The Need for Flexible Interfaces for Text-to-Image Auditing: A Case Study of DALL·E 2 and DALL·E 3

CLARE PROVENZANO, PARSA RAJABI, DIANA CUKIERMAN, and NICHOLAS VINCENT, Simon Fraser University, School of Computing Science, Canada

With increased ubiquity of text-to-image systems and other generative AI technologies, potential representational harms from these systems have become high stakes. Past work has documented issues with such systems along the lines of gender and race. We present a comparative audit of the DALL·E 2 and DALL·E 3 text-to-image systems. We find that the new "prompt grounding" feature incorporated in DALL·E 3 substantially improves image diversity, though this approach generally did not show people with identities outside a fixed set of labels. We conclude with suggestions for designing flexible interfaces that could be used to audit new text-to-image models.

1 INTRODUCTION

As AI systems that enable text-to-image (and even text-to-video [5]) generation of anything from photo-realistic portraits to complicated comics become more advanced, understanding the potential representational harms [13] of these systems has become a pressing issue. There has been a surge in attention to these concerns. For instance, venues such as Bloomberg [10], the Washington Post [14], and NPR [4] have reported on biases in system outputs. Luccioni et al.'s research [9] (also covered in the press [7]) found a strong tendency of text-to-image systems to output images of white men when prompted with high prestige job titles such as "CEO" or "director". AI apps that take in photos of the users and edit them also display concerning behaviors. A case study showed that Lensa, an AI app that generates avatars of users, generated sexualized and topless photos for a women with Asian heritage, but significantly fewer sexual images for white women or white men [6].

In this short paper, we describe a small scale audit aimed at comparing the DALL·E 2 and DALL·E 3 systems in terms of bias along lines of race, gender, and other social dimension, and discuss implications for the design of new auditing interfaces. In October 2023, OpenAI released the DALL·E 3 system, a successor to the DALL·E 2 text-to-image generator. This system was notable for explicit acknowledgement of bias-related concerns: the system card [11] states that "DALL·E 3 has the potential to reinforce stereotypes or have differential performance in domains of relevance for certain subgroups."

Overall, we observed improvements in the diversity of results, likely a result of the DALL·E 3 system's new "prompt grounding" feature that edits vague prompts to include specific details about race, gender, the location of the image, and so on. For instance, "teacher" might become "Photo of a female East Asian teacher, in a lab coat, conducting a science experiment for intrigued students in a well-equipped laboratory." DALL·E 3 will "conditionally transform a provided prompt if it is ungrounded to ensure that DALL·E 3 sees a grounded prompt at generation time." This prompt adjustment strategy seems like it may reduce representational harms in the short term, but opens the door for new societal discussions about balancing data-related interventions and post-hoc prompting.

However, we saw that DALL·E 3 struggled with depicting "non-binaryness", in both gender and in race. Similar to prior work [12], we also saw evidence that other biases fell through the cracks. For instance: women tended to appear more often smiling.

Based on our audit, we report some early design suggestions for new interfaces that might help users to conduct open-ended audits. Specifically, we suggest that any interfaces for labeling support a deal of flexibility with regards to the columns and labels that are applied. Our study also adds to calls for HCI and social computing to be centered in efforts to create fair generative AI systems that balance diverse preferences for model behavior. Our hope is that new interfaces for labeling generative AI outputs will complement new efforts to put together fairness focused benchmarks for generative AI [8].

2 RELATED WORK

A number of studies support the general trend of text-to-image models displaying biases in outputs with potential representational harms. A large scale audit study of Dall-E 2 from Sun et al. (15300 images from 153 occupations) found evidence of gender bias relative to labor statistics [12]. They also found women tended to be depicted smiling more and pitching their heads downwards. Nicoletti and Bass conducted an audit of Stable Diffusion, finding similar evidence of bias [10]. They considered 14 jobs (300 images each), and specifically considered a distinction between high-income and low-income jobs, and three categories related to crime. Luccioni et al.'s work proposed a new bias exploration method and tested the approach on Dall-E 2 and two different Stable Diffusion models [9]. Alenichev, Kingori, and Grietens studied Midjourney and found that the system struggled to produce an image of a Black doctor treating a white patient [1]. Finally, Tiku, Schaul, and Chen also audited Stable Diffusion, finding that "detailed prompts didn't mitigate this bias"[14]. Outside of audit studies, there is a line of work on feminist AI that provides a framework for thinking about many of these findings are their implications [15]. A key idea from this line of work is that visibility of bias matters.

3 METHODS

We planned our audit in early October 2023 when Dall-E 3 was the horizon, but not yet publicly available. To prepare, we first collected a small set of comparison images from Dall-E 2. Then, after Dall-E 3 was released, we collected a corresponding set of images from the newer system.

Specifically, we selected a set of 15 prompts, focusing on occupational categories as in prior work [10, 12]. We manually entered each prompt into each interface twice, generating a total of 8 images (barring three cases in which only 3 images were returned for a single issuance).This led to about 16 images for each of 15 prompts (237 images in total). While small for any kind of computational analysis, our goal was to focus on manual investigation of all the images. We wanted to produce a dataset that would be useful for direct manual comparison of DALL·E 2 and DALL·E 3 images, and in particular to reflect on the design of interfaces for future audit work.

The first author first tagged the images based on gender and race. Additional tags were added and refined in an open-ended and iterative fashion, intending at first to capture general characteristics like background, art style, and attire. Then, the author team discussed these tags. The goal of this open coding was to identify factors that might be useful to add to labeling/auditing interfaces in the future.

3.1 Prompts and audit data

Our prompts were intentionally generic and "ungrounded" [11] (e.g. simply "doctor" as opposed to "a doctor,young woman, in an operating room"). As in prior work, we sought to capture job categories with potentially high-stakes representational harms (e.g., an AI system might communicate the idea that women cannot be doctors or CEOs).

All our prompts related to people. We expect that some companies will use AI generated art rather than hiring human illustrators. This may lead to content across a variety of domains – such as journalism, entertainment, informational content – containing images that reinforce stereotypes.

Our full prompt list was: *doctor, CEO, beautiful woman, handsome man, model, construction worker, criminal, teacher, homeless person, scientist, politician, artist, social worker, musician, computer programmer.*

Our generated images were stored in a table with their date of creation, prompt, the image. Then, the first author tagged them based on gender, race, and additional labels. At the time of our study, DALL·E 3 provided the user with the full modified “grounded prompt” [11], which we stored.

4 RESULTS

To first contextualize our results, we compared the relative frequency of different identities from each system. For Dall-E 2, we analyze our own annotations of race[1] and gender. For DALL·E 3, we focus specifically on the *provided* identities filled in by the prompt grounding procedure.

In short, DALL·E 2 saw a major preference towards white and male images: “white” made up 65% of race-related tags and 62% of gender-related tags. Most individual prompts showed imbalances (in general, each of the four images would show the same identity). DALL·E 3 provided very gender balanced results (54% women) and trended towards more balanced representation of ethnicity and geography (the prompt grounding included words typically related to ethnicity such as “Latina” but also broad geographic categories such as “North American”).

We focused most on our analysis and open coding to identify relevant themes for future audit interface design. First, DALL·E 2 tended to generate four variations of the same image (e.g. Asian female doctor in scrubs), while Dall-E 3 made four distinct and different images from the same prompt (as we might expect from prompt grounding). Second, images from DALL·E 3 had more diversity in gender, race, art style, and backgrounds, as evidenced by a much larger number of labels. DALL·E 3 produced more “thoughtful”, vibrant, detailed, appropriate backgrounds, where as DALL·E 2 tended to just have blank backgrounds. Third, we perceived a tendency to avoid ambiguity in people’s appearance. A gender binary was clear in all images (except in a few cases, the person in an image was far away no gender was distinguishable). Furthermore, based on our labels, no individuals that were explicitly biracial or of mixed heritage were created. We also observed a binary approach to skin tone (although we did not attempt to systematically label skin tone, which recent work has highlighted requires serious consideration [2]). Finally, we also observed that as in [12] there were affective differences between men and women, especially around women being more likely to be shown smiling.

5 DISCUSSION

Overall, these results suggest that the DALL·E 3 prompt grounding approach does address some of the very serious concerns from prior auditing work. Future work may help to tease out the specific role that prompt modification plays relative to training data changes or other interventions. One important factor to consider is that prompt based fairness interventions may hide inequalities in the training data. Navigating this new trade-off will be an important ongoing discussion for designing AI systems, and will likely be very domain specific.

Our results do suggest some specific areas where prompt adjustment many not work, or may require more manual or data-driven adjustment of the prompt adjustment pipeline. Specifically, in order to capture non-binary gender, mixed race, affective aspects (e.g. smiling women), and other dimensions likely to be identified by future audits, an updated approach may be needed. This raises a broader question: if we assume more systems will incorporate prompt adjustment to increase generated image output diversity, how should the list of words and identities that are added to prompts be governed? Can we just keep updating our prompt grounding “vocabulary” to address such issues?

[1] We note that prompt grounding seemed to provide a mix of terms that could map to race, ethnicity, ancestry, and geographic.

Our results reflect the experiences and perspectives of our author team. There may be several additional dimensions of bias that prompt grounding would not address that we missed, but that another set of labelers might identify. Ideally, this question might be answered in a substantially more participatory manner. Drawing on social computing and crowdsourcing, we might answer this question *at scale*, and allow a wide variety of users to curate prompt grounding pipelines that fit their needs.

Allowing a wide variety of potential users to audit generated images in a flexible manner (i.e. with "open coding" philosophy) will allow product managers, software engineers and other internal stakeholders to easily check for biases – some subtle – within the image generating system.

Our results suggest immediate value in developing auditing/labeling tools that meet the following criteria: (1) emphasizing flexibility in labeling what a given user sees as important and (2) enabling users to share their labels to platforms where they will be seen and used by AI developers. Ideally, generative AI developers and the generative AI research community might invest resources in soliciting participation in using this kind of auditing tool (perhaps via a mix of a peer production style model and through fairly paid crowdwork tasks). These tools could help auditors quickly navigate through pre-generated images and tag various factors such as art style, number of people, race, gender, facial expression, and more, based on what feels salient to that auditor. Critically, when working with the outputs of such auditing tools, AI developers should be prepared to adopt techniques for handling disagreement between labelers [3].

In some cases, it could help to incorporate a social component into the labeling interface, so that users might understand how other labelers have viewed a given image. This social component might draw on research on human computation for complex tasks with global constraints, such as itinerary planning [16].

6 CONCLUSION

In a small comparative audit of the DALL·E 2 and DALL·E 3 text-to-image systems, we found that prompt grounding can substantially diversify image outputs along the lines of gender, race, ethnicity, and geography, but there are some limitations around representing identities that do not map to a fixed set of labels. We highlight the importance of auditing tools that allow for open and flexible coding from a variety of users.

REFERENCES

[1] Arsenii Alenichev, Patricia Kingori, and Koen Peeters Grietens. 2023. Reflections before the storm: the AI reproduction of biased imagery in global health visuals. *The Lancet Global Health* 11, 10 (Oct. 2023), e1496–e1498. https://doi.org/10.1016/S2214-109X(23)00329-7 Publisher: Elsevier.

[2] Teanna Barrett, Quanze Chen, and Amy Zhang. 2023. Skin Deep: Investigating Subjectivity in Skin Tone Annotations for Computer Vision Benchmark Datasets. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1757–1771.

[3] Scott Allen Cambo. 2021. *Model Positionality: A Novel Framework for Data Science with Subjective Target Concepts.* Ph. D. Dissertation. Northwestern University.

[4] Carmen Drahl. 2023. AI was asked to create images of Black African docs treating white kids. How'd it go? *NPR* (Oct. 2023). https://www.npr.org/sections/goatsandsoda/2023/10/06/1201840678/ai-was-asked-to-create-images-of-black-african-docs-treating-white-kids-howd-it-

[5] Omar H. Fares. 2024. OpenAI's new generative tool Sora could revolutionize marketing and content creation. http://theconversation.com/openais-new-generative-tool-sora-could-revolutionize-marketing-and-content-creation-223806

[6] Melissa Heikkilä. 2022. The viral AI avatar app Lensa undressed me—without my consent. *Technology Review. https://www. technologyreview. com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent* (2022).

[7] Melissa Heikkilä. 2023. These new tools let you see for yourself how biased AI image models are. https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are/

[8] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. 2023. Holistic Evaluation of Text-To-Image Models. https://doi.org/10.48550/arXiv.2311.04287 arXiv:2311.04287 [cs].

[9] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable Bias: Analyzing Societal Representations in Diffusion Models. https://doi.org/10.48550/arXiv.2303.11408 arXiv:2303.11408 [cs].

[10] Leonardo Nicoletti and Dina Bass Technology + Equality. 2023. Humans Are Biased. Generative AI Is Even Worse. *Bloomberg.com* (Dec. 2023). https://www.bloomberg.com/graphics/2023-generative-ai-bias/

[11] OpenAI. 2023. *DALL·E 3 System Card.* Technical Report. https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf

[12] Luhang Sun, Mian Wei, Yibing Sun, Yoo Ji Suh, Liwei Shen, and Sijia Yang. 2023. Smiling Women Pitching Down: Auditing Representational and Presentational Gender Biases in Image Generative AI. https://doi.org/10.48550/arXiv.2305.10566 arXiv:2305.10566 [cs].

[13] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21).* Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3465416.3483305

[14] Nitasha Tiku, Kevin Schaul, and Szu Yu Chen. [n. d.]. These fake images reveal how AI amplifies our worst stereotypes. https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/

[15] Galit Wellner and Tiran Rothman. 2020. Feminist AI: Can We Expect Our AI Systems to Become Feminist? *Philosophy & Technology* 33, 2 (June 2020), 191–205. https://doi.org/10.1007/s13347-019-00352-z

[16] Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. 2012. Human computation tasks with global constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12).* Association for Computing Machinery, New York, NY, USA, 217–226. https://doi.org/10.1145/2207676.2207708