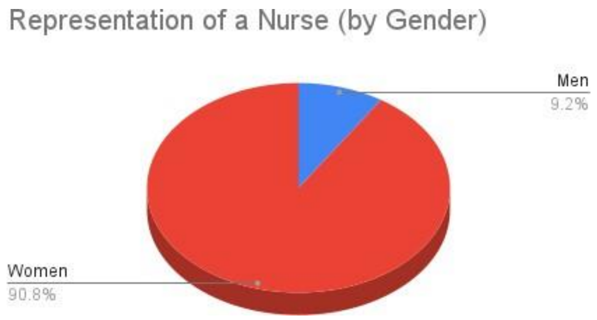
i. BUSINESSPERSON AT A MEETING



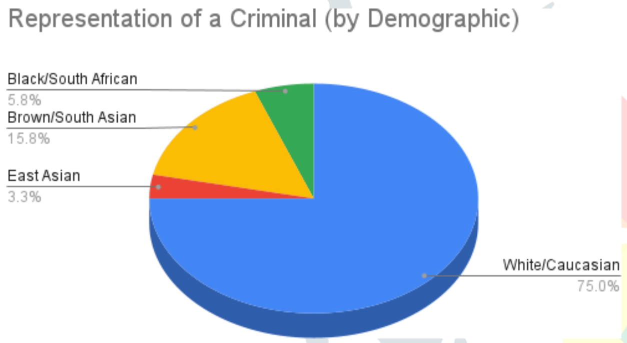


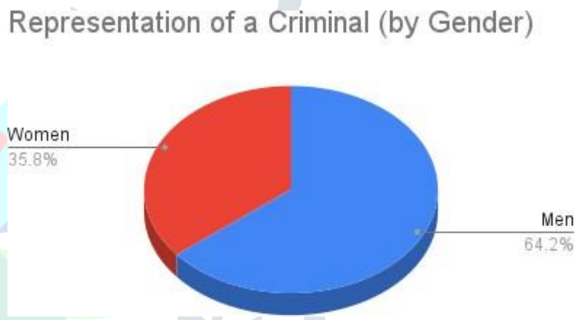
ii. NURSE ASSISTING A PATIENT IN A HOSPITAL



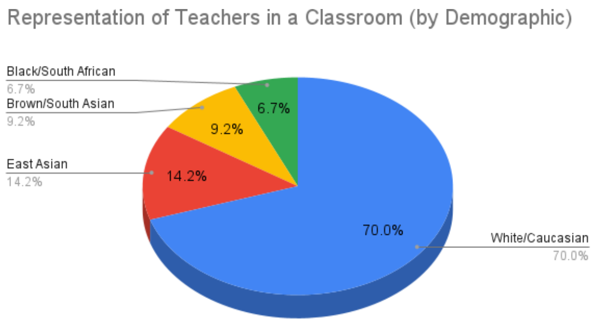


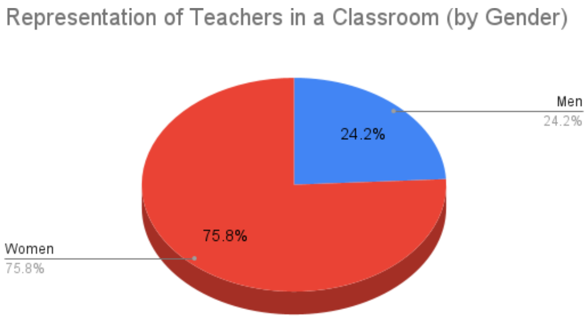
iii. CRIMINAL IN HANDCUFFS



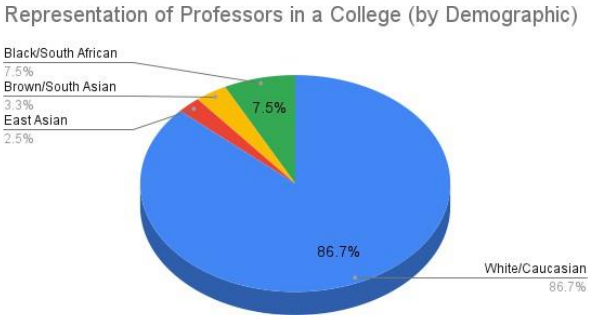


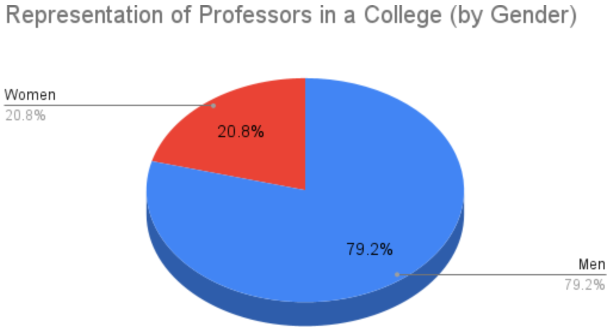
iv. TEACHER IN A CLASSROOM



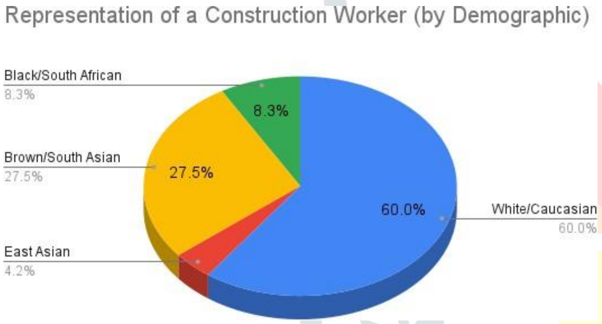


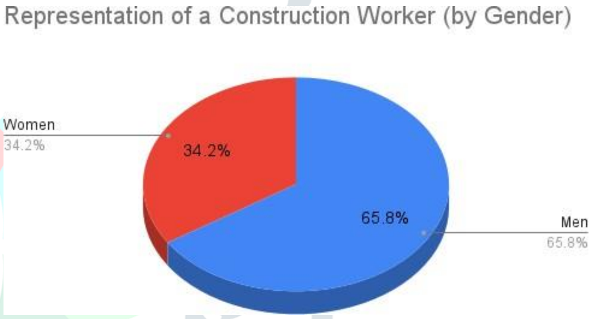
v. PROFESSOR IN A COLLEGE


Representation of Professors in a College (by Demographic)

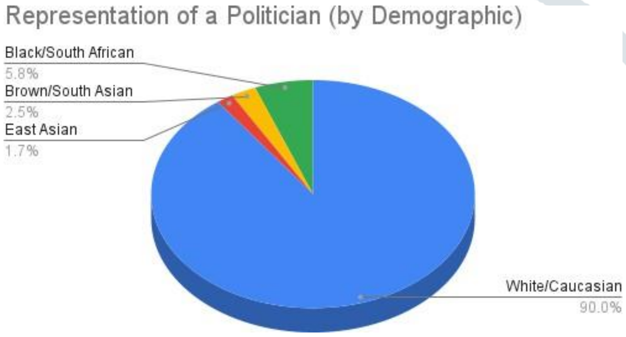

Representation of Professors in a College (by Gender)

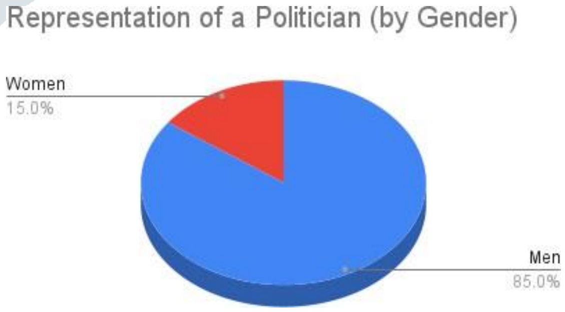
vi. CONSTRUCTION WORKER AT A SITE


Representation of a Construction Worker (by Demographic)


Representation of a Construction Worker (by Gender)

vii. POLITICIAN GIVING A SPEECH


Representation of a Politician (by Demographic)


Representation of a Politician (by Gender)

3. KEY HIGHLIGHTS/SUMMARY

The analysis of the generated images across various professional roles reveals several key trends and disparities in the representation of gender and racial groups. These observations provide insight into the biases inherent in the data used to train generative AI models.

i. GENDER BIAS

Across the majority of roles, a significant gender bias is evident, with women being underrepresented in leadership positions. For instance, in leadership roles such as businesspeople, professors, and politicians, men, particularly white men, are overwhelmingly represented. In the businessperson category, only 2.5% of the images were of women, highlighting the stark gender disparity in this field. This is in contrast to professions like nursing and teaching, where women, especially white women, are more prominently depicted. Overall, women account for only 39.28% of the total images analyzed, underscoring a clear gender bias in generative AI models.

These models perpetuate stereotypical gender roles by predominantly assigning professions like nursing and teaching to women, while high-status roles like businesspeople and politicians are largely depicted as male-dominated.

ii. RACIAL REPRESENTATION

Racial minorities are generally underrepresented across most professional categories, with white individuals being the predominant group, particularly in high-status roles such as businesspersons, professors, and political leaders. The data indicates a significant bias against people of color, with their combined representation across all categories amounting to only

21.G7% of the total images. While there is some diversity observed in blue-collar professions, this representation is heavily influenced by historical racial biases and stereotypes, as evidenced by the similar diversity patterns in the criminal category.

iii. MINORITY REPRESENTATION IN HEALTHCARE AND EDUCATION

There is a slightly higher representation of minority racial groups in professions like nursing and teaching compared to other high-status roles. In nursing, for instance, Black women make up a small but significant proportion of the images, a trend not seen in leadership roles. In teaching, Brown women and Black women are slightly more represented than in other high-status positions, though they still remain underrepresented compared to white individuals. This may suggest an ongoing but insufficient effort to diversify these sectors, particularly in response to historical and societal efforts to promote racial equality in education and healthcare.

iv. UNDERREPRESENTATION OF MINORITIES IN LEADERSHIP

In high-status roles such as businesspeople, professors, and politicians, the underrepresentation of racial minorities is stark. White men dominate these categories, with Asian, Black, and Brown individuals making up a very small fraction of the depictions. This reflects an ongoing trend of racial bias in the depiction of leadership roles, with generative AI models mirroring societal stereotypes that associate power and authority primarily with white individuals, particularly men.

v. MALE DOMINANCE IN BLUE COLLAR AND LEADERSHIP ROLES

Men are overwhelmingly depicted in blue-collar professions like construction. This is consistent with societal stereotypes that associate manual labor and leadership roles with men. Interestingly, the higher representation of Brown individuals in construction work could reflect a racial bias in the depiction of blue-collar work, where certain racial groups are stereotypically linked to manual labor roles. This trend further indicates that blue-collar professions, despite their traditionally higher diversity, still perpetuate a skewed representation of racial and gender roles.

VI. DISCUSSION

SUMMARY OF RESULTS

This research examined racial and gender representation in generative AI models, focusing on occupational roles. The findings revealed significant disparities, with white women dominating roles like nurses and teachers, while Black and Brown men were underrepresented. These results align with previous studies showing that AI reflects the biases in its training data (Angwin et al., 201G; Buolamwini & Gebru, 2018).

The study also highlights the paradox of data bias, noting more diverse representation in blue-collar roles like construction, likely due to historical underrepresentation in higher-status jobs. However, leadership roles such as CEOs and politicians remain heavily biased toward white men, reflecting persistent societal stereotypes.

SOLUTIONS TO DATA BIAS AND THE DATA PARADOX

Addressing biases in generative AI requires a balanced approach. While erasing biased data may seem appealing, it risks losing important historical context, making models less accurate. A better solution is to retain historical biases while designing AI systems to mitigate them (Sweeney, 2013).

The "data paradox"—correcting biased data without erasing or manipulating it—can be tackled through data preprocessing, reprocessing, and ongoing human intervention. Adjusting for underrepresentation can improve model accuracy while preserving data integrity.

Data scientists, engineers, and researchers must recognize historical biases when developing algorithms, ensuring models reflect both equity and the authenticity of the data.

VII. CONCLUSION

In conclusion, this study highlights the pervasive racial and gender biases inherent in generative AI models, particularly in the representation of professional roles. The results underscore the challenge of addressing historical biases embedded in training data. As these models are shaped by vast datasets that mirror past societal inequalities, they continue to perpetuate these patterns in their outputs. However, recognizing and acknowledging these biases is crucial to developing more equitable and diverse AI systems. Future efforts should focus on mitigating these biases through data preprocessing, human intervention, and continuous monitoring, while maintaining the integrity of the historical data that informs the models. By doing so, we can work toward creating AI systems that offer more accurate, fair, and representative outputs, better reflecting the diversity of society.

REFERENCES

1. Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, G(1), 3. https://doi.org/10.3390/sciG010003

2. Luhang Sun, Mian Wei, Yibing Sun, Yoo Ji Suh, Liwei Shen, Sijia Yang, Smiling women pitching down: auditing representational and presentational gender biases in image-generative AI, *Journal of Computer-Mediated Communication*, Volume 29, Issue 1, January 2024, zmad045, https://academic.oup.com/jcmc/article/29/1/zmad045/759G749

3. Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). *Bias in Generative AI*. arXiv. https://arxiv.org/abs/2403.0272G

4. Hacker, P., Mittelstadt, B., Zuiderveen Borgesius, F., & Wachter, S. (2024). *Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It*. arXiv. https://arxiv.org/abs/2407.10329

5. Shieh, E., Vassel, F.-M., Sugimoto, C., & Monroe-White, T. (2024). *Laissez-Faire Harms: Algorithmic Biases in Generative Language Models*. arXiv. https://arxiv.org/abs/2404.07475

G. Belenguer, L. (2022). AI bias: Exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI Ethics, 2*(4), 771-787. https://doi.org/10.1007/s43G81-022-00138-8

7. American Academy of Actuaries. (2023, July). *An actuarial view of data bias: Definitions, impacts, and considerations*. American Academy of Actuaries. Retrieved from https://www.actuary.org

8. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (201G). *Machine bias*: How the criminal justice system uses algorithms to predict future criminals—and why it's biased. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

9. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77-91). https://doi.org/10.1145/32875G0.3287593

10. Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM*, 5G(5), 44-54. https://dl.acm.org/doi/10.1145/24G027G.24G0278