From Development to Dissemination: Social and Ethical Issues with Text-to-Image AI-Generated Art

Sharon Chee Yin Ho [†, *]

[†] Department of Computer Science and Software Engineering (CSSE), Concordia University, Montreal, Quebec, Canada

Abstract

Text-to-image generative artificial intelligence (AI) have made global news headlines for not only having the ability to generate high-fidelity artworks, but also for causing increased discussion on the ethicality of its impact on living artists, the automation and commodification of art production, the frequent non-consensual collection and usage of sensitive and copyrighted images as training data, and the routinely exhibited cultural and social biases in their generated outputs. In addition, there are concerns that open-sourced text-to-image generative AI models, such as Stable Diffusion, and techniques like Textual Inversion, allow for technical restrictions on the content subject matter to be removed and for generated images to be subject specific, which could be utilized as a new medium for disinformation and sexual or targeted abuse. Because ethical discussions on AI-generated art using text-to-image generative AI models have only come to light in the last quarter of 2022, academic research on the social and ethical implications of this technology have yet to be thoroughly explored. Therefore, it is imperative for research to be done on these implications with regards to the technological development, evaluation, perception, creation, and moderation of AI-generated artworks while text-to-image generative AI systems are still in the preliminary stages of public dissemination and adoption.

Keywords: ethics, bias, text-to-image generation, AI-generated art, open-source software, generative AI

1. Research Problem and Contributions

This abstract presents ongoing thesis work that aims to scientifically investigate the **social and ethical issues** regarding the technological development, evaluation, perception, creation, and moderation **of AI-generated art produced by text-to-image generative AI models.** The novelty of the thesis research is to address the amount of ethical awareness, or lack thereof, from practitioners, developers, and prospective researchers in the field of text-to-image generative AI, which to our knowledge, has not been investigated thus far. There is future intention to publish the findings in academic journals and to present them at conferences on "AI, Ethics and Society," or on other relevant research fields. In addition, the findings from this thesis have potential to build towards developing a workshop for computer science and software engineering graduate students to develop skills to assess and critique the impact of technology used to create text-to-image AI-generated art.

2. Research Questions and Approach

Three research questions (RQs) have been outlined for the thesis. This will be done through a triangulation [1] of the following research methods: 1) systematic literature review, 2) statistical analysis, and 3) interviews. These research methods will tackle the following dimensions and social aspects of text-to-image generative AI: **research and development (RQ3), implementation and application (RQ2), and creation, dissemination, and perception of AI-generated art (RQ1).** An upstream investigation on

[*] sharoncheeyin.ho@concordia.ca

the technology's outcomes to its inception and development will be done by answering RQ1 to RQ3. This methodological triangulation will develop a comprehensive understanding of the social and ethical issues of AI-generated art with text-to-image generative AI models.

The first research question (RQ1) will cover existing knowledge regarding the social and ethical implications of text-to-image generative AI models for AI-generated art. The findings for RQ1 will reflect existing viewpoints from the phenomenon's users, observers, and critics with regards to the technology's public perception, dissemination, and social repercussions. The second research question (RQ2) will investigate ethical values held by developers and practitioners in open-source communities on text-to-image generative AI models. The findings for RQ2 will reflect the viewpoints from people involved with the implementation and application of technology commonly used to produce AI-generated art. Lastly, the third research question (RQ3) will involve interviewing individuals who may become prospective researchers and developers of text-to-image generative AI: computer science and software engineering graduate students. These interviews will assess whether these students raise any social and ethical issues with text-to-image generative AI models. The findings for RQ3 will reflect the values and ethical awareness brought by prospective researchers towards the research and development of the technology. These findings may also identify gaps in computer science and software engineering higher education that could underprepare students with skills to assess and critique technologies used to produce text-to-image AI-generated art during the research and development of such technologies.

2.1. RQ1: What are the current acknowledged social and ethical issues with text-to-image AI-generated art?

A systematic literature review will be done to thematically identify the social and ethical implications of AI-generated art with text-to-image generative AI models in existing academic literature and scientific journalism.

2.1.1. Methodology for Academic Literature

Citing articles of academic publications on state-of-the-art research and development towards text-to-image AI models and techniques will be reviewed. The citing articles for each of these academic papers will be found using Google Scholar. To clarify, a citing article is a term and feature in Google Scholar to describe a publication that cites a published research work. From these citing articles, research publications focusing on the social and ethical issues produced by such models and techniques will be manually collected for analysis. Notable discussions raised in each of the collected publications on social and ethical issues will then be extracted and compiled thematically.

2.1.2. Methodology for Scientific Journalism

Articles covering social issues and ethical discourses on text-to-image AI-generated art from major and notable news publishers on technology and society-oriented topics will be collected and analyzed. Additional news articles from other notable sources found using Google News may also be further collected and analyzed as supplementary material. To identify emerging themes of social and ethical concerns during the text analysis, we will apply an inductive coding method and develop a coding scheme consisting of categories, definitions, examples, and coding rules as an instrument. Notable excerpts from these articles will be thematically and manually coded by the author of the thesis. To assert a level of confidence towards the author's manual coding, five sample articles will be blindly cross-checked by two individuals, using Cohen's kappa coefficients to assess the interrater reliability.

2.2. RQ2: What ethical values are found in GitHub discussions among open-source developers and practitioners of text-to-image generative AI models commonly used for AI-generated art?

Open-source software community discussions will be examined through various user activities, such as mined issue comments, pull request comments, and README files from various releases of Stable Diffusion [2]. In addition, this process will also be done for a derivative open-source and community-driven web implementation of the model with additional features that abstracts many of the technical components for the model so that it is presented in a user-friendly interface. Despite several mainstream text-to-image generative AI models having large-scale online communities of their own, only discussions amongst Stable Diffusion developers and practitioners will be solely analyzed. This is because it is the only mainstream state-of-the-art text-to-image generative AI model that has open-sourced its source code and has encouraged community engagements and contributions to the technology. This is contrary to how other state-of-the-art systems are proprietary and have not disclosed their source code, nor the dataset used. Findings for this research question will provide insight into whether open-source communities present ethical awareness and hold certain ethical values during the technical implementation and application of text-to-image generative AI models beyond what has been documented thus far in existing literature and discourse.

2.2.1. Methodology for Software Repository Mining and Analysis

User activities (issue comments, pull request comments, and README file) for the following three software repositories will be mined. This methodology is based on two existing research works by Winter and Salter [3], and Newton and Stanfill [4], both published in 2020, investigating moral values and discourse among developers in deepfake open-source communities.

1) **CompVis/stable-diffusion (i.e., Stable Diffusion v1)** [5]: With compute donation from AI startups Stability AI and Runway, and support from LAION, the Latent Diffusion Model developed by the Computer Vision & Learning Group (CompVis) lab at Ludwig Maximilian University of Munich was trained with a subset of the LAION-5B dataset [6], comprising of about 2.3 billion CLIP-filtered image-text pairs by parsing files in the Common Crawl dataset. There are four training versions associated with this model: v1-1 through v1-4.
2) **Stability-AI/stablediffusion (i.e., Stable Diffusion v2)** [7]: Stable Diffusion v2 is a fine-tuned configuration of the previous model architecture, Stable Diffusion v1. Most notable to v2 is the v2-2 release that has provided pretrained weights on a less restrictive NSFW filtering of the LAION-5B dataset.
3) **AUTOMATIC1111/stable-diffusion-webui (i.e., Stable Diffusion Web UI)** [8]: A browser interface implementation of Stable Diffusion. This project is community driven, and it is not created, monitored, nor officially endorsed by the creators of Stable Diffusion. It contains additional fine-tuning features for more styling and customization in the generated outputs, through the implementation of various open-source techniques and libraries.

Quantitatively, a frequency analysis of the most active commenters and most frequently used words will be done for data collected in each software repository. Qualitatively, a manual coding process will be done for each software repository for 50 issues and 50 pull requests with the most reactions for each reaction type (i.e., +1, -1, laugh, confused, heart, hooray, rocket, eyes). GitHub reactions reflect user sentiment and may be indicative of notable community discussions [9], which have potential to showcase various values and conflicts among developers. Timestamps for contributions and discussion posts can also be utilized to determine intersections between user discussions and user actions, such as commits, pushes, and pulls within the software repository itself. The qualitative analysis will then be utilized as a method to contextualize notable keywords extracted through the quantitative analysis to display their context in user discussions.

2.3. RQ3: How much awareness do prospective researchers (i.e., computer science and software engineering graduate students) have on social and ethical issues embedded in text-to-image generative AI models for AI-generated art?

Computer science (CS) and software engineering (SE) graduate students will be asked to interact with three different popular and publicly available state-of-the-art text-to-image generative AI models, with the intention to create AI-generated art. The findings from the interview sessions will provide insight into: 1) whether CS and SE graduate students raise any social and ethical issues with text-to-image AI-generated art, 2) whether any of the social and ethical issues on text-to-image AI-generated art mentioned by CS and SE graduate students are inside or outside existing literature, and 3) if any recommendations are raised by CS and SE students on strategies or solutions to mitigate social and ethical issues of text-to-image generative AI in the domain of AI-generated art.

2.3.1. Methodology for Interviews

Interview participants will interact with the following text-to-image generative AI models popularly used to create AI-generated artworks: 1) **Stable Diffusion Web UI** [8]: an open-sourced browser interface of Stable Diffusion, 2) **DALL-E 2** [10]: OpenAI's text-to-image generative AI model hosted as a beta web and cloud service, and 3) **Midjourney** [11]: a Discord server and bot service for text-to-image AI-generated art. These systems were chosen because they are popular implementations of different state-of-the-art text-to-image generative AI techniques using diffusion models [12] and CLIP image embeddings [13] that were developed primarily by researchers and academics. These models have also been studied in academic research for the content nature, and cultural and social biases of their generated outputs [14,15,16,17,18,19,20]. For this experiment, Stable Diffusion's v1-4 and v2-1 pretrained weights will each be used independently in conjunction with Stable Diffusion Web UI. This is so that participants can compare generated results with the same text prompts and configurations of their choosing. About 10 CS and SE students will be interviewed. They will be asked to provide a summary of their knowledge of the domain and their opinion on the technology, whether it be on the technical and/or non-technical aspects. They will then generate up to 20 different artworks with their choice of text prompt and configurations for each tool used. Afterwards, participants will describe their workflow, thought process, and reactions. The interviews will be transcribed and then analyzed for themes and concepts.

3. Progress to Date

To date, systematic literature review for RQ1 is in progress, while research towards RQ2 and RQ3 has not yet started. One survey paper on diffusion models by Yang et al. [21] and a literature review section in a research paper on the creativity of text-to-image generation for digital images and artworks with prompt engineering by Oppenlaender [22] were used to find the most notable and influential academic papers on state-of-the-art text-to-image generative AI techniques and models. From Yang et al.'s survey, 10 notable research papers were extracted, of which Stable Diffusion [2], Google's DreamBooth [23], and OpenAI's DALL-E 2 [10] are mentioned. From Oppenlaender's work, 5 research papers were extracted. Because OpenAI's DALL-E [24] and Textual Inversion [25] were not mentioned in either of the related works, these papers were also chosen for their relevancy. Therefore, to date, 17 academic papers are being utilized as a basis for finding citing articles on the social and ethical implications of text-to-image generative AI systems and their generated outputs. Articles in journalism are selected manual analysis from notable and well-established online publishers that cover technology and society-oriented topics and current events (i.e., Ars Technica, MIT Technology Review, TechCrunch, The Verge, VICE, and WIRED). These news sites are also covering the current and ongoing social and ethical public discourses with regards to text-to-image AI-generated art and technologies. Other news articles from notable news sources may also be further collected and analyzed as supplementary material.

Acknowledgements

This thesis is co-supervised[1] at Concordia University (Montreal, Quebec, Canada) by Dr. Tanja Tajmel (EDI Lab) and by Dr. Emad Shihab (Data-driven Analysis of Software Lab). This thesis is also funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Collaborative Research and Training Experience (CREATE) SE4AI program.

[1] As the supervisors of the author, Sharon Chee Yin Ho, I am aware of the acceptance of the paper titled "*From Development to Dissemination: Social and Ethical Issues with Text-to-Image AI-Generated Art*" in the GSS track of 2023 Canadian AI, and approve that it can be published on the PubPub open access online publication platform.

References

[1] L. Cohen, L. Manion, and K. Morrison. *Research methods in education*. Brit J Educ Stud. 2000.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. "High-resolution image synthesis with latent diffusion models". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684-10695.

[3] R. Winter and A. Salter. "DeepFakes: uncovering hardcore open source on GitHub". In: Vol. 7-4. Porn Studies. 2020, pp. 382-397.

[4] O. B. Newton and M. Stanfill. "My NSFW video has partial occlusion: deepfakes and the technological production of non-consensual pornography". In: Vol. 7-4. Porn Studies. 2020, pp. 398-414.

[5] *CompVis/stable-diffusion*. url: https://github.com/CompVis/stable-diffusion.

[6] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. "LAION-5B: An open large-scale dataset for training next generation image-text models". In: 36th Conference on Neural Information Processing Systems (NeurIPS 2022), Track on Datasets and Benchmarks. 2022.

[7] *Stability-AI/stablediffusion*. url: https://github.com/Stability-AI/stablediffusion.

[8] *AUTOMATIC1111/stable-diffusion-webui*. url: https://github.com/AUTOMATIC1111/stable-diffusion-webui.

[9] J. Boxer. "Add Reactions to Pull Requests, Issues, and Comments". The GitHub Blog. GitHub, 11 March 2016. url: https://github.blog/2016-03-10-add-reactions-to-pull-requests-issues-and-comments/.

[10] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. "Hierarchical text-conditional image generation with clip latents". In: arXiv preprint arXiv:2204.06125. 2022.

[11] *Midjourney*. url: https://www.midjourney.com/.

[12] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. "Deep unsupervised learning using nonequilibrium thermodynamics". In: Proceedings of the International Conference on Machine Learning. 2015, pp. 2256–2265.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning transferable visual models from natural language supervision". In: Proceedings of the International Conference on Machine Learning. 2021, pp. 8748–8763.

[14] L. Struppek, D. Hintersdorf, and K. Kersting. "The Biased Artist: Exploiting Cultural Biases via Homoglyphs in Text-Guided Image Generation Models". In: arXiv preprint arXiv:2209.08891. 2022.

[15] H. Bansal, D. Yin, M. Monajatipoor, and K.-W. Chang. "How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions?". In: arXiv preprint arXiv:2210.15230. 2022.

[16] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan. "Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale". In: arXiv preprint arXiv:2211.03759. 2022.

[17] S. Park, S. Moon, and J. Kim. "Judge, Localize, and Edit: Ensuring Visual Commonsense Morality for Text-to-Image Generation". In: arXiv preprint arXiv:2212.03507. 2022.

[18] Z. Liu, Y. Shin, B.-C. Okogwu, Y. Yun, L. Coleman, P. Schaldenbrand, J. Kim, and J. Oh. "Towards Equitable Representation in Text-to-Image Synthesis Models with the Cross-Cultural Understanding Benchmark (CCUB) Dataset". In: arXiv preprint arXiv:2301.12073. 2023.

[19] K. C. Fraser, S. Kiritchenko, and I. Nejadgholi. "A Friendly Face: Do Text-to-Image Systems Rely on Stereotypes when the Input is Under-Specified?". In: The AAAI-23 Workshop on Creative AI Across Modalities. 2023.

[20] Y. Zhang, L. Jiang, G. Turk, and D. Yang. "Auditing Gender Presentation Differences in Text-to-Image Models". In: arXiv preprint arXiv:2302.03675. 2023.

[21] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang. "Diffusion models: A comprehensive survey of methods and applications". In: arXiv preprint arXiv:2209.00796. 2022.

[22] J. Oppenlaender. "The Creativity of Text-to-Image Generation". In: Proceedings of the 25th International Academic Mindtrek Conference. 2022, pp. 192-202.

[23] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation". In: arXiv preprint arXiv:2208.12242. 2022.

[24] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. "Zero-shot text-to-image generation". In: Proceedings of the International Conference on Machine Learning. 2021, pp. 8821-8831.

[25] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. "An image is worth one word: Personalizing text-to-image generation using textual inversion". In: arXiv preprint arXiv:2208.01618. 2022.