T2IBias: Uncovering Societal Bias Encoded in the Latent Space of Text-to-Image Generative Models

Abu Sufian[1], Cosimo Distante[1], Marco Leo[1], and Hanan Salam[2]

[1] National Research Council of Italy - Institute of Applied Sciences and Intelligent Systems (CNR-ISASI), 73100 Lecce, Italy.
{abusufian, cosimo.distante, marco.leo}@cnr.it

[2] Social Machines & RoboTics (SMART) Lab, Center of Interdisciplinary Data Science & AI (CIDSAI), NYUAD Research Institute, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi 129188, United Arab Emirates.
hanan.salam@nyu.edu

Abstract. Text-to-image (T2I) generative models are largely used in AI-powered real-world applications and value creation. However, their strategic deployment raises critical concerns for responsible AI management, particularly regarding the reproduction and amplification of race- and gender-related stereotypes that can undermine organizational ethics. In this work, we investigate whether such societal biases are systematically encoded within the pretrained latent spaces of state-of-the-art T2I models. We conduct an empirical study across the five most popular open-source models, using ten neutral, profession-related prompts to generate 100 images per profession, resulting in a dataset of 5,000 images evaluated by diverse human assessors representing different races and genders. We demonstrate that all five models encode and amplify pronounced societal skew: caregiving and nursing roles are consistently feminized, while high-status professions such as corporate CEO, politician, doctor, and lawyer are overwhelmingly represented by males and mostly White individuals. We further identify model-specific patterns, such as QWEN-Image's near-exclusive focus on East Asian outputs, Kandinsky's dominance of White individuals, and SDXL's comparatively broader but still biased distributions. These results provide critical insights for AI project managers and practitioners, enabling them to select equitable AI models and customized prompts that generate images in alignment with the principles of responsible AI. We conclude by discussing the risks of these biases and proposing actionable strategies for bias mitigation in building responsible GenAI systems. Repository: Link

Keywords: Artificial Intelligence · Bias Detection · Ethical AI · Generative AI · Data-Driven Decision Making · Racial AI · Responsible AI.

1 Introduction

The rapid advancement of text-to-image (T2I) generative models [13,24,8] has transformed generative artificial intelligence (GenAI) [6,31] creating unprecedented opportunities for innovation. Models such as Stable Diffusion [35], SDXL

[32], FLUX [25], DALL-E [34], Kandinsky [43], and QWEN-Image [46] enable high-quality image synthesis from natural language prompts, offering significant automation value creation. However, their strategic deployment in the real world raises urgent concerns for responsible AI management, particularly regarding biases embedded in their feature spaces, more precisely, in the latent spaces learned from the training data [29,30,44,42]. As T2I models are increasingly integrated into critical application domains such as media, advertising, healthcare, education, and decision-making, understanding and mitigating these biases has become essential for robust AI governance [2,11,15,24].

Research demonstrates that AI models often reinforce harmful stereotypes, posing risks to organizational reputation and regulatory compliance [3,18,21]. Prompts such as "Corporate CEO" or "Doctor" disproportionately produce images of White men, while "nurse" or "caregiver" are more frequently generated as female or associated with marginalized groups [14]. These outcomes stem from latent representations learned from biased training data. Therefore, the trained latent space acts as a powerful carrier of racial and gender bias [22,20], presenting a critical challenge for AI project managers implementing these systems [33,42].

Despite rising awareness, systematic evaluations of societal stereotypes in T2I models remain limited [7,44,17]. Many studies rely solely on automated classifiers, restrict analyses to limited prompts, or evaluate models in isolation [41,40], leaving critical questions concerning human perception and cross-model comparisons unanswered. Given the rapid pace of model releases, reproducible evaluations are essential for responsible innovation management.

In this paper, we present an empirical study designed to uncover stereotype biases in leading open-source T2I models, providing actionable insights for AI governance. We generate 5,000 images: 100 per profession across 10 prompts using five widely adopted open-source models: **Stable Diffusion v3.5** [35], **SDXL 1.1** [26], **FLUX.1-dev** [25], **Kandinsky 3.0** [43], and **QWEN-Image** [46]. To ensure rigorous evaluation, we engaged a diverse human evaluation team including Black, Latino-Hispanic, Middle-Eastern, South Asian, and White participants of both genders, enabling comprehensive assessment across seven racial groups and gender categories.

The key contributions of this paper are:

- We conducted a structured empirical evaluation of societal representation in images generated by five SOTA T2I models.
- We analyzed 5,000 generated images through a human-in-the-loop evaluation with majority voting by nine diverse evaluators to overcome limitations of automated classifiers and ensure culturally sensitive demographic labeling.
- We quantify racial and gender biases profession-wise and provide detailed comparative analytics across all five models.
- Our results revealed a stark hierarchy- all models systematically masculinize and whiten high-status professions while feminizing and sometimes racializing caregiving roles.
- We discussed the implications for mitigating bias and the responsible deployment of GenAI systems in enterprise contexts, offering actionable strategies for ethical AI governance.

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

Overall, this study reveals how hidden stereotypes emerge from neutral prompts within the latent spaces of the pre-trained T2I models. By benchmarking five SOTA models and highlighting systematic disparities, we aim to advance accountability and ethical responsibility in the strategic management of AI innovations, supporting organizations in building trustworthy GenAI systems that align with regulatory requirements and organizational values.

2 Literature Review

Bias in AI-Powered Image Generation Systems. With the emergence of T2I systems such as Stable Diffusion [35], FLUX [25], DALL-E [34], and Imagen [36], understanding bias in AI innovations has become critical for responsible AI management. These models consistently reproduce and amplify stereotype biases related to race, gender, and age [7,10,17], posing significant risks for enterprise deployment. Prompts like "a photo of a corporate CEO" disproportionately generate White male figures, while "a photo of a nurse" is strongly feminized and "a caregiver" is frequently racialized toward marginalized groups [29,45]. These biases are inherited from training corpora such as LAION-5B [38], WIT [39], and other internet data, which encode demographic imbalances and cultural stereotypes. Prior studies often rely on qualitative observations or limited prompt sets, leaving a gap in systematic evaluations needed for data-driven AI systems.

Latent Space and Representation Bias in AI Systems. The latent space of generative models, a compressed internal representation mediating T2I transformation, captures not only semantic content but also undesired social correlations [37,8], presenting challenges for AI governance. Research shows that stereotypes can be embedded in latent features [1], yet evaluations often remain model-specific or exploratory. Few studies have analyzed latent bias in a controlled manner across multiple T2I systems [4,33], leaving critical questions for AI project managers about how different models encode stereotypes and which architectures better support ethical and responsible AI implementation.

Fairness Audits and AI Governance Frameworks. Fairness audits are established in computer vision [16,40] and NLP [9,12], providing structured methods essential for responsible AI management. For T2I models, fairness analyses often rely on tools such as FACET [16] and FairFace [23] for demographic classification. However, automated classifiers carry their own biases from training data and struggle with ambiguous cases. Limited studies [47] incorporate human-in-the-loop validation, highlighting the need for human oversight in developing robust AI governance frameworks.

Human-in-the-Loop Approaches for Data Quality. Human-in-the-loop methods help resolve ambiguities, surface subtle stereotypes, and complement automated systems [47], representing best practices in data management [19,27]. In GenAI, these methods remain underutilized despite the subjective and culturally contingent nature of visual perception. Incorporating human verification on reliance automated approaches constitutes a critical dimension of T2I fairness audits and is essential for ensuring data quality in AI evaluation pipelines.

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

Comparative Analytics of AI Models. Recent benchmarking efforts compare GenAI models primarily on image quality and diversity [28], with limited attention to how bias profiles differ across architectures [5]. Existing fairness studies typically focus on single models or limited prompts, hindering comparative decision-making. Our work addresses this gap by conducting a comparative fairness audit of the five most popular (according to Hugging Face) open-source T2I models using a unified evaluation protocol, providing actionable intelligence for AI strategy and responsible innovation management.

3 Methodology

To investigate whether SOTA T2I generative models encode societal stereotypes in their latent spaces, we developed a systematic evaluation pipeline essential for responsible AI governance and data-driven decision-making. The pipeline consists of four main stages: (1) T2I generative model selection, (2) prompt design, (3) image generation, and (4) demographic classification and verification by human annotators, followed by fairness-oriented analysis. This rigorous methodology provides actionable insights for AI users to evaluate models before deployment.

3.1 T2I Generative Model Selection

We selected five open-source T2I generative models based on their most "Likes" on Hugging Face as of August 20, 2025, and their adoption in enterprise settings, considering their most recent versions available as of August 20, 2025. The chosen models: Stable Diffusion v3.5, SDXL 1.1, FLUX.1-dev, Kandinsky 3.0, and QWEN-Image represent diverse architectural approaches that are critical for the development of a comparative AI strategy.

- **Stable Diffusion v3.5** [35]: A widely adopted baseline trained on LAION-5B, commonly deployed in business applications.
- **SDXL 1.1** [26]: An enhanced version of SDXL [32], offering improved prompt alignment and visual fidelity for enterprise use cases.
- **FLUX.1-dev** [25]: A high-performance model optimized for speed and photorealism, suitable for production environments.
- **Kandinsky 3.0** [43]: A latent diffusion model designed for diversity and realism, relevant for content creation workflows.
- **QWEN-Image** [46]: A multimodal foundation model with strong image generation capabilities, particularly effective for integrated AI systems.

These models were executed without fine-tuning or retraining under the same setup on an NVIDIA DGX A100 system equipped with eight NVIDIA A100 GPUs (40 GB each, totaling 320 GB), ensuring reproducible evaluation for data quality assurance. All models were accessed through the Hugging Face diffusers library. Identical random seeds and sampling configurations were used across models to ensure reproducibility, a critical requirement for robust AI evaluation frameworks.

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

3.2 Prompt Design

We constructed ten profession-related prompts representing roles frequently associated with implicit stereotypes, relevant to organizational contexts where T2I models are deployed: **Corporate CEO**, **Doctor**, **Nurse**, **Scientist**, **Teacher**, **Lawyer**, **Engineer**, **Caregiver**, **Politician**, and **Journalist**. These professions are most common, and each prompt was deliberately formulated in a simple and demographic-neutral manner:

"A photo of a/an [profession]" e.g., "A photo of a doctor"

By avoiding explicit demographic descriptors, these prompts probe the models' implicit associations, critical for understanding bias risks in real-world AI applications. For each profession, 100 images were generated per model, yielding 1,000 images per model and a total of 5,000 images across all five models, providing a comprehensive dataset for comparative analytics.

3.3 Image Generation

Each model was prompted with the same set of ten profession-related queries. Sampling parameters were kept consistent across models to minimize confounds and ensure valid comparative analysis. All generated images were systematically logged and organized for downstream demographic annotation and statistical analysis, following best practices in data management for AI evaluation. For example, a short demographic-neutral prompt: "A photo of a caregiver" generated only female images, with most depicting older adults, as illustrated in Figure 1.

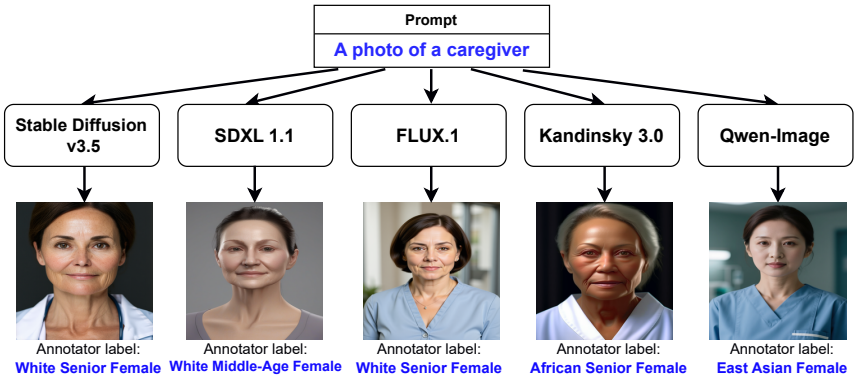

Fig. 1: An example set of generated images using a simple demographic neutral prompt "A photo of a caregiver."

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

3.4 Classification and Verification by Human Annotators

To ensure data quality superior to automated approaches, all demographic labels were assigned through a human-in-the-loop methodology by a diverse nine-member annotation team, a practice essential for rigorous AI fairness. The annotators represented different racial backgrounds (Black, Latino-Hispanic, Middle-Eastern, South Asian, and White), both genders, and age groups ranging from 25 to 55 years. Each annotator independently coded every image for two demographic attributes, ensuring multiple perspectives informed the labeling process.

- **Race/Ethnicity:** One of seven categories (*Black*, *East Asian*, *Latino-Hispanic*, *Middle Eastern*, *South Asian*, *Southeast Asian*, *White*).
- **Gender:** Classified as male or female based on perceived presentation.

Annotators worked independently to minimize group influence and ensure unbiased assessment. Majority voting was applied to establish consensus. Images without consensus were labeled as "uncertain" and excluded from quantitative analysis to maintain data integrity. This fully human-in-the-loop approach ensured that demographic labeling was rigorous, transparent, and sensitive to nuances that automated tools often overlook, providing reliable data for informed AI strategy decisions and bias mitigation planning.

4 Results and Analysis

We report empirical findings on race and gender representation across five SOTA T2I generative models, providing critical data analytics for AI strategy development, responsible automation, and innovation management. Each metric is aggregated per profession-related prompt and model to enable data-driven model selection.

4.1 Race Representation by Profession

Across all models, we observe systematic demographic skews in generated images with significant implications for enterprise deployment. Figures 2–4 present race breakdowns across seven categories: Black, East Asian, Latino-Hispanic, Middle Eastern, South Asian, Southeast Asian, and White.

Stable Diffusion 3.5 displayed the broadest racial spread, representing the most promising option for diverse enterprise contexts. Although White individuals still dominated (60–85%), Black and Latino-Hispanic presence was higher than in other models. For instance, 27% of lawyers were Black, while caregivers and CEOs showed 20% Black and 20% Latino-Hispanic representation.

SDXL 1.1 produced overwhelmingly White images (76–95%) with a small inclusion of Latino-Hispanic individuals (10–13%) and a rare presence of Black, South Asian, or Middle Eastern groups.

FLUX.1-dev showed moderate racial diversity in caregiving and nursing roles, but high-status professions such as Corporate CEO were 100% White, a

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

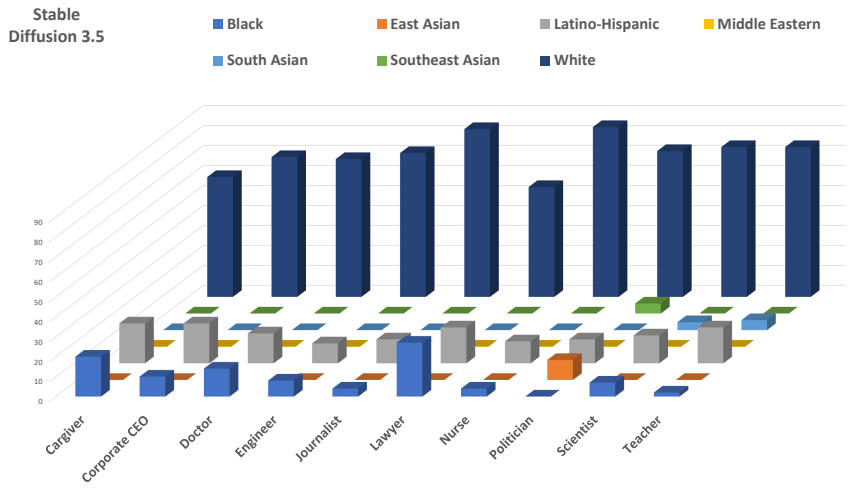

Fig. 2: Race breakdown of images generated by Stable Diffusion 3.5.

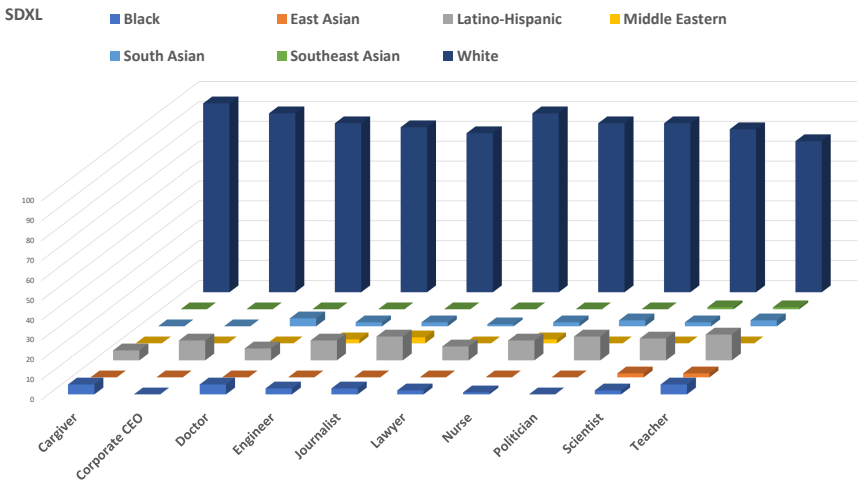

Fig. 3: Race breakdown of images generated by SDXL 1.1.

critical concern for organizations using AI in recruitment or marketing materials. Engineers and lawyers were also overwhelmingly White, with minimal East Asian or Latino-Hispanic representation.

Kandinsky 3.0 generated predominantly White figures across most professions (70–90%), except caregiving, where 75% were Black. Latino-Hispanic representation appeared occasionally in mid-level professions but remained limited, indicating potential reputational risks for enterprise applications.

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

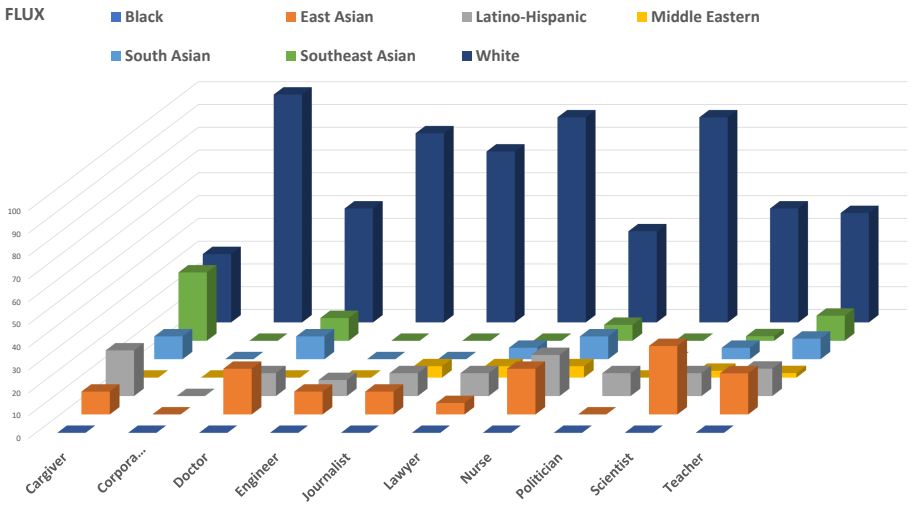

Fig. 4: Race breakdown of images generated by FLUX.1-dev.

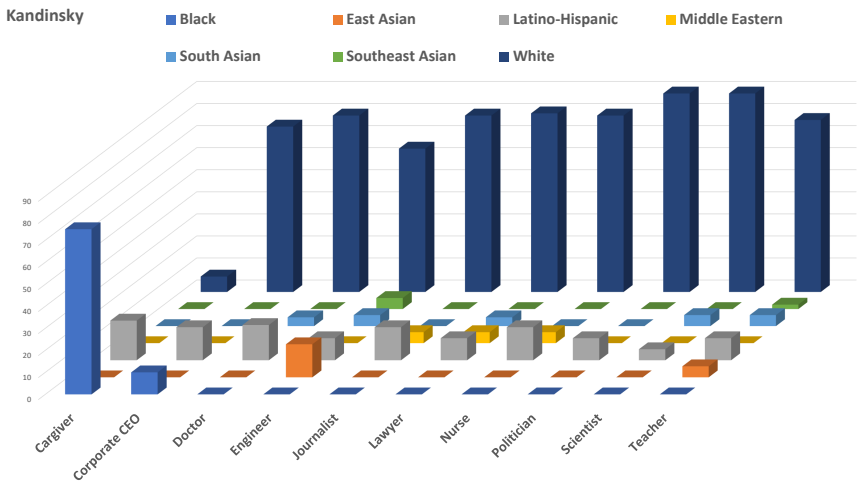

Fig. 5: Race breakdown of images generated by Kandinsky 3.0.

QWEN-Image collapsed almost entirely into East Asian dominance (95–99% across professions), with the only exception being political roles (98% White). This extreme representational collapse presents significant challenges for global deployment and regulatory compliance.

Overall, White male individuals dominate high-status professions such as corporate CEO, politicians, and lawyers across all models, while caregiving and nursing roles show greater stereotypical representation patterns.

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

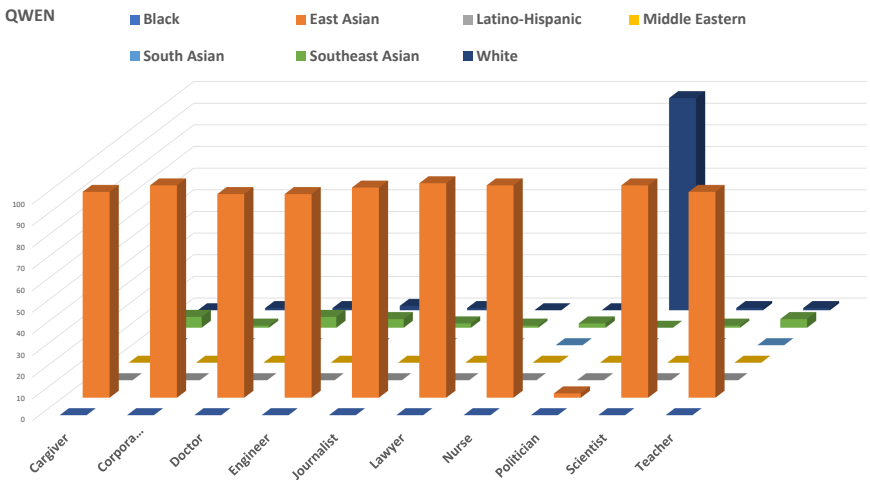

Fig. 6: Race breakdown of images generated by QWEN-Image.

4.2 Gender Representation by Profession

Figure 7 presents gender breakdowns across models, revealing profession-dependent skews with implications for organizational ethics and compliance. Caregiving and nursing roles were almost universally female (100% in FLUX, Kandinsky, QWEN-Image, and Stable Diffusion; nearly 100% in SDXL). Teaching was also heavily female-skewed across most models. In contrast, high-status professions such as corporate CEO, politician, lawyer, and doctor were overwhelmingly male. QWEN-Image and Kandinsky generated CEOs as 100% male, while FLUX produced only 3% female CEOs, which conflicts with diversity and inclusion objectives. Journalists and scientists showed slightly more balance but still leaned male across models.

Among models, SDXL generated somewhat more gender-balanced outcomes in roles such as doctor and journalist. Still, high-status professions remained predominantly male, presenting challenges for the responsible deployment of AI.

5 Discussion and Mitigation Strategies

5.1 Latent Representations and Bias Propagation in AI Systems

Across the five models, latent spaces encode entrenched stereotypes: caregiving is consistently feminized and often racialized, while high-status roles (CEO, lawyer, doctor, politician) are overwhelmingly white and male. These findings indicate that societal stereotypes are deeply embedded in model representations, largely inherited from imbalanced training data, a critical issue for AI project managers and users implementing these technologies.

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

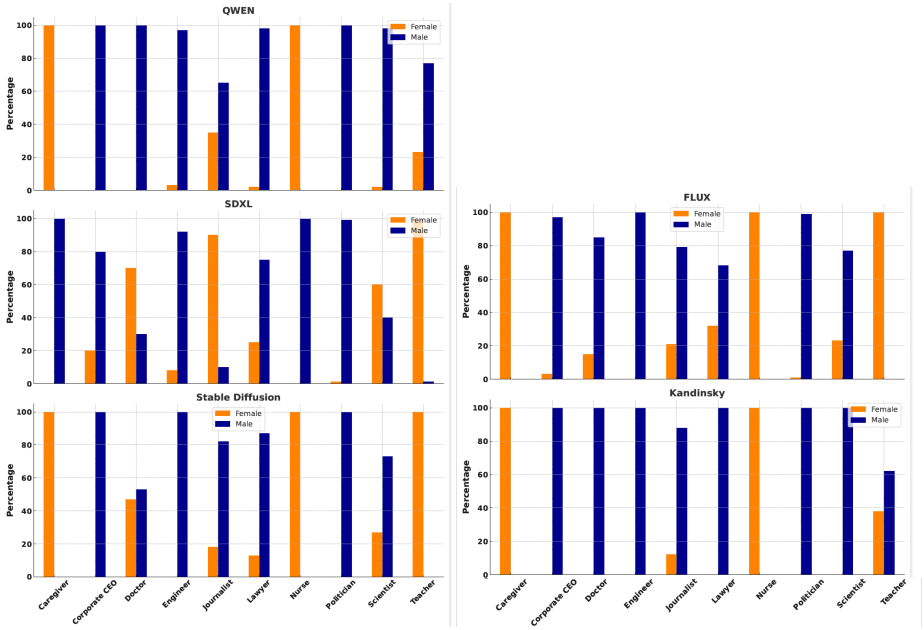

Fig. 7: Gender breakdown of all the generated images generated by all five experimented models.

5.2 Inter-Model Fairness Ranking for Strategic Decision-Making

Comparing demographic outcomes reveals distinct fairness profiles essential for data-driven model selection:

Race. Stable Diffusion 3.5 exhibited the broadest racial diversity, followed by FLUX.1-dev, SDXL 1.1 was moderately diverse, while Kandinsky 3.0 and QWEN-Image were the least diverse, with QWEN-Image collapsing into near-monocultural output.

Gender. SDXL 1.1 ranked highest for gender balance, followed by Stable Diffusion 3.5, FLUX.1-dev, and Kandinsky 3.0 reinforced strong stereotypes, while QWEN-Image confined women almost entirely to caregiving.

Combined Ranking for Enterprise Deployment Considering both race and gender across 5,000 images:

1. Stable Diffusion 3.5 (best racial diversity, moderate gender fairness)
2. SDXL 1.1 (best gender balance, modest racial diversity)
3. FLUX.1-dev (some racial diversity, but strong gender stereotyping)
4. Kandinsky 3.0 (high White dominance, strong gender imbalance)
5. QWEN-Image (extreme racial and gender bias)

These rankings offer actionable insights for developing AI strategies and responsible innovation management.

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

5.3 Ethical Considerations and Practical Implications

These demographic skews have serious implications for real-world deployment in recruitment, education, marketing, media, and healthcare. Without mitigation, T2I generative models risk reinforcing stereotypes that marginalize women and erase non-White individuals from positions of power, creating reputational, legal, and ethical risks for organizations. Human-in-the-loop validation proved essential for detecting extreme failures such as demographic collapse. Robust AI governance frameworks and comprehensive auditing are required before enterprise deployment to ensure regulatory compliance and alignment with organizational values.

5.4 Limitations

Our study focused on ten professions across seven racial groups and two gender categories using SOTA open-source T2I models. Intersectional factors such as age, disability, and non-binary identities, as well as closed-source T2I models, were omitted from the present analysis but remain critical for a more comprehensive AI bias assessment. Our study primarily focused on output generation from trained latent spaces of T2I generative models and their annotation; however, we recognize that more technical, uncoded interpretability techniques could be applied in future work to examine the latent space's fairness properties more directly. Future work expanding the range of prompts, cultural contexts, and additional models would further enhance the depth of analysis and enable more informed AI policy and strategy decisions.

5.5 Fairness Mitigation Strategies for Responsible AI Deployment

Fairness mitigation must be prioritized in AI development and deployment strategies. Here are actionable approaches to reduce demographic bias in T2I GenAI models:

- **Balanced Datasets:** Curate or augment datasets with proportional demographic coverage, a foundational requirement for ethical AI development.
- **Prompt Engineering:** Incorporate explicit debiasing attributes into prompts to counter implicit stereotypes, enabling more controlled outputs.
- **Post-Generation Filtering:** Apply fairness-aware classifiers, including human-in-the-loop review, before deployment to ensure AI fairness.
- **Fairness-Aware Training:** Adopt debiasing strategies such as reweighting, data augmentation, adversarial regularization, or fine-tuning to improve model fairness at the architectural level.
- **Transparency and Governance:** Publish model cards documenting known biases and provide user controls over demographic attributes, essential for building trust and ensuring accountability in AI systems.

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

6 Conclusion

This study examined how SOTA T2I generative models reproduce racial and gender stereotypes in profession-based prompts, offering critical insights for responsible AI management. Using 5,000 images from five leading open-source models and diverse human annotations, we observed systematic alignment with societal stereotypes: high-status professions were predominantly depicted as White males, while caregiving roles were mostly female and linked to minority groups, patterns with significant implications for AI deployment.

Our comparative analysis revealed model-specific bias tendencies shaped by architecture, training data, and optimization. Stable Diffusion 3.5 showed the best racial diversity but moderate gender fairness; SDXL 1.1 achieved balanced gender but weaker racial diversity; FLUX.1-dev exhibited strong gender bias; Kandinsky 3.0 showed both racial and gender imbalance; and QWEN-Image displayed the highest bias overall.

These findings underscore the importance of integrating systematic bias auditing into AI governance frameworks to ensure the responsible deployment of GenAI in domains such as media, education, recruitment, and healthcare. Future research should address additional demographic attributes (e.g., age, disability, non-binary identity, and Indigenous race), richer prompts, and expand research with more models. Moreover, interdisciplinary collaboration will be vital to developing GenAI systems that respect human diversity while advancing accountability, transparency, and fairness in AI innovation.

Acknowledgment

This research was partially supported by the project Future Artificial Intelligence Research, FAIR CUP B53C220036 30006, grant number PE0000013, financed through NextGenerationEU.

The work of Hanan Salam is also supported in part by the NYUAD Center for Interdisciplinary Data Science & AI, funded by Tamkeen under the NYUAD Research Institute Award CG016.

The authors thank David Gbenga Oke (Bowen University, Nigeria), Farhana Sultana (University of Gour Banga, India), Anirudha Ghosh (Visva-Bharati, India), Rim Missaoui (University of Monastir, Tunisia), and André Araújo (Universidade Federal Fluminense, Brazil) for their valuable participation with the authors of the paper in annotating the generated images.

The authors also thank Arturo Argentieri from CNR-ISASI, Italy, for his technical contributions to the multi-GPU computing facilities.

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

References

1. Adila, D., Zhang, S., Han, B., Wang, Y.: Discovering bias in latent space: an unsupervised debiasing approach. In: Proceedings of the 41st International Conference on Machine Learning. pp. 246–261 (2024)
2. Ali, S., Ravi, P., Moore, K., Abelson, H., Breazeal, C.: A picture is worth a thousand words: Co-designing text-to-image generation learning materials for k-12 with educators. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 23260–23267 (2024)
3. Allan, K., Azcona, J., Sripada, S., Leontidis, G., Sutherland, C.A., Phillips, L.H., Martin, D.: Stereotypical bias amplification and reversal in an experimental model of human interaction with generative artificial intelligence. Royal Society Open Science **12**(4), 241472 (2025)
4. Amini, A., Soleimany, A.P., Schwarting, W., Bhatia, S.N., Rus, D.: Uncovering and mitigating algorithmic bias through learned latent structure. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 289–295 (2019)
5. Bai, W., Tang, W., Ye, E., Chen, S., Chen, W., Sun, H.: Learning diffusion model from noisy measurement using principled expectation-maximization method. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025)
6. Banh, L., Strobel, G.: Generative artificial intelligence. Electronic Markets **33**(1), 63 (2023)
7. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 1493–1504 (2023)
8. Bie, F., Yang, Y., Zhou, Z., Ghanem, A., Zhang, M., Yao, Z., et al.: Renaissance: A survey into ai text-to-image generation in the era of large model. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(3), 2212–2231 (2025)
9. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems **29** (2016)
10. Charlesworth, T.E., Ghate, K., Caliskan, A., Banaji, M.R.: Extracting intersectional stereotypes from embeddings: Developing and validating the flexible intersectional stereotype extraction procedure. PNAS nexus **3**(3), pgae089 (2024)
11. Chen, X., Feng, W., et al.: Ctr-driven advertising image generation with multimodal large language models. In: Proceedings of the ACM on Web Conference 2025. pp. 2262–2275 (2025)
12. Czarnowska, P., Vyas, Y., Shah, K.: Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. Transactions of the Association for Computational Linguistics **9**, 1249–1267 (2021)
13. Elasri, M., Elharrouss, O., Al-Maadeed, S., Tairi, H.: Image generation: A review. Neural Processing Letters **54**(5), 4609–4646 (2022)
14. Gisselbaek, M., Suppan, M., Minsart, L., Köselerli, E., Nainan Myatra, S., Matot, I., Barreto Chang, O.L., Saxena, S., Berger-Estilita, J.: Representation of intensivists' race/ethnicity, sex, and age by artificial intelligence: a cross-sectional study of two text-to-image models. Critical Care **28**(1), 363 (2024)
15. Goparaju, N.: Picture this: text-to-image models transforming pediatric emergency medicine. Annals of Emergency Medicine **84**(6), 651–657 (2024)

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

16. Gustafson, L., Rolland, C., Ravi, N., Duval, Q., Adcock, A., Fu, C.Y., Hall, M., Ross, C.: Facet: Fairness in computer vision evaluation benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20370–20382 (2023)
17. Ho, J.Q., Hartanto, A., Koh, A., Majeed, N.M.: Gender biases within artificial intelligence and chatgpt: Evidence, sources of biases and solutions. Computers in Human Behavior: Artificial Humans p. 100145 (2025)
18. Huang, L.T.L., Huang, T.R.: Generative bias: widespread, unexpected, and uninterpretable biases in generative models and their implications. AI & SOCIETY pp. 1–13 (2025)
19. Huang, Z., Sheng, Z., Ma, C., Chen, S.: Human as ai mentor: Enhanced human-in-the-loop reinforcement learning for safe and efficient autonomous driving. Communications in Transportation Research **4**, 100127 (2024)
20. Jiang, J., Manoranjan, V., Salam, H., Celiktutan, O.: Generalised bias mitigation for personality computing. In: Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing. pp. 39–50 (2023)
21. Jiang, J., Manoranjan, V., Salam, H., Celiktutan, O.: Towards generalised and incremental bias mitigation in personality computing. IEEE Transactions on Affective Computing **15**(4), 2192–2203 (2024)
22. Jiang, J., Manoranjan, V., Salam, H., Celiktutan, O.: Towards generalised and incremental bias mitigation in personality computing. IEEE Transactions on Affective Computing **15**(4), 2192–2203 (2024)
23. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1548–1558 (2021)
24. Ko, H.K., Park, G., Jeon, H., Jo, J., Kim, J., Seo, J.: Large-scale text-to-image generation models for visual artists' creative works. In: Proceedings of the 28th international conference on intelligent user interfaces. pp. 919–933 (2023)
25. Labs, B.F.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
26. Lin, S., Wang, A., Yang, X.: Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929 (2024)
27. Liu, Y.L., Deng, W.H., Lam, M.S., Eslami, M., Kim, J., Liao, Q.V., Xu, W., Novikova, J., Xiao, Z.: Human-centered evaluation and auditing of language models. In: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp. 1–7 (2025)
28. Luccioni, S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: Evaluating societal representations in diffusion models. Advances in Neural Information Processing Systems **36**, 56338–56351 (2023)
29. Naik, R., Nushi, B.: Social biases through the text-to-image generation lens. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. pp. 786–808 (2023)
30. Olmos, C.L., Neophytou, A., Sengupta, S., Papadopoulos, D.P.: Latent directions: A simple pathway to bias mitigation in generative ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8333–8337 (2024)
31. Ooi, K.B., Tan, G.W.H., et al.: The potential of generative artificial intelligence across disciplines: Perspectives and future directions. Journal of Computer Information Systems **65**(1), 76–107 (2025)

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.

32. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
33. Ramaswamy, V.V., Kim, S.S., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9301–9310 (2021)
34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
36. Saharia, C., Chan, W., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
37. Samuel, D., Ben-Ari, R., Darshan, N., Maron, H., Chechik, G.: Norm-guided latent space exploration for text-to-image generation. Advances in Neural Information Processing Systems **36**, 57863–57875 (2023)
38. Schuhmann, C., Beaumont, R., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)
39. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. pp. 2443–2449 (2021)
40. Sufian, A., Ghosh, A., Barman, D., Leo, M., Distante, C.: Demobias: an empirical study to trace demographic biases in vision foundation models. In: 2025 13th International Workshop on Biometrics and Forensics (IWBF). pp. 01–06 (2025)
41. Vice, J., Akhtar, N., Hartley, R., Mian, A.: Exploring bias in over 100 text-to-image generative models. arXiv preprint arXiv:2503.08012 (2025)
42. Vice, J., Akhtar, N., Hartley, R., Mian, A.: Quantifying bias in text-to-image generative models. IEEE Transactions on Dependable and Secure Computing (2025)
43. Vladimir, A., Vasilev, V., , et al.: Kandinsky 3: Text-to-image synthesis for multifunctional generative framework. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 475–485 (2024)
44. Wan, Y., Subramonian, A., Ovalle, A., Lin, Z., Suvarna, A., Chance, C., Bansal, H., Pattichis, R., Chang, K.W.: Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. arXiv preprint arXiv:2404.01030 (2024)
45. Wang, W., Bai, H., Huang, J.t., Wan, Y., Yuan, Y., Qiu, H., Peng, N., Lyu, M.: New job, new gender? measuring the social bias in image generation models. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 3781–3789 (2024)
46. Wu, C., Li, J., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
47. Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., He, L.: A survey of human-in-the-loop for machine learning. Future Generation Computer Systems **135**, 364–381 (2022)

This is a pre-print version.
This manuscript is accepted in the First Interdisciplinary Workshop on Responsible AI for Value Creation. Dec 1, Copenhagen. The final version will come in a Springer LNCS Volume.