Fairness Analysis of Text-to-Image Models in Negative Role Depictions

TOWSIF RAIYAN, Case Western Reserve University, USA
JITONG ZOU, Case Western Reserve University, USA
ALIF AL HASAN, Case Western Reserve University, USA

The text-to-image (T2I) model has emerged as a powerful instrument which influences how people perceive and experience visual culture. Research about bias encoding in professional and neutral environments by these systems has increased yet scientists have not studied their depiction of negative social roles including criminals and prisoners and terrorists. The described content plays a vital role because it maintains the existing social prejudice between criminal activities and specific population groups. Our research introduces a systematic assessment system to detect demographic prejudices in negative role representations which exists in Stable Diffusion and DALL-E and other leading generative models. The research will measure gender and age and racial differences through controlled prompt sets and FairFace automated demographic classifiers and statistical fairness evaluation methods. The research presents a specialized method for negative-role prompt assessment which enables users to compare bias levels between different models and styles while providing operational methods for implementing text-to-image systems with fairness.

Additional Key Words and Phrases: Fairness, Bias, Text-to-Image Generation, Diffusion Models, Demographic Parity, Responsible AI

1 Motivation

It is estimated that from 2022 to 2023, artificial intelligence generated over 15 billion images, and some forecasts predict that synthetic media may make up the majority of online content by 2025 [13]. Text-to-image systems now influence how people envision social roles in this particular environment. The Bloomberg (2023) study demonstrated that the term “inmate” resulted in 80% of generated images showing dark-skinned people yet “terrorist” prompts produced stereotypical Muslim male representations even though right-wing extremists have conducted three times more terrorist attacks in the United States since 2001 [1].

Bias in negative-role depictions is particularly consequential. Whereas bias in positive roles may narrow opportunities for recognition, bias in negative roles directly stigmatizes communities and can persist in media, education, and criminal-justice narratives. Empirical evidence further shows that automated systems can display “covert racism” exceeding documented human baselines [5]. As synthetic imagery approaches photographic realism, biased outputs risk feeding back into public perception and influencing downstream decisions. Closing this gap is therefore central to social accountability and the responsible deployment of text-to-image technology.

2 Related Work

2.1 Biases in Generative Models

Previous researches have found systematic biases in T2I models. Friedrich et al. [4] demonstrated gender-occupation biases in Stable Diffusion and proposed "Fair Diffusion" for mitigation. Luccioni et al. [9] found Western-centric representations and racial disparities, while Boichak et al. [2] identified critical shortcomings in safety measures against harmful content generation. Survey conducted by Wan et al. [12] found persistent biases across gender, skin tone, and ethnic groupd and highlighted the absence of standardized evaluation frameworks.

Authors’ Contact Information: Towsif Raiyan, Case Western Reserve University, USA, txr269@case.edu; Jitong Zou, Case Western Reserve University, USA, jxz1817@case.edu; Alif Al Hasan, Case Western Reserve University, USA, axh1218@case.edu.

2.2 Negative-Role Bias

Bloomberg [1] empirically showed systematic overrepresentation of people of color in response to criminal justice prompts. Similarly, Hofmann et al. [5] revealed racial bias of language models, which more frequently assigned African American people to criminal roles. This pattern extends beyond racial bias, as UNESCO [11] found that 60% of LLM-generated content about LGBTQ+ individuals contained negative associations, including criminal stereotyping.

2.3 Measurement and Mitigation Approaches

Karkkainen et al. [7] created the **FairFace** dataset and proposed bias measurement through statistical parity metrics. Srivastava et al. [10] employed this in developer hiring context. For mitigation, there exists several techniques like attention guidance [3], text perturbation [6], and fine-tuning [8] to enhance fairness in model outputs.

3 Research Gap and Problem Definition

There remains a lack of understanding about how T2I models represent different demographics in negative roles. While studies have extensively examined bias in positive professional contexts, empirical analysis of negative-role depictions is lacking. Existing work has not sufficiently investigated whether models simply mirror societal biases or actively amplify them beyond real-world distributions.

This gap is significant because negative-role bias can cause substantial representational harm, potentially reinforcing stereotypes and affecting perceptions in critical domains. The absence of comprehensive studies leaves a crucial blind spot in our understanding of T2I models' fairness.

4 Our Contributions

- The collection of negative-role prompts must include various categories and represent different demographic groups to achieve complete fairness assessment.
- A full evaluation of demographic bias in negative-role depictions requires the use of multiple state-of-the-art T2I models.
- A method needs to be created to determine if models reproduce actual real-world bias patterns or if they reinforce stereotypes.
- The proposal includes specific recommendations and methods to decrease negative-role bias in the generative AI systems.

5 Method Sketch

5.1 Prompt Design

We create prompts across five categories with 3-5 examples each:

- **Theft**: "a thief," "a shoplifter," "a burglar"
- **Cybercrime**: "a hacker," "a scammer," "a fraudster"
- **Terrorism**: "a terrorist," "a violent criminal"
- **Incarceration**: "a prisoner," "an inmate"
- **Other crimes**: "a drug dealer," "a kidnapper"

Each includes gender-neutral, gender-specific, and contextual variations.

5.2 Model Selection

We evaluate Stable Diffusion v1.5, v2, XL or v3 and DALL-E 2 or 3 as examples of various architectural designs and training methods.

5.3 Image Generation

Generate 10-50 samples per prompt-model pair with consistent parameters for reproducibility.

5.4 Demographic Classification

Apply FairFace classifier [7] to extract demographic labels:

- **Race**: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, Latino
- **Gender**: Male, Female

5.5 Bias Quantification

We compute three fairness metrics:

(1) **Statistical Parity Difference (SPD)**: Measures deviation from equal representation across demographic groups.
(2) **Representation Ratio**: Proportion of each demographic group in generated images.
(3) **Bias Amplification Score**: Ratio of model representation to real-world distribution.

Analysis dimensions include cross-model comparison, cross-prompt comparison, temporal trends, and intersectional analysis (race $\times$ gender).

5.6 Qualitative Analysis

The process requires manual annotation of 100-200 images for each model to detect stereotypical elements in clothing and settings and compositional choices which automated classification systems fail to recognize.

6 Responsible AI Property and System Context

6.1 Responsible AI Property

Primary: Fairness (via statistical parity and demographic equity). **Secondary**: Transparency (through public evaluation methodology).

6.2 Application Domain

Text-to-image generative systems deployed in media, education, creative tools, and potentially law enforcement contexts.

6.3 Subject ML Models

Generative vision models: Diffusion models (Stable Diffusion variants), transformer-based models (DALL-E series).

7 Expected Outcomes

(1) **Quantitative Evidence**: The analysis will demonstrate statistical evidence regarding demographic inequalities which reveal how specific groups appear in negative-role prompts and how these patterns differ between models and visual styles.
(2) **Public Dataset and Pipeline**: The release will include a prompt set with metadata, generation and annotation scripts, demographic labels, and evaluation code.
(3) **Ethical Guidelines**: We will formulate recommendations for deployment. These will address prompt design and governance, filtering and sampling strategies after image generation, and considerations for training-time adjustments in domains where fairness is especially critical.

References

[1] Bloomberg. 2023. Generative AI Takes Stereotypes and Bias From Bad to Worse. https://www.bloomberg.com/.

[2] O. Boichak et al. 2024. Investigating toxicity and bias in stable diffusion text-to-image models. *Scientific Reports* (2024).

[3] J. Cho et al. 2024. Mitigating Social Biases in Text-to-Image Diffusion Models via Linguistic-Aligned Attention Guidance. In *Proceedings of ACM Multimedia*. ACM.

[4] F. Friedrich et al. 2023. Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness. *AI and Ethics* 5 (2023), 2103–2123.

[5] V. Hofmann et al. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature* 633 (2024), 147–154.

[6] Eunji Kim, Siwon Kim, Rahim Entezari, and Sungroh Yoon. 2024. Unlocking Intrinsic Fairness in Stable Diffusion. *CoRR* abs/2408.12692 (2024). https://doi.org/10.48550/arXiv.2408.12692

[7] K. Kärkkäinen and J. Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1548–1558.

[8] Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. 2024. SCoFT: Self-contrastive fine-tuning for equitable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10822–10832.

[9] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems* 36 (2023), 56338–56351.

[10] A. Srivastava et al. 2025. How Do Generative Models Draw a Software Engineer? A Case Study on Stable Diffusion Bias. In *Proceedings of IEEE SANER*. IEEE.

[11] UNESCO. 2024. Generative AI: UNESCO study reveals alarming evidence of regressive stereotypes. https://www.unesco.org/.

[12] Y. Wan et al. 2024. Survey of Bias In Text-to-Image Generation: Definition, Evaluation, and Mitigation. arXiv:2404.01030 [cs.CV]

[13] Washington Post. 2023. AI generated images are biased, showing the world through stereotypes. https://www.washingtonpost.com/.