The Revealed Preferences of Pre-authorized Licenses and Their Ethical Implications for Generative Models

Anonymous Authors[1]

1. Introduction

Generative AI models, which create text, images, audiovisual works, and other multimedia resembling human creativity introduce new challenges and prospects for artists and rightsholders. As AI systems become more sophisticated, their ability to autonomously generate works indistinguishable from those created by humans raises critical questions about copyright protection (originality, authorship, direct and indirect liability, defenses, and remedies), the industrial organization of creative industries, and ultimately social justice in a post-generative AI world.

Ongoing lawsuits between generative AI developers (such as OpenAI, Google, Meta, Stable Diffusion) and rightsholders (such as The New York Times, Getty Images, and author classes broadly) have amplified debate in the legal community about the propriety of ingestion of copyright-protected internet "data" without authorization and the generation of outputs remixing such "data" (Samuelson, 2023; 2024). While the legal question of what constitutes fair use is central to how these tools will develop, we focus here on the ethical question of what it would mean to respect artist and rightsholder preferences.

Initial attempts to understand what is ethical and desirable have either surveyed artists and rightholders (Lovato et al., 2024) or speculated about their preferences based on existing academic sources (Jiang et al., 2023; Attard-Frost, 2023; Latikka et al., 2023; Brunder). These surveys and analyses represent a skewed sample of artists/rightsholders and are only able to investigate the *stated* preferences of these artists. Yet, it is well-known that the stated preferences of individuals does not always align with their actions (De Corte et al., 2021; Craig et al., 2017). Thus, investigating the *revealed* preferences of artists and rightsholders would provide a clearer understanding of the ethics of ingestion and generative outputs.

Our work takes a step towards understanding this ethical landscape by examining the revealed preferences of creators as reflected in open and quasi-open licensing regimes. We analyze the most commonly used licenses by copyright holders of images in the Creative Commons and copyright holders of code in GitHub code repositories. We discuss the ramifications that these licenses might have on the existing generative AI training models. Finally, we discuss the technical affordances needed from the AI community to meet the artists' and rightsholders' license conditions.

License	Attribution Required	Remixing Allowed	Commercial Use Allowed
CC BY	✓	✓	✓
CC BY-NC	✓	✓	✗
CC BY-SA	✓	✓	✓
CC BY-NC-SA	✓	✓	✗
CC BY-ND	✓	✗	✓
CC BY-NC-ND	✓	✗	✗
CC0	✗	✓	✓

Table 1. Creative Commons licenses and their requirements.

2. Analysis

Images (Creative Commons) The Creative Commons (CC) is an international non-profit organization established in 2001 with the mission of enabling easier and more ethical use of copyrighted works. The CC organization generally reflects an open philosophy, although it offers users a range of pre-authorized licensing options. Table 1 presents the the six CC licenses. The salient features of the licenses for generative models are: waiver of rights (CC0), attribution (BY), authorization for editing (or remixing) (preparation of derivative works) (ND), authorization for commercial use (NC), and requirement to share alike (SA). Licensors may waive all rights or pre-authorize usage with one or more reservations of rights.

There are currently over 2.5 billion works across the internet that use CC licenses. These span text, audio, and images. Many of these works are scraped by Common Crawl (Patel & Patel, 2020), a commonly used tool for obtaining internet-based data, and subsequently used as training data for most generative models. For our analysis, we will focus on CC-licensed images curated by Openverse (`openverse.org`). Openverse sources over 700 million CC-licensed images

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

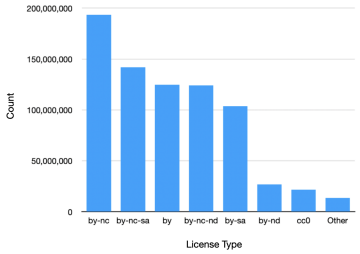

Figure 1. Breakdown of CC licensed used by images on Openverse.org.

License	Copyright Notice	Modification Allowed	Commercial Use Allowed	Same License
MIT	✓	✓	✓	✗
GPLv2	✓	✓	✓	✓
Apache	✓	✓	✓	✗
GPLv3	✓	✓	✓	✓
BSD 3-clause	✓	✓	✓	✗
BSD 2-clause	✓	✓	✓	✗
LGPLv3	✓	✓	✓	✓
AGPLv3	✓	✓	✓	✗

Table 2. Open Source Software licenses and their requirements.

from open APIs (e.g. Flickr) and Common Crawl. From this database of images sourced by Openverse and engineers at Openverse we calculate the breakdown of the CC licenses used (Figure 1).

The preferences revealed by this breakdown present important ethical considerations for training generative models on these images and generating outputs based on them. First, more than 90% of theses licenses require attribution. Additionally, the majority of licenses chosen do not allow commercial uses or remixing of the copyright images.

Open-Source Code (GitHub) GitHub is a cloud-based platform that allows developers to store Git repositories for their code. It is the most popular platform for open-sourcing code throughout the software community. We present the most commonly used open-source licenses and their differences in Table 2. Similar to CC-licensed images, we focus on the most salient permissions and conditions of the licenses for generative models: whether license and copyright notices are required, if modification of the code is allowed, and whether commercial use is allowed. Notably, all of the licenses allow commercial use.

As of May 2024, over 420 million repositories are stored on GitHub. Lacking direct API access, we focus on an analysis of license usage conducted by GitHub in 2015 (Balter). The breakdown presented in Figure 2 indicates a much different set of preferences than the CC-licensed breakdown. The MIT and Apache licenses have similar permissions and conditions in regards to those that will have the most impact on generative model training and outputs.

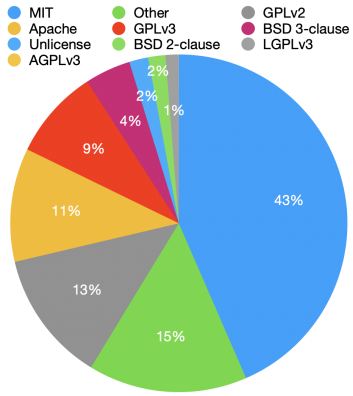

Figure 2. Breakdown of OSS licenses used by repos on GitHub.

3. Potential Impact on Generative AI

We briefly discuss some of the potential impacts based on these revealed preferences.

State of the art and commonly used generative models for image generation such as DALLE (Ramesh et al., 2021), Stable Diffusion (Rombach et al., 2022), and Midjourney (mid) were trained on images gathered from Common Crawl, many of which are CC-licensed images captured in the Openverse database. As evidenced in Figure 1, many of the CC-licensed images do not allow commercial use. What is currently unclear is how these individual licenses affect the overall use of a generative model. For example, does the use of these CC-licensed images with an NC license in training mean that these models are not allowed to be distributed for commercial use? If so, absent a determination that such training and associated outputs constituted fair use, OpenAI and StabilityAI are already in violation of these licenses due to the commercial nature of their activities. In contrast with the impact of CC on image generation models, the majority of licenses used for GitHub repos which have been cited as a primary source for training code generation models such as GPT-4 (Chen et al., 2021) and CoPilot (Gershgorn, 2021) allow commercial use. Thus, there is no impact on the monetization of GPT-4 and CoPilot for these purposes.

Across both regimes, it is clear that attribution is an important technical affordance, whether to satisfy the CC license conditions, to understand what copyright notices would need to be included for generated code, or to ensure that generated code has the same license in cases where the outputs are based on code that has GPL license. Similarly, many images use a `CC ND` license which prohibits generation of derivative works. How this applies to image generation models is dependent on how we view the mechanism by which outputs are created. If we view the mechanism as simply an interpolation of the training data, that interpretation would bar most outputs that are similar to images restricted by a `CC ND` license.

References

Midjourney. https://www.midjourney.com/home. Accessed: 2024-6-3.

Attard-Frost, B. Generative ai systems: impacts on artists & creators and related gaps in the artificial intelligence and data act. *Available at SSRN*, 2023.

Balter, B. [no title]. https://github.blog/2015-03-09-open-source-license-usage-on-github-com/. Accessed: 2024-6-10.

Brunder, K. M. " ai art and its implications current and future artists".

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code. July 2021.

Craig, A. C., Garbarino, E., Heger, S. A., and Slonim, R. Waiting to give: stated and revealed preferences. *Management Science*, 63(11):3672–3690, 2017.

De Corte, K., Cairns, J., and Grieve, R. Stated versus revealed preferences: An approach to reduce bias. *Health economics*, 30(5):1095–1123, 2021.

Gershgorn, D. GitHub and OpenAI launch a new AI tool that generates its own code. https://www.theverge.com/2021/6/29/22555777/github-openai-ai-tool-autocomplete-code, June 2021. Accessed: 2024-6-3.

Jiang, H. H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J., and Gebru, T. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 363–374, 2023.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Latikka, R., Bergdahl, J., Savela, N., and Oksanen, A. Ai as an artist? a two-wave survey study on attitudes toward using artificial intelligence in art. *Poetics*, 101:101839, 2023.

Lovato, J., Zimmerman, J., Smith, I., Dodds, P., and Karson, J. Foregrounding artist opinions: A survey study on transparency, ownership, and fairness in ai generative art. *arXiv preprint arXiv:2401.15497*, 2024.

Patel, J. M. and Patel, J. M. Introduction to common crawl datasets. *Getting structured data from the internet: running web crawlers/scrapers on a big data production scale*, pp. 277–324, 2020.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and 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.

Samuelson, P. Generative ai meets copyright. *Science*, 381 (6654):158–161, 2023.

Samuelson, P. Thinking about possible remedies in the generative ai copyright cases. *Communications of the ACM, forthcoming*, 2024.