“Heart on My Sleeve”: From Memorization to Duty

Anonymous Author(s) [1]

Abstract

Can machine learning models store protected content? This early-stage law review Article answers that question by developing several diffusion models and comparing generator outputs with training data in a measurement study. As the empirical data shows, the question is not *can* models contain copyrighted works, but *do* models contain copyright works. In turn, the Article argues for a series of duties to be placed on model owners.

Introduction. On April 4, 2023, an individual known as Ghostwriter977 uploaded a video to YouTube titled “Heart on My Sleeve” (Alexander, 2024). The song was an instant success, racking up 600,000 Spotify plays, 275,000 YouTube views, and 15 million TikTok views (Snapes, 2023). The track, featuring musical artists Drake and The Weeknd, included lyrical callbacks similar to those found on a Drake song, a Metro Boomin tagline, and the Weeknd’s signature falsetto. In a 21st-century twist, however, neither of the artists credited for the track had ever heard of it. Ghostwriter977 used generative artificial intelligence (AI) to nearly perfectly reproduce the voices of Drake and The Weeknd. Thirteen days after the song’s release, Universal Music Group filed Digital Millennium Copyright Act takedown requests to all sites hosting the content (Congress, 1998; Patel, 2023). “Heart on My Sleeve” died nearly instantly.

Julia Bausenhardt is a blogger and illustrator who creates nature sketches (Bausenhardt, 2023). Ms. Bausenhardt previews her illustrations on her website and makes the illustrations available for purchase in a variety of formats. Having recently heard about generative AI’s ability to copy artwork, Ms. Bausenhardt decided to see if generative AI had ingested her own work, thinking the possibility slim given her relative obscurity as an artist. To her surprise—and disconcert—Ms. Bausenhardt found “countless examples” of her work in the training datasets used by popular AI image generators (Spawning, 2022). Ms. Bausenhardt swiftly leveraged the self-governance tools available to her and politely requested that every one of her pieces of artwork be excluded from future training datasets. This process of data erasure may not show effects for over a year, and is dependent on a continued relationship between the opt-out mechanism and all of the model providers who used Ms. Bausenhardt’s protected expressions.

The juxtaposition of these stories is the impetus for this Article. On the one hand, the embattled forces behind Napster are, yet again, sounding alarm bells, this time focusing on no-name artists’ ability to create works that compete with hits that can cost over a million dollars to make (Chace, 2011). On the other hand, millions of no-name artists’ work is being ingested by models, used in training, and then sold by companies, like openAI, that are now valued at over 80 billion dollars (Pequeño IV, 2024). In both cases, the least advantaged individuals are the ones experiencing the negative externalities brought on by AI. The lever used to encourage creativity in this context—copyright—is not working. The root of the problem comes from an argument that machine learning models do not store protected content, and therefore cannot infringe, meaning there is no need for permission or compensation. This position, however, disincentivizes creativity, and, because the lifeblood of machine learning models is high-quality data, incentives model rot (synthetic data causing model collapse (Srivastava et al., 2017; Bellovin et al., 2019)).

This Article will take a closer look at the question: Do machine learning models store protected content? The Article answers that question in the affirmative, noting that this question is both simple and complex. It is simple because, factually, machine learning models can memorize content, and therefore can contain copies of works from training data. It is complex because: (1) memorization may occur infrequently; (2) what the model possesses may be construed as ideas, which are not protected by copyright; and (3) the process of training may be considered learning, which, like facts, are not protected by copyright. The Article engages with these questions through a series of empirical measurements.

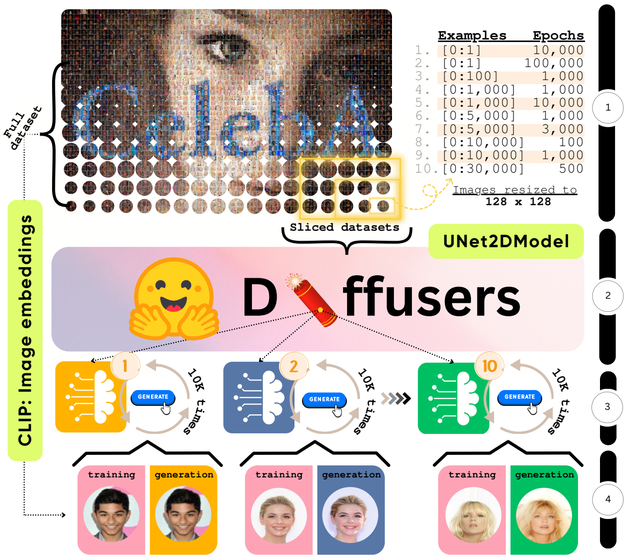
Background. Prior work has engaged with the question of whether models store protected content (Lee et al., 2024;

[1] [redacted]. Correspondence to: [redacted] <[redacted]>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Cooper & Grimmelmann, 2024; Samuelson, 2023), but few have deeply investigated these questions using empirical data. For starters, OpenAI, has clearly declared that "[m]odels do not contain or store copies of information that they learn from (OpenAI, 2024)." This finding has, in some ways, been echoed by legal scholars. On the far end of the spectrum, there are claims that image generators are mere collage machines parroting protected content (Bozard, 2024; Sobel, 2024; Vincent, 2023; Sobel, 2024). On the other, scholars argue that models can never infringe because models only ever possess floating-point numbers (Lindberg, 2022; Murray, 2024; 2023). Still, others have softened these positions, though remain skeptical of model infringement and consider it a theoretical exercise because, it is claimed, models learn latent features, (i.e., "informational patterns") or infringe only when prompted (Sag, 2023; Lemley & Casey, 2020; Lindberg, 2024; Bracha, 2024; Lemley, 2024). Contrary to these arguments, years of computer science research has shown that models do memorize training data, either at the detriment of privacy or, more recently, at the infringement of copyright (Carlini et al., 2023; Meehan et al., 2020; Somepalli et al., 2023a; Mireshghallah et al., 2020; Somepalli et al., 2023b; McCoy et al., 2023; Feng et al., 2021; Feldman & Zhang, 2020). This set of work, largely aimed at either reducing overfitting or conducting real-world measurements on as-deployed models, helps identify the theoretical possibility of infringement, but has a few limitations: (1) production models frequently change, and third-party researchers have very little control or knowledge over the full corpus of training data and post-generation filtering; (2) assessing model output per copyright is not the same as a privacy analysis (e.g., inversion attacks may not mean copyright infringement); and (3) prompting is often used to produce copyrighted works, muddying a legal analysis. Our empirical work controls the entire training dataset, assesses outputs in terms of copyright infringement, and builds models that do not rely on user input. In this way, our results are precise to the question of whether and when models store offending content.

Methods and Architecture. This Article addresses the complexity of copyright and generative AI by discussing empirical work aimed at model memorization. Several machine learning models—unconditional image generators—were built with the purpose of assessing their ability to store protected works. These models are similar to those found in production, like StableDiffusion or DALL-E (Ramesh et al., 2022; Rombach et al., 2022). Our pipeline takes small slices of the CelebA dataset (Liu et al., 2018), uses that data to train a series of diffusion models, and then checks the output of those models against training data, made efficient via CLIP embeddings to create a finite search space of neighboring images (Almeida, 2024). An overview of the pipeline may be seen in Figure 1.



	Examples	Epochs
1.	[0:1]	10,000
2.	[0:1]	100,000
3.	[0:100]	1,000
4.	[0:1,000]	1,000
5.	[0:1,000]	10,000
6.	[0:5,000]	1,000
7.	[0:5,000]	3,000
8.	[0:10,000]	100
9.	[0:10,000]	1,000
10.	[0:30,000]	500

Figure 1. To start, the CelebA dataset ① is broken into smaller pieces (Liu et al., 2018). We then train ② diffusion models ($n = 10$), generate ③ 10K images, and ④ compare generations to 1K CLIP neighbors using similarity metrics (Müller et al., 2020).

For each of the models' outputs (10K generations) we look up the closest 1,000 images per CLIP embeddings. When looping through neighbors, we keep track of the most similar image per the Structural Similarity Index (SSIM) (Brunet et al., 2011). If the SSIM is higher than the previously highest SSIM score, we update the similarity score and measure that generation per a suite of similarity metrics (i.e., SSIM, RMSE, PSNR, FSIM, SRE, SAM, and UIQ) (Müller et al., 2020). This gives us a rudimentary assessment of substantial similarity.

Discussion and Conclusion. Although our results are preliminary, we find blatant cases of copying (i.e., memorization (Cooper & Grimmelmann, 2024)) when a model is trained on a small number of examples ($\leq 1,000$). We also find a non-trivial amount of copying in larger datasets. Storing protected content inside a neural network, therefore, does not erase the protected content. We argue that this finding calls for a duty to be placed on model owners, a duty extending to both input and output. For input, model owners should (for starters) maintain detailed training data records, an archival hyperlink to the source of the data, information on who was responsible for obtaining the data, and a reasonable amount of effort put into ascertaining who may be the owner of the data. Likewise, there should be a duty triggered on output: is the output similar to any data from training; if so, how similar is the data, and is this a case where a creator may deserve compensation? Though imperfect, these duties will better balance the state of creativity today, encouraging progress of the sciences and useful arts.

Acknowledgements and Impact Statement

This paper was not supported by any specific grants. The impacts of this work pertain to advancing the ethical and responsible use of machine learning.

References

Alexander, A. *"Heart on My Sleeve": An AI-Created Hit Song Mimicking Drake and The Weeknd Goes Viral*. SAGE Publications: SAGE Business Cases Originals, 2024.

Almeida, D. R. Clip embeddings to improve multimodal rag with gpt-4 vision. https://cookbook.openai.com/examples/custom_image_embedding_search, 2024.

Bausenhardt, J. How ai is stealing your art. https://juliabausenhardt.com/how-ai-is-stealing-your-art/, 2023.

Bellovin, S. M., Dutta, P. K., and Reitinger, N. Privacy and synthetic datasets. *Stan. Tech. L. Rev.*, 22:1, 2019.

Bozard, Z. What does it mean to create art? intellectual property rights for artificial intelligence generated artworks. *South Carolina Journal of International Law and Business*, 20(1):10, 2024.

Bracha, O. Generating derivatives: Ai and copyright's most troublesome right. *North Carolina Journal of Law and Technology*, 25(3), 2024.

Brunet, D., Vrscay, E. R., and Wang, Z. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, 21(4):1488–1499, 2011.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Chace, Z. How much does it cost to make a hit song? *Planet Money Blog*, 30, 2011.

Congress, U. Digital millennium copyright act. *Public Law*, 105(304):112, 1998.

Cooper, A. F. and Grimmelmann, J. The files are in the computer: Copyright, memorization, and generative ai. *arXiv preprint arXiv:2404.12590*, 2024.

Feldman, V. and Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Feng, Q., Guo, C., Benitez-Quiroz, F., and Martinez, A. M. When do gans replicate? on the choice of dataset size. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6701–6710, 2021.

Lee, K., Cooper, A. F., and Grimmelmann, J. Talkin"bout ai generation: Copyright and the generative-ai supply chain (the short version). In *Proceedings of the Symposium on Computer Science and Law*, pp. 48–63, 2024.

Lemley, M. How generative ai turns copyright law upside down. *Science and Technology Law Review*, 25(2), 2024.

Lemley, M. A. and Casey, B. Fair learning. *Tex. L. Rev.*, 99: 743, 2020.

Lindberg, M. Applying current copyright law to artificial intelligence image generators in the context of anderson v. stability ai, ltd. *Cybaris®*, 15(1):3, 2024.

Lindberg, V. Building and using generative models under us copyright law. *Rutgers Bus. LJ*, 18:1, 2022.

Liu, Z., Luo, P., Wang, X., and Tang, X. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15 (2018):11, 2018.

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., and Celikyilmaz, A. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11:652–670, 2023.

Meehan, C., Chaudhuri, K., and Dasgupta, S. A non-parametric test to detect data-copying in generative models. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Mireshghallah, F., Taram, M., Vepakomma, P., Singh, A., Raskar, R., and Esmaeilzadeh, H. Privacy in deep learning: A survey. *arXiv preprint arXiv:2004.12254*, 2020.

Müller, M. U., Ekhtiari, N., Almeida, R. M., and Rieke, C. Super-resolution of multispectral satellite images using convolutional neural networks. *arXiv preprint arXiv:2002.00580*, 2020.

Murray, M. D. Generative ai art: Copyright infringement and fair use. *SMU Sci. & Tech. L. Rev.*, 26:259, 2023.

Murray, M. D. Comment to the united states copyright office re: Notice of inquiry on copyright and artificial intelligence, questions 18 and 21 (authorship of works created with the assistance of generative ai). 2024.

OpenAI. How chatgpt and our language models are developed. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed, 2024.

Patel, N. Ai drake just set an impossible legal trap for google. *The Verge, https://www. theverge. com/2023/4/19/23689879/ai-drake-song-google-youtube-fair-use*, 2023.

Pequeño IV, A. Openai reaches $80 billion valuation in venture firm deal, report says. https://www.forbes.com/sites/antoniopequenoiv/2024/02/16/openai-reaches-80-billion-valuation-in-venture-firm-deal-report-says, 2024.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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.

Sag, M. Copyright safety for generative ai. *Forthcoming in the Houston Law Review*, 2023.

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

Snapes, L. Ai song featuring fake drake and weeknd vocals pulled from streaming services. *The Guardian. Artikkeli*, 18:2023, 2023.

Sobel, B. Elements of style: Copyright, similarity, and generative ai. *Harvard Journal of Law & Technology, Forthcoming*, 38, 2024.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023a.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023b.

Spawning, A. Have i been trained?, 2022.

Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., and Sutton, C. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017.

Vincent, J. Ai art tools stable diffusion and midjourney targeted with copyright lawsuit. https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart, 2023.