An Exploration of Prompt-Based Biases in AI Art-Generated Tools

NABILA CHOWDHURY, University of Toronto, Canada
ANASTASIA KUZMINYKH, University of Toronto, Canada

With the growing popularity of AI art-generating tools, the biases in their outcomes have become an increasingly important issue. While prior research focused on how to generate more realistic or aesthetic art, more work is required on the techniques for mitigating biased outcomes. Given that these systems commonly take text-based input, we explore the effects of prompt formulation of the appearance of such biases. We first discuss the early results of the analysis of public discourse and users' observations on these effects; we then illustrate the identified associations through a comparative analysis of outcomes from two popular prompt-based AI art-generating tools, showing gender and racial bias variations based on the use of certain keywords in prompts.

Additional Key Words and Phrases: AI, image, Bias, AI, User-generated Prompts, Generative Art

ACM Reference Format:
Nabila Chowdhury and Anastasia Kuzminykh. 2023. An Exploration of Prompt-Based Biases in AI Art-Generated Tools . 1, 1 (March 2023), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND BACKGROUND

AI-powered tools have spread in every aspect of life, including algorithmic decision-making for job [12], mortgage approval [39], higher education [5], criminal prediction [20], support creativity through AI art generation [7], collaborative writing [17], etc. Recent research, however, repeatedly demonstrated that the outcomes from such algorithm-driven tools are often biased [4, 6, 23, 33]. For instance, in existing AI decision-making tools, researchers have identified gender, race, and cultural biased tools [21, 23, 33]. Similarly, biases were noted in the outcomes of AI art-generating tools [31, 35], recently gaining in popularity in supporting creative tasks. However, the majority of the recent research on AI art generation tools was conducted to understand how to generate accurate images [10, 15, 18, 27], and less attention was paid to biases in AI art generation outcomes [7, 10, 15, 18, 27].

With the emerging use of AI-based tools, it becomes increasingly important to mitigate the potential harm introduced by biased outcomes [22]. In the effort to reduce such biases, researchers have been actively exploring several approaches, including the development of bias mitigation frameworks [30], designing explainable AI systems [19] to increase systems' transparency [3], etc. This research is predominantly focused on the system-driven sources of the biases. For instance, exploring biases in outcomes of AI systems for decision-making [16], prior research has identified such sources as biased data sets, an underrepresented group in the trained data, inappropriate data labeling, wrong data analysis model for wrong data or circumstances, biases of the system programmers, etc. [4, 6, 31, 33]. At the same time, one of the important characteristics of AI art-generation tools is defined by the format of user input for these systems: most commonly through free-form text prompts. User observations reflected in the public media discourse (e.g. [28, 29])

Authors' addresses: Nabila Chowdhury, nabila.chowdhury@mail.utoronto.ca, University of Toronto, Toronto, Canada; Anastasia Kuzminykh, University of Toronto, Toronto, Canada, anastasia.kuzminykh@mail.utoronto.ca.

suggest potential effects of the text prompts used with the art-generation tools and the appearance of biases in their outcomes, however, currently, there is a paucity of research on such effects.

Most research on the relationships between user-formulated queries and the specifics of the corresponding outcomes can be found in the field of information retrieval. For example, Sanchiz et al. [32] identified in their study that longer formulation of the queries perform worse than shorter ones during web searches. Other research shows that formulating queries in a request format (e.g., “may I, should I") might not give the expected output [34]. Keyvan and Huang [14] showed that reformulating queries with certain keywords (e.g., instead of searching “java", searching “java programming language") might give users better web results. In a similar vein, Zamani et al. [38] identified that the query formulation that often lacks context and keywords due to users’ lack of expertise in the area results in ambiguous output. Papenmeier [25] discussed how the pattern of query formation varies between a novice and an expert in the context of online retail, particularly due to the incorporation of certain keywords, which results in better output.

To begin exploring the potential association of prompt formation with the appearance of biases in AI-generated art, we conducted an initial study by analyzing online discussions about prompts and AI art generation tools on publicly available media outlets. We focused on the following questions: 1) Whether and how people discuss biases in association with prompts; and 2) What prompt variations do people note in association with those biased outcomes? We found that people discuss biases in association with the prompt variations, most commonly noting gender and racial bias. People also discuss how varying certain words associated with cultural stereotypes in a prompt generate different outcomes. We then compared the outcomes generated by common AI art generation tools (DALL-E[1] and Stable Diffusion[2]) in response to the prompts variations informed by the first study. We discuss our early findings, provide early illustrations of the prompt variation effects through a comparative analysis of the art outcomes, and outline the implications of this research direction for human-AI interaction.

2 EARLY ANALYSIS OF PUBLIC DISCOURSE

To understand public discourse around biases in prompt-based AI art-generation tools, we collected discussions from Reddit[3], Twitter[4], and blog platforms, using the following search keywords: prompts for DALL-E, prompts for Stable Diffusion, prompts for AI art generation, biases in AI art generation tool, biases in DALL-E, biases in Stable Diffusion, DALL-E, and Stable Diffusion. Through this process, we identified 145 unique public conversation points (Reddit: 76, Twitter:52, Blogs: 17) regarding AI art generation and biases. We then excluded general discussions around bias and AI, posts purely sharing prompts, and posts in which the comment section had age restrictions, which resulted in a final dataset of 51 unique discussions (Reddit: 21, Twitter: 20: Blogs: 10) containing 102 unique discussion points. The final dataset was then thematically analyzed by two members of the research team using reflexive analysis approach.

Our analysis first showed that people discuss biases in AI art generation tools in association with prompts, most commonly gender and racial bias, along with economical and political biases. For instance, discussing the association of prompts with gender-biased outcomes, one user shared different outcomes of "handsome man" vs. "beautiful woman": highly edited, portrait-style images of women compared to the images generated for men. In this discussion, other users commented on their own idea of what "handsome" and "beautiful" means [26] and noted that most of the outputs would not include arts of certain demographics (e.g., Asian). We also found prominent discussions of racial bias, e.g., “Where demographic parity = 25%, perceived female figures with darker skin tones are produced 4% of the time; perceived male

[1]https://openai.com/dall-e-2/
[2]https://stablediffusionweb.com/
[3]https://https://www.reddit.com/
[4]https://https://twitter.com/

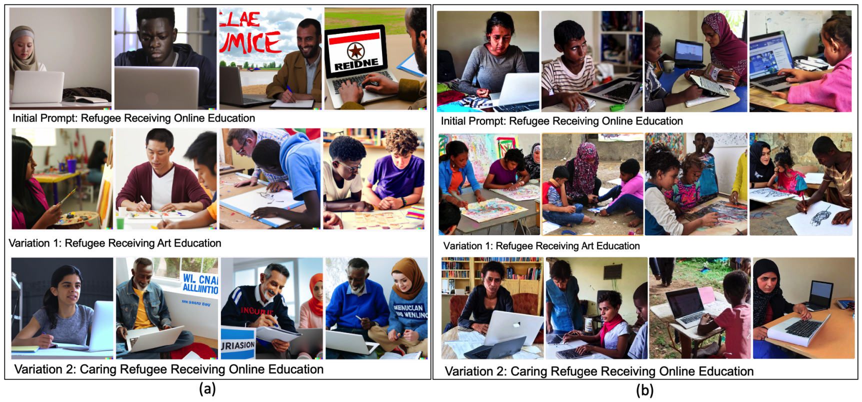

Fig. 1. (a) Example of Racial Bias in DALL-E and (b) Example of Racial Bias in Stable Diffusion

figures with darker skin tones are produced 3% of the time." Another user noted, “Westerners post Asian people only if they specifically looked for art with them." Racial bias is also discussed in these platforms in relation to the styles of the outcome. For example, articles have noted how outputs of these art-generating tools might seem Westernized or European even though the style is not mentioned in the prompts [8, 36].

We also analyzed the prompt variations discussed by users. We first found the discussed associations between sexist outcomes use of certain job types with assertive words as prompts [28, 29], e.g., “Including search terms like “CEO” exclusively generates images of white-passing men in business suits, while using the word “nurse” or “personal assistant” prompts the system to create images of women" [28]. Stereotypes on gender-based roles are also noted in the analyzed blog articles [1, 24, 36, 37], e.g., “To portray gender biases in Stable Diffusion, prompts are selected whose outputs reflect possible gender biases: a face of an intelligent person, a face of a kind person, [...] a face of a passionate person" [24]. Another example is, "[...] these generators can often be based on stereotypical biases, [...] images can often be Westernized, or show favor to certain genders or races, depending on the types of phrases used [1]." We also found this theme in the user discussions on certain prompts, e.g. “Even the weakest link to womanhood or some aspect of what is traditionally conceived as feminine returned pornographic imagery". Overall, users discuss the relationship between biased generated art and variation of culturally stereotyped words (mostly adjectives) used in prompts.

3 COMPARATIVE ANALYSIS OF AI ART GENERATION TOOLS

We conducted a comparative analysis of AI-generated art prompted by the corresponding variations of prompts. We choose to use DALL-E and Stable Diffusion, as they generate outcomes by taking text-based prompts. We formulated two prompts under the two common themes identified from the media discourse analysis: gender and racial bias. We formulated an initial prompt and two variations of it, using early findings from phase 1, to see if that varies the appearance of biases in the outcomes. The initial prompts we used were: racial bias - “Refugee receiving online education”; gender bias - “Assertive Professor”; along with two variations for each prompt (see Fig 1 and Fig 2) The prompt variations were introduced through alternating certain adjectives, identified as stereotypical in phase 1. We first provided both tools with the initial prompts followed by two prompt variations.

For the first set of prompts, all the prompts contained the word ‘refugee’ which is found (from early analysis of public discourse) to have a racial stereotypical meaning. This racial stereotypical meaning was also reflected in outcomes generated both by DALL-E and Stable Diffusion for the initial prompt. In variation 1, only by changing the keyword

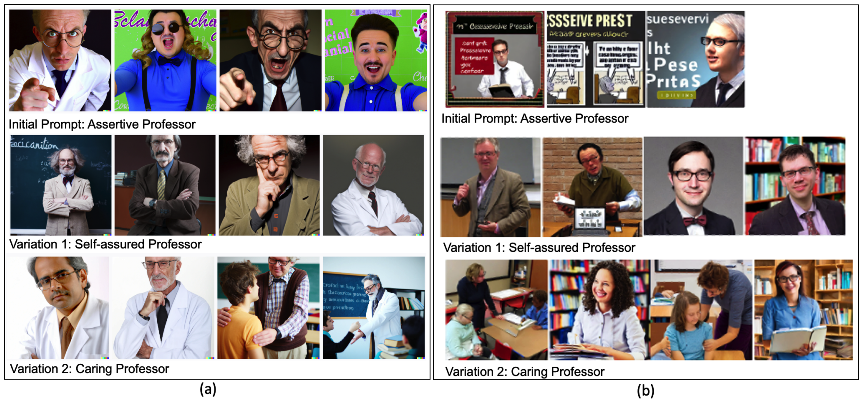

Fig. 2. (a) Example of Gender Bias in DALL-E and (b) Example of Gender Bias in Stable Diffusion

'online' to 'art' in the initial prompt, we noticed for DALL-E the appearance of racial outcomes changed by incorporating 1 different racial representative image. Although for Stable Diffusion the biased racial outcomes did not change with the prompt variation, the appearance of the bias reflected a certain demographic (only women and children). In variation 2, we combined the word 'caring' with refugee, as from early analysis of public discussion we identified gender bias associated with the word 'caring'. In association with this prompt, the appearance of biased outcomes changed for DALL-E from racial to racial and age-related. The appearance of racial and demographic bias remains the same for Stable Diffusion in relation to the prompt variation 2.

For the second set of prompts, which contained the keyword 'professor' (associated with gender stereotypes related to profession from our early analysis), we varied three adjectives to see their relationship to the appearances of biases in the generated outcomes. In association with the initial prompt variation (a stereotypical adjective 'assertive' with gender stereotypes associated word 'professor') we saw gender- and racially-biased appearances in both DALL-E and Stable Diffusion generated outcomes. With the adjective 'self-assured', the appearance of racial and gender bias remained the same in both AI tools and with the adjective 'caring', while the gender-biased appearance remained the same in DALL-E generated outcomes, the racial appearance had changed. The appearance of gender bias in Stable Diffusion generated outcomes had significantly changed, however, the racial bias in the appearance of those outcomes was prevalent.

This early comparative analysis illustrates that the appearance of biased outcomes varies in association with the prompt variations. We found when the prompts are formulated with varied stereotyped adjectives, the appearance of the biased outcome also changed for both DALL-E and Stable Diffusion, although varies across the two selected AI art generation platforms.

4 CONCLUSION

Current AI art generation tools provide a creative collaboration between humans and AI, although the outcome has biases. Such biases were discussed by designers, artists, and people. In July 2022, OpenAI [2] shared the news on implementing a mitigation technique to reflect diversity on the outcomes when input is given a generic word such as, "Firefighter" and "Teacher"; however, in our prompt exploration, we found how on the "Professor" DALL-E generated biased results. Thus, it is important to continue exploring the appearance of biases in association with prompts used for art-generating tools, especially as AI art is being integrated into popular platforms like TikTok [11], Canva [9], and being used to design content on the internet [13]. The development of a better understanding of how users formulate their prompts to generate art and which aspects of a prompt trigger biased outcomes would allow us to design for guiding users in formulating prompts to avoid biased outcomes.

REFERENCES

[1] Fionna Agomuoh. 2022. AI image generators appear to propagate gender and race stereotypes. https://www.digitaltrends.com/computing/ai-tool-reveals-biases-in-text-to-image-generators/.

[2] Open AI. 2022. Reducing Bias and Improving Safety in DALL·E 2. https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/.

[3] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[4] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (may 2018), 54–61. https://doi.org/10.1145/3209581

[5] Ryan S Baker and Aaron Hawn. 2022. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* 32, 4 (2022), 1052–1092.

[6] Amin Bigdeli, Negar Arabzadeh, Shirin SeyedSalehi, Morteza Zihayat, and Ebrahim Bagheri. 2022. Gender Fairness in Information Retrieval Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3436–3439.

[7] Eva Cetinic and James She. 2022. Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 2 (2022), 1–22.

[8] Neel Dhanesha. 2022. AI art looks way too European. https://www.vox.com/recode/23405149/ai-art-dall-e-colonialism-artificial-intelligence.

[9] Larry Ferlazzo. 2022. TEACHING ENGLISH THROUGH USING AI FOR ART CREATION. https://larryferlazzo.edublogs.org/2022/10/02/teaching-english-through-using-ai-for-art-creation/.

[10] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131* (2022).

[11] Harry Guinness. 2022. TikTok's new AI art filter riffs on your text. https://www.popsci.com/technology/tiktoks-ai-filter-text-to-image-generator/.

[12] Claretha Hughes, Lionel Robert, Kristin Frady, and Adam Arroyos. 2019. Artificial intelligence, employee engagement, fairness, and job outcomes. In *Managing technology and middle-and low-skilled employees*. Emerald Publishing Limited.

[13] Shelly Tan Monique Woo Kevin Schaul, Hamza Shaban and Nitasha Tiku. 2022. AI can now create images out of thin air. See how it works. https://www.washingtonpost.com/technology/interactive/2022/ai-image-generator/?tid=ss_tw.

[14] Kimiya Keyvan and Jimmy Xiangji Huang. 2022. How to Approach Ambiguous Queries in Conversational Search: A Survey of Techniques, Approaches, Tools, and Challenges. *Comput. Surveys* 55, 6 (2022), 1–40.

[15] Kyungsun Kim, Jeongyun Heo, and Sanghoon Jeong. 2021. Tool or Partner: The Designer's Perception of an AI-Style Generating Service. In *International Conference on Human-Computer Interaction*. Springer, 241–259.

[16] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[17] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[18] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. 2019. Controllable text-to-image generation. *Advances in Neural Information Processing Systems* 32 (2019).

[19] Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 147–159.

[20] Carolyn McKay. 2020. Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making. *Current Issues in Criminal Justice* 32, 1 (2020), 22–39.

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

[22] Safiya Umoja Noble. 2018. Algorithms of oppression. In *Algorithms of Oppression*. New York University Press.

[23] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[24] Alberto Osorio. 2022. Exploring GenderBias in StableDiffusion and StableDiffusion-2. https://medium.com/@osoriomunozalberto/exploring-genderbias-in-stablediffusion-and-stablediffusion-2-4b8bb2fa6c01.

[25] Andrea Papenmeier, Alexander Frummet, and Dagmar Kern. 2022. "Mhm..."–Conversational Strategies For Product Search Assistants. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. 36–46.

[26] Reddit Post. 2022. Why does dalle give such different results. https://www.reddit.com/r/dalle2/comments/x2xv7f/why_does_dalle_give_such_different_results_for/.

[27] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. Learn, imagine and create: Text-to-image generation from prior knowledge. *Advances in neural information processing systems* 32 (2019).

[28] Janus Rose. 2022. The AI That Draws What You Type Is Very Racist, Shocking No One. https://www.vice.com/en/article/wxdawn/the-ai-that-draws-what-you-type-is-very-racist-shocking-no-one.

[29] Janus Rose. 2022. This Tool Lets Anyone See the Bias in AI Image Generators. https://www.vice.com/en/article/bvm35w/this-tool-lets-anyone-see-the-bias-in-ai-image-generators.

[30] Drew Roselli, Jeanna Matthews, and Nisha Talagala. 2019. Managing bias in AI. In *Companion Proceedings of The 2019 World Wide Web Conference*. 539–544.

[31] Joni Salminen, Soon-gyo Jung, Shammur Chowdhury, and Bernard J Jansen. 2020. Analyzing demographic bias in artificially generated facial pictures. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[32] Mylène Sanchiz, Franck Amadieu, P-V Paubel, and Aline Chevalier. 2020. User-friendly search interface for older adults: supporting search goal refreshing in working memory to improve information search strategies. *Behaviour & Information Technology* 39, 10 (2020), 1094–1109.

[33] Vivek K Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. 2020. Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms. *Journal of the Association for Information Science and Technology* 71, 11 (2020), 1281–1294.

[34] Amanda Spink and H Cenk Ozmultu. 2002. Characteristics of question format web queries: An exploratory study. *Information processing & management* 38, 4 (2002), 453–471.

[35] Ramya Srinivasan and Kanji Uchino. 2021. Biases in generative art: A causal look from the lens of art history. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 41–51.

[36] Kyle Wiggers. 2022. Researchers find race, gender, and style biases in art-generating AI systems. https://venturebeat.com/ai/researchers-find-evidence-of-bias-in-art-generating-ai-systems/.

[37] Elliot Wong. 2022. AI art promises innovation, but does it reflect human bias too? https://superrare.com/magazine/2022/10/18/ai-art-promises-innovation-does-it-reflect-human-bias-instead/.

[38] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020*. 418–428.

[39] Leying Zou and Warut Khern-am nuai. 2022. AI and housing discrimination: the case of mortgage applications. *AI and Ethics* (2022), 1–11.