“Heart on My Sleeve”: From Memorization to Duty

Nathan Reitinger [1]

Abstract

Can a machine learning model infringe on a copyright—do machine learning models store protected content? This work-in-progress law review Article focuses on empirical data developed, in part, to answer that question: yes. A set of unconditional image generators, diffusion models ($n = 14$), are trained on small slices of a dataset consisting of celebrities’ faces. The synthetic data output from these generators is then compared to training data using a variety of similarity metrics. As the empirical data shows, the question is not *can* models contain copyrighted works, but *do* models contain copyright works. In some cases, there is a 99% chance that a model will generate an image nearly identical to its training data; in other cases, even after 10,000 generations, a model does not produce any images that may be considered identical (though finding similarity is nonetheless possible). This Article uses the empirical data to argue for a series of duties to be placed on model owners.

1. Introduction

On April 4, 2023, an individual known only 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 signature “Metro Boomin” tagline, and the Weeknd’s unmistakable falsetto. In a 21st-century twist, neither of the artists credited for the track had ever heard of it.

Ghostwriter977 used generative artificial intelligence (AI) to nearly perfectly replicate the voices of Drake and The Weeknd. In a seemingly Napster-inspired reaction, just thirteen days after the song’s release, Universal Music Group filed Digital Millennium Copyright Act takedown requests to all sites hosting the song (Congress, 1998; Patel, 2023). “Heart on My Sleeve” died[1] 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, for all of her pieces, that the artwork be excluded from future training datasets. She also updated her website’s `robots.txt` file to request—a voluntary request—that no AI-purposed scrapers visit her website content.[2] This process of data erasure is voluntary, may not show effects for over a year, and is dependent on a continued relationship between the opt-out provider 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 and then sold by companies, like openAI, which is now valued at over 80 billion dollars (Pequeño IV, 2024). In both cases, the least advantaged

[1] University of Maryland. Correspondence to: Nathan Reitinger <nathan.reitinger@gmail.com>.

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

[1] In true Napster fashion, the song’s clones may still be found online (Dodes, 2002).

[2] Martijn Koster proposed the idea of `robots.txt` back in 1994, supposedly after Koster’s website slowed to a crawl during a bot-driven denial-of-service attack (Elmer, 2008). `Robots.txt` opens a communication channel between bots, who are scraping the web to ingest content, and website owners, who might not want their content ingested. This communication channel tells a bot which files (i.e., web pages) it can visit and which files are off-limits (Monaco & Woolley, 2022).

individuals are the ones experiencing most of the negative externalities created by AI.

One of the levers used to encourage creativity in this context—copyright—is not working. Part of the problem comes from an argument that has been picking up steam in the legal literature: Machine learning models do not store protected content, and therefore, cannot infringe. In other words, model-washing is used to transform protected content into unprotected content. This position, however, has serious consequences. It disincentivizes creativity given that generative AI can reproduce an artist's work after just a few training examples, and, because the lifeblood of machine learning models is high-quality data, it incentives model rot (i.e., synthetic data, in extreme cases, causes model collapse (Srivastava et al., 2017; Bellovin et al., 2019)). Moreover, model washing fails on technical grounds vis-a-vis memorization, as recent literature has shown (Carlini et al., 2023; Cooper & Grimmelmann, 2024).

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, using a series of empirical measurements.[3] To be sure, this question turns out to be 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. At the same time, it is complex because: (1) memorization may occur infrequently or not at all; (2) what the model possesses may be construed as ideas, that are not protected by copyright; and (3) the process of training may be considered learning, which, like facts, are not protected by copyright. This work-in-progress piece will consider only the first of these issues, leaving the other two to future developments. The conclusion of these measurements, the key lesson from complexity plus ambiguity, drives a proposed solution: duties placed on model owners related both to training data management and royalty-related payments per specific model output.

The Article proceeds as follows. After Section 2's discussion of the legal and technical background, Section 3 will provide details on the architecture used to measure a variety of diffusion models for their ability to generate copyright-offending output. Section 4 presents the results of those measurements. The Article then offers a preliminary discussion, Section 5, regarding the results and why the findings necessitate a set of duties to be placed on model owners.

2. Background

Copyright. In the United States, works of expression fixed in a tangible medium find themselves protected by copyright—a series of restrictions controlled by an author.[4] These restrictions relate to rights of reproduction, distribution, performance, and dissemination (Fromer & Sprigman, 2024). The rights are created to promote "the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries (US Constitution, Article I, Section 8, Clause 8, 1787)." The infringement of these rights occurs when: (1) a valid copyright exists and (2) one of the rights owed to an author was violated (Fruehwald, 1992). In the context of models possessing copyrighted works, the question is likely one of reproduction: Has a copyrighted work been copied?

"Copying" turns on direct or indirect evidence of copying (Asay, 2022; Rogers, 2013). Direct evidence would be relatively non-contested; you watched someone copy your work or someone admits to copying from your work. Indirect evidence of copying is a much more difficult question and generally occurs when it is shown that a defendant had (1) access to the protected work and (2) the copied work was substantially similar to the protected work. Substantial similarity is one of those doctrines infamous in the legal literature for its resistance to clarity (Lim, 2021). Although many legal tests exist to identify what is or is not substantially similar, it is nonetheless difficult to provide assurances a priori regarding the outcome of such a determination (Cohen, 1986).

To assess copying at the scale used in this work, therefore, the piece sidesteps the substantial similarity question and instead looks to striking similarity: when copies are so nearly indistinguishable to the point where the similarities preclude the possibility of independent creation (Latman, 1990; Autry, 2002; Fruehwald, 1992). In the case of nearly identical works, access is (mostly (Lanzalottie, 2002)) assumed, and the analysis becomes scalable given an ability to hunt for near duplicates as opposed to specific elements or stylistic considerations. Looking at striking similarity, therefore, permits the creation of a floor, but not a ceiling—a model's generations may engage in more examples of copying than identified, but the instances where copying is occurring are more clear-cut.

What do legal scholars say about model possession? OpenAI was one of the first AI-reliant companies to take a strong stance on models and copyright: "Models do not contain or store copies of information that they learn from . . . as a model learns, some of the numbers that make up the model change slightly to reflect what it has learned.

[3] All models used in this study are available at: `https://huggingface.co/nathanReitinger`.

[4] "The Congress shall have Power . . . [t]o promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries" (US Constitution, Article I, Section 8, Clause 8, 1787).

(OpenAI, 2024)." Some scholars[5] have embraced OpenAI's position and argue that models can never infringe because models only ever possess floating-point numbers (Lindberg, 2022; Murray, 2024; 2023). Opposing this, there are (debated (Sobel, 2024)) claims that image generators are mere collage machines that can only parrot protected content (Bozard, 2024; Vincent, 2023). Still others have found a middle ground, though remain skeptical because infringement is seen as a theoretical exercise; models learn latent features, "informational patterns," or only infringe when prompted by a user (Sag, 2023; Lemley & Casey, 2020; Lindberg, 2024; Bracha, 2024; Lemley, 2024). In turn, the general legal position has leaned away from infringement and toward either fair use or theoretical, but not practical, infringement.

How has possession been analyzed technically? Contrary to some of the legal arguments, years of computer science research has shown that models can and do infringe—called memorization in the technical literature—on training data, either at the detriment of privacy or, more recently, 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; Gu et al., 2023). This set of work, largely aimed at reducing overfitting (i.e., a model learns from examples too well and fails to generalize that knowledge, resulting in poor performance on new data (Howard & Gugger, 2020; Reitinger, 2024)) or conducting real-world measurements on production models, helps identify the theoretical possibility of infringement, but has three primary limitations when it comes to a legal application.

First, analyzing production models causes both a red herring and incomplete measurements. On the one hand, labeling problematic output in production models incentivizes quick-fix solutions, like filtering out undesired generations, over slower solutions, like filtering out undesired training data. As adversarial machine learning has shown, the quick solution is likely to be leaky, as tricking "protected" models into producing undesired output is often still possible (Terekhov et al., 2023). Additionally, finding examples of problematic output flags a problem, but does not provide an overall understanding of that problem. The scale of the problem is still unknown, and that scale may be important for a legal analysis. Likewise, if the model requires prompting, as most production models do, then the problem inherits a liability wrinkle given the required user input to produce offending content.

Second, prior work in understanding problematic model output has focused on privacy concerns (Carlini et al., 2021), but these are not the same as copyright concerns. True enough, areas of copyright, like substantial similarity, may be riddled with uncertainty, but copyright law nonetheless provides a rich literature to draw from, one that is not the same as the privacy literature. Moreover, a privacy violation is not always a copyright problem; for example, learning an image was used to train a model in an inversion attack may be a privacy issue, but that does not mean the model has stored protected content (Shokri et al., 2017; Khosravy et al., 2022).

Third, many works in this domain define memorization relative to an entire dataset: A generated image is considered a duplicate if it is more similar to an image in training than most images in the dataset (sometimes called (k,ℓ,δ)-Eidetic Memorization) (Yoon et al., 2023; Gu et al., 2023; Carlini et al., 2023; 2021). Even in cases of measurement-style assessments of diffusion model copying (Gu et al., 2023), this definition is used. The problem is that this definition may work well when analyzing a model for a propensity to overfit or privacy implications, but it is less well suited in a legal similarity analysis—given the non-trivial amount of non-similar, false positives along with a focus on overall model properties. The legal analysis would prioritize a reduction of false positives and highlight single instances of copying (Webster et al., 2023; Cooper & Grimmelmann, 2024).

This Article operates in the computer science domain, but with an eye toward copyright. The entire training dataset is controlled, meaning that assessments may be made on the model and training data as a whole (e.g., searching the entire dataset for the "most" similar image to a generation is possible, see Figure 2, and so is a statement about how often the model generates possibly infringing content). The classification of memorization uses pixel-level similarity metrics to reduce false positives, and the models are unconditional to remove any inclination that the user is somehow responsible for what a model possesses. The next section introduces the methods used to assess whether models store protected works.

3. Methods and Architecture

Several machine learning models—unconditional image generators—were built to assess 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), but differ in terms of train-

[5] It is worth mentioning that many varieties of legal analysis, as is common with new technologies (Reitinger, 2015b), have been applied to machine learning. From fair learning to the copyrightability of output to the international perspective, generative AI is giving pause to a growing number of intellectual property scholars (Alhadeff et al., 2024; Lemley & Casey, 2020; Abbott & Rothman, 2023; Gao et al., 2022; Gillotte, 2019; Lee et al., 2024). Few of these works, however, and what is the raison d'être for this Article, engage with an empirical analysis of models and copyright.

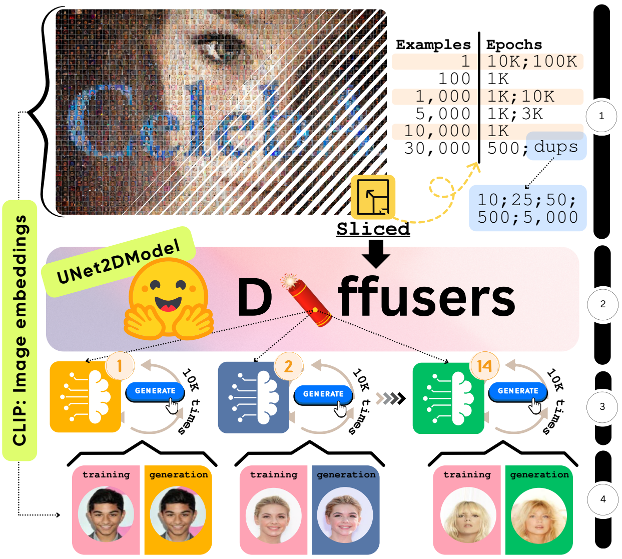

Figure 1. The CelebA dataset (Liu et al., 2018) is broken into smaller pieces ① (variations on dataset size, training epochs, and number of duplicates “dups” found in the slice). Diffusion models ② are trained ($n = 14$) and set to generate ③ 10K images. The generated images ④ are compared to 1K CLIP neighbors using similarity metrics like SSIM (Müller et al., 2020).

ing time, training dataset, hyperparameter tuning, and other variables. Although these variables differ, using a similar architecture to production models (i.e., `UNet2DModel`) provides a way to make a comparison without the nation-state-level resources required to train production models. Moreover, having a smaller dataset paired with smaller computing power has parallels to using a larger dataset with more computing power.

A pipeline was used to create a testing ground for model memorization, as shown in Figure 1. Slices ① of the CelebA dataset (Liu et al., 2018) are used for training. Examples per slice were picked to provide a variety, both with nonsensical example counts (e.g., one example) and large example counts (e.g., 30,000). Duplicate images were included in one of the models that had the least likely chance of producing offending content. This would provide a good test case for the idea that duplicates can taint an otherwise “clean” (i.e., a model producing no protected content) model (Webster et al., 2023). The number of duplicates was picked arbitrarily (i.e., one of the images was copied in the model 10, 25, 50, 500, and 5,000 times). Likewise, training epochs were varied, though limited by resource constraints (i.e., a limit of three days was set on training time, to create realistic limitations on training). Training images were reduced to 128x128 pixels. All models ② are Denoising Diffusion Probabilistic Models (Ho et al., 2020) and use the vanilla HuggingFace implementation. Each model has a batch size of 16, a learning rate of $1\text{e}{-4}$, and 500 warm-up steps. The models used six output channels (e.g., 128, 128, 256, 256, 512, 512) with two ResNet layers per UNet block.

Each model was set to unconditionally generate 10,000 images ③. The number of generations was picked pursuant to prior work, and provides a relatively stable estimate of how often a model may produce offending images (Somepalli et al., 2023a). Images generated by the models ④ were compared to all images in the CelebA dataset using CLIP embeddings (Radford et al., 2021). Each image in CelebA was processed with CLIP to produce a vector that could be inserted into a local `faiss` database (Douze et al., 2024). This allowed for fast image similarity lookup. For each new generation, the new image went through the same process, creating a CLIP embedding and finding pixel-similar neighbors in the `fiass` database (Almeida, 2024). Up to 1,000 neighbors were considered, with each image assessed using the Structural Similarity Index (SSIM) (Wang et al., 2004; Brunet et al., 2011).

When are images similar? The computational metric used by the pipeline to define striking similarity, SSIM, takes into account variations in images that may be missed with traditional hashing-based image comparisons (Ke et al., 2004; Nilsson & Akenine-Möller, 2020). For each generation, the pipeline identified “most similar” images from CLIP embeddings ($n = 1,000$). Then, for each pair of images, the SSIM score was calculated. Every time a higher score was identified (i.e., closer to perfectly similar), a log was updated, and upon updating the log, a suite of similarity metrics was applied (i.e., SSIM, RMSE, PSNR, FSIM, SRE, SAM, and UIQ) (Müller et al., 2020). In this way, the more recent additions to the log represent more similar images between training and generation, per SSIM. To be sure, a looser, semantic similarity concept could have been used to assess the substantial similarity standard. However, these assessments would be much more theoretical and produce false positives, as substantial similarity is difficult to define computationally in a way that would be convincing to a judge or jury (Scheffler et al., 2022).

The exact value of SSIM used to deem two images as strikingly similar was $\left(SSIM \geq 0.981\right)$. This range was picked by manually looking through image comparisons in varying ranges of similarity (e.g., > 0.980 *and* $<= 0.982$). As shown in Figure 2, some number of false positives (i.e., an SSIM score more than .0981 where the images are not nearly identical) existed in all ranges. Therefore, SSIM scores of at or more than 0.981, though false positives were found, was deemed reasonable. Examples of false negatives could be observed as well. SSIM scores above 0.970 and below 0.980 included images that were likely substantially similar, though given how few of these existed, the cutoff was set higher.

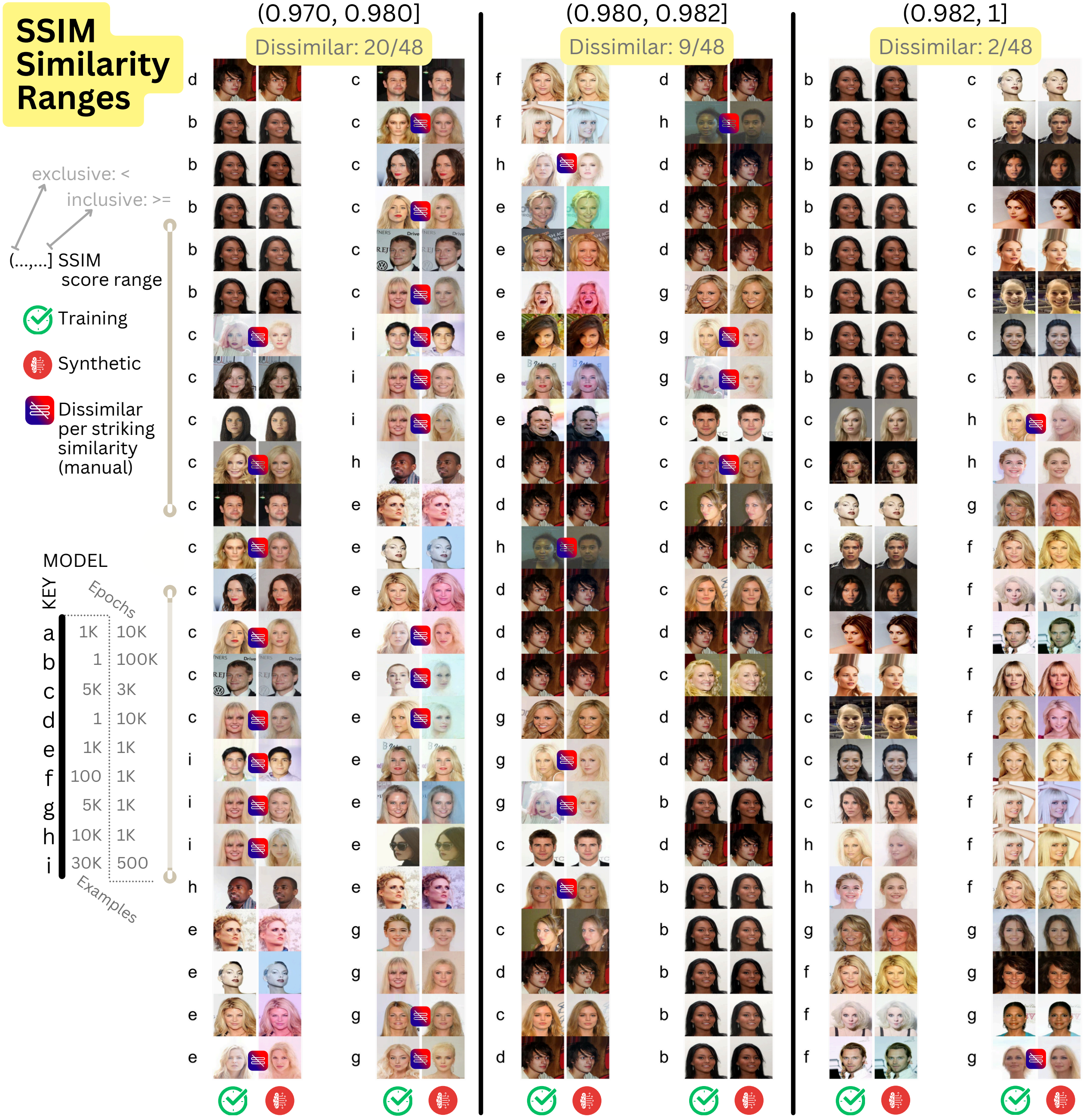
SSIM Similarity Ranges
exclusive: <
inclusive: >=
(...,...] SSIM score range
✓ Training
Synthetic
Dissimilar per striking similarity (manual)
MODEL
KEY
Epochs
Examples
a 1K 10K
b 1 100K
c 5K 3K
d 1 10K
e 1K 1K
f 100 1K
g 5K 1K
h 10K 1K
i 30K 500
(0.970, 0.980]
Dissimilar: 20/48
d b b b b b c c c c c c c c c c i i i h e e e e
c c c c c c i i i h e e e e e e e e e e g g g g
(0.980, 0.982]
Dissimilar: 9/48
f f h e e e e e e d d h d d d g g g c c c d c d
d h d d d g g g c c c d c d c d d b d b b b b b
(0.982, 1]
Dissimilar: 2/48
b b b b b b b b c c c c c c c c c c h h g f f f
c c c c c c c c h h g f f f f f f f f f g g g g

Figure 2. Showing the comparison between training data (left column) to generated image (right column) per SSIM score range. Models are tagged with letters *a* through *i* according to the key noted on the left. Dissimilarity (i.e., not striking similarity) was determined by manual review. Notably, although there are a high number of dissimilar images in the first range, > 0.970 *and* $<= 0.980$, the range nonetheless includes strikingly similar images—in fact, 28 out of 48 images may be determined to be strikingly similar: 58%. This shows that while similarity may be a more clear-cut case with higher SSIM scores, lower scores do not necessarily mean a generated image is not memorized. On the other hand, even in the higher SSIM score ranges, > 0.980 *and* $<= 0.982$, there exist image pairs that are not strikingly similar. This means that a purely computational test, like an SSIM score threshold, is unsuited for a thorough analysis. Some type of human review, aided by SSIM score cutoffs, is likely necessary. Lastly, it is important that even in the most similar case, with pairs that have an SSIM score above 0.982, there exist generations that are likely not strikingly similar. Though these generations may be determined to be substantially similar, this adds to the importance of including a human review at some point during a similarity determination.

In short, SSIM was used to identify generations that are nearly indistinguishable from training data to the point where the similarities preclude the possibility of independent creation. However, a finding of nearly identical should only be viewed as a proxy for similarity—see Figure 2—as there may exist generated images that are similar to training data, but, for one reason or another, had lower SSIM scores. In this way, the pipeline produces a floor rather than a ceiling.

3.1. Assumptions and limitations

Several assumptions regarding the empirical measurements should be highlighted. First, images in the training dataset are assumed to be protected by copyright. While this may not be the case for all images in a production model's training data, and is surely not the case for the training data in the CelebA dataset, this assumption is made for the sake of simplicity in identifying problematic generations and scale in making these assessments for 10,000 generations over 14 models. Additionally, varying definitions of memorization, regurgitation, and copying appear in the literature (Cooper & Grimmelmann, 2024). Here, memorization is defined as strikingly similar: the case where a generation is nearly identical to an image found in training data. To define this term technically, we use an SSIM value of 0.981 or greater. Although this technical definition is arbitrary, it is necessary to reduce false positives and analyze models at scale. Finally, for each generated image, we only consider the similarity of up to 1,000 CLIP-encoding neighbors. A more thorough search would involve the entire training dataset, although this approach would add a heavy computational burden and may not provide a worthwhile benefit on balance. More importantly, the duties discussed in Section 5 in part require companies to engage in a similar output-similarity assessment. In production, that type of assessment must occur nearly instantaneously. Using a `faiss` database with CLIP embeddings provides speed and is efficient, making this burden more practical.

4. Results

Two primary findings are evidenced by the empirical measurements. First, models memorize potentially protected content. As shown in Figure 3, nearly all models generated images that were strikingly similar to images found in training data. This was not only true in the "sanity" case where the training dataset consisted of a single image, but it was also true when a model trained on a large number of images (e.g., 10,000).

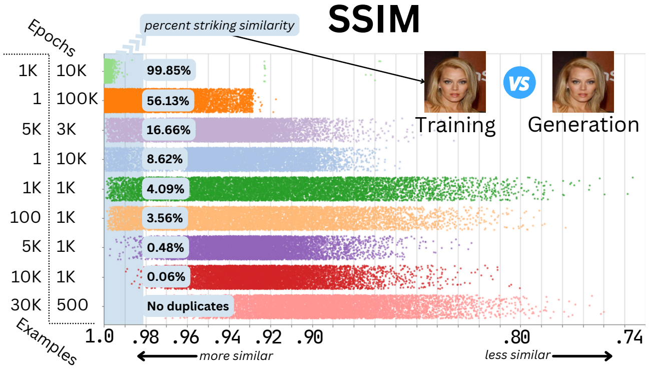

Figure 3. Each dot represents a generated image, per SSIM value, when compared to the most similar image found in training data; with 10,000 dots per model, noted on the y-axis as variations in example and epoch count. The light blue band represents similarity scores that signify near-perfect copies (i.e., ≥ 0.981). Percentages relate to the number of generations that were offending out of the total number of generations. In the most offending case (i.e., 1,000 training examples at 10,000 epochs), ∼100% of the model's outputs are nearly identical; in the least offending case (i.e., 30,000 examples with 500 epochs), no output could be found that constituted a copy—per pixel-level similarity.

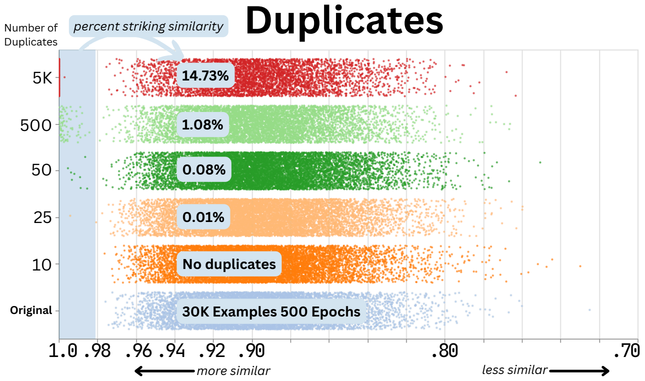

Figure 4. Showing the effect that duplicates (i.e., the same image found multiple times in training data) have on the model's likelihood of memorization. The original model, 30,000 examples at 500 epochs, did not memorize any training data, but if there are 25 duplicate images in the training data, then there exists a chance (albeit small: .01%) that the model will engage in memorization.

The measurements further identified a wide variety in the likelihood of memorization per model. Some models had a fairly high chance of producing an image copied from training data (e.g., ∼100% in the case of a model trained on 1,000 images). Other models produced no images that could be said to be *strikingly* similar to training data (though it is notable that there may be substantially similar images produced). In other words, even though a model may produce—from storage—a memorized training data image, the likelihood of this happening is affected by many factors. Clear divisions between factors like example counts or number of epochs did not drive clear differences in the likelihood of memorization.

Second, as Figure 4 visualizes, duplicates have a large impact on a model's likelihood of memorization. If a mere 25 images in a dataset are duplicated, or are similar enough to be considered a duplicate, then an otherwise "clean" model may nonetheless produce a potentially infringing image. This finding is not novel (Webster et al., 2023; Ren et al., 2024), but the measurements conducted here occurred using an unconditional generator, which would have less of a likelihood of memorization.

It is also notable that a high number of duplicated images (i.e., up to 16% in the case of 5K duplicates) can dominate the output of a model, leading to nearly 15% of a model's outputs being problematic. On the other side, the likelihood of memorization can drop quickly to nearly 1% in the case of 500 duplicates and $\leq .01\%$ in the case of 25 duplicates. If ten or fewer duplicates around found in training data, then the model maintains its ability to produce no striking similar content.

5. Discussion

Models can and do memorize protected content. This is true even when over 10,000 images can be found in a training dataset or when a small number of images in a large training dataset are duplicates. The last point is particularly concerning as it is estimated that some popular databases for production models, like LAION-5B, are made up of $\sim$30% duplicated images (Webster et al., 2023). In other words, storing protected content inside a neural network (i.e., inside a machine learning model (Reitinger, 2024)), even a complex model or a model trained with a massive amount of data, should not be deemed to void issues of copyright. The question is not *can* a model memorize, but *did* the model memorize. This turns a theoretical inquiry into a factual one.

Moreover, the factual inquiry, as evidenced in Section 4, is burdened by unclear variables affecting memorization. If both large and small training datasets can be used to produce copied, protected content by generative AI, then this only further incentivizes the assessment of what components went into training a model. However, this type of nuance enters a more closely guarded realm for the owners of production models. While openAI and other AI-facing companies may be willing to share details on model weights or even model architecture, datasets and fine-tuning have so far been left off the table (Liesenfeld et al., 2023; Widder et al., 2023). Further, the pay-to-generate models do not permit an easy way to measure a large number of generations in terms of similarity to training data.

What this means is that not only are models likely to produce offending content every nth generation, but it will also be very difficult for outsiders to gauge when or how often this might occur. In order to reduce or appropriately handle issues of copyright infringement, therefore, model owners need to play a role. Duties should be placed on model owners, the entities that are capable of making detailed assessments of their products. These duties should stretch to both model maintenance and offending generations.

On the maintenance side, typical hygiene should be in place. During a scrape for data used in training, there should be a log concerning where training data came from, who the likely owner of the content is, whether the content could be dangerous or sensitive or unsuitable for certain audiences, when the data was gathered, and who was responsible for gathering the data (Sag, 2023).

On prohibiting or fairly handling copyright, a more difficult dilemma ensues: Should questions of copyright infringement be assessed on the front end or the back end? A front-end solution would be to remove all protected content from a model's training data. While ideal in theory, this approach makes the large assumption that protected and unprotected content is easily distinguishable. Notice that the maintenance duties relate to logging, not the removal of content. Given the difficulty of substantial similarity discussed in this short work-in-progress piece, it is easy to see how that is likely not the case.

The second option is to handle these issues on the back end, after a model is fully baked. One option here might be to mandate transparency in model memorization, making model owners engage in the type of measurements conducted in this Article. That option, however, may not be practical, especially for models that require user input. Not only would the generations be hypothetical, as they would be based on hypothetical user input, but the number of tests that would need to be run may be large (e.g., it might be appropriate to scale the number of generations to the number of training-data examples, which could become impractical).

A more sustainable solution would be to assess model outputs before those outputs are provided to a user. This solution is similar to the filtering that occurs now (Henderson et al., 2022), but could be triggered in royalty-style payments. Whenever a model produces content that is strikingly similar to training data, a model owner may make the assumption that this content is protected and therefore permission would be needed. That permission may come in the form of prior agreed-upon compensation schemes, similar to how royalties operate. This would allow content owners to continue profiting from their expressions, model owners to continue using the data required to make these models work, and set payments on an as-needed basis tied to actual outputs rather than theoretical outputs.

Conclusion. If generative models did not memorize content or produce content that was substantially similar to training data, then model washing (i.e., protected data can be made unprotected by training a machine learning model on the data) would be a viable option to escape copyright liability.[6] However, as the empirical data has shown, models do memorize training data, even without user prompts and even with a large amount of training data. Additionally, the very secrets model owners are keen to keep are the same variables affecting how likely it is that a model produces memorized content. In turn, model owners are needed, and should be required, to take part in ensuring that their models do not offend copyright. A series of duties related to model maintenance and the protection of copyrighted works should be encumbered by model owners.

Acknowledgements and Impact Statement

This paper was not supported by any specific grants. The impacts of this work pertain to using machine learning in a way that encourages the progress of the sciences and useful arts.

References

Abbott, R. and Rothman, E. Disrupting creativity: Copyright law in the age of generative artificial intelligence. *Fla. L. Rev.*, 75:1141, 2023.

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.

Alhadeff, J., Cuene, C., and Del Real, M. Limits of algorithmic fair use. *Wash. JL Tech. & Arts*, 19:1, 2024.

Almeida, D. R. CLIP embeddings to improve multimodal rag with GPT-4 Vision. https://cookbook.openai.com/examples/custom_image_embedding_search, 2024.

Asay, C. D. An empirical study of copyright's substantial similarity test. *UC Irvine L. Rev.*, 13:35, 2022.

Autry, J. R. Toward a definition of striking similarity in infringement actions for copyrighted musical works. *J. Intell. Prop. L.*, 10:113, 2002.

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., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

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.

Cohen, A. B. Masking copyright decisionmaking: The meaninglessness of substantial similarity. *UC Davis L. Rev.*, 20:719, 1986.

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.

Dodes, J. L. Beyond Napster, beyond the United States: The technological and international legal barriers to on-line copyright enforcement. *NYL Sch. L. Rev.*, 46:279, 2002.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

Elmer, G. Exclusionary rules?: The politics of protocols. In *Routledge handbook of internet politics*, pp. 376–383. Routledge, 2008.

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.

[6] In some ways, model washing is similar to the idea that you can escape liability when an AI (i.e., an AI agent) creates a liability gap. This ground is far from new (Santoni de Sio & Mecacci, 2021; Reitinger, 2015a; Tigard, 2021) though seems to be an appealing trend for those wishing to alter certain preexisting restrictions.

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.

Fromer, J. C. and Sprigman, C. J. *Copyright law: Cases and materials*. Independent, 2024.

Fruehwald, E. S. Copyright infringement of musical compositions: A systematic approach. *Akron L. Rev.*, 26:15, 1992.

Gao, Y., Kossof, P., and Dong, Y. Research on the dilemma and improvement of the copyright fair use doctrine related to machine learning in China. *UIC Rev. Intell. Prop. L.*, 22:vii, 2022.

Gillotte, J. L. Copyright infringement in AI-generated artworks. *UC Davis L. Rev.*, 53:2655, 2019.

Gu, X., Du, C., Pang, T., Li, C., Lin, M., and Wang, Y. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.

Henderson, P., Krass, M., Zheng, L., Guha, N., Manning, C. D., Jurafsky, D., and Ho, D. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Howard, J. and Gugger, S. *Deep Learning for coders with fastai and PyTorch*. O'Reilly Media, 2020.

Ke, Y., Sukthankar, R., Huston, L., Ke, Y., and Sukthankar, R. Efficient near-duplicate detection and sub-image retrieval. In *ACM multimedia*, volume 4, pp. 5. Citeseer, 2004.

Khosravy, M., Nakamura, K., Hirose, Y., Nitta, N., and Babaguchi, N. Model inversion attack by integration of deep generative models: Privacy-sensitive face generation from a face recognition system. *IEEE Transactions on Information Forensics and Security*, 17:357–372, 2022.

Lanzalottie, H. J. Is proof of access still required? Proving copyright infringement using the "strikingly similar"' doctrine: An analysis of the Fourth Circuit's decision in Bouchat v. Baltimore Ravens, Inc. *Vill. Sports & Ent. LJ*, 9:97, 2002.

Latman, A. "Probative similarity" as proof of copying: Toward dispelling some myths in copyright infringement. *Columbia Law Review*, 90(5):1187–1214, 1990.

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.

Liesenfeld, A., Lopez, A., and Dingemanse, M. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th international conference on conversational user interfaces*, pp. 1–6, 2023.

Lim, D. Saving substantial similarity. *Fla. L. Rev.*, 73:591, 2021.

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.

Monaco, N. and Woolley, S. *Bots*. John Wiley & Sons, 2022.

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.

Nilsson, J. and Akenine-Möller, T. Understanding SSIM. *arXiv preprint arXiv:2006.13846*, 2020.

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.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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.

Reitinger, N. Algorithmic choice and superior responsibility: Closing the gap between liability and lethal autonomy by defining the line between actors and tools. *Gonz. L. Rev.*, 51:79, 2015a.

Reitinger, N. CAD's parallel to technical drawings: Copyright in the fabricated world. *J. Pat. & Trademark Off. Soc'y*, 97:111, 2015b.

Reitinger, N. Artificial intelligence is like a perpetual stew. *American University Law Review*, 73, 2024.

Ren, J., Li, Y., Zen, S., Xu, H., Lyu, L., Xing, Y., and Tang, J. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. *arXiv preprint arXiv:2403.11052*, 2024.

Rogers, E. Substantially unfair: An empirical examination of copyright substantial similarity analysis among the federal circuits. *Mich. St. L. Rev.*, pp. 893, 2013.

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.

Santoni de Sio, F. and Mecacci, G. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, 34(4): 1057–1084, 2021.

Scheffler, S., Tromer, E., and Varia, M. Formalizing human ingenuity: A quantitative framework for copyright law's substantial similarity. In *Proceedings of the 2022 Symposium on Computer Science and Law*, pp. 37–49, 2022.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Snapes, L. AI song featuring fake Drake and Weeknd vocals pulled from streaming services. *The Guardian*, 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.

Terekhov, M., Graux, R., Neville, E., Rosset, D., and Kolly, G. Second-order jailbreaks: Generative agents successfully manipulate through an intermediary. In *Multi-Agent Security Workshop@ NeurIPS'23*, 2023.

Tigard, D. W. There is no techno-responsibility gap. *Philosophy & Technology*, 34(3):589–607, 2021.

US Constitution, Article I, Section 8, Clause 8. Constitution of the united states, 1787.

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.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Webster, R., Rabin, J., Simon, L., and Jurie, F. On the de-duplication of LAION-5B. *arXiv preprint arXiv:2303.12733*, 2023.

Widder, D. G., West, S., and Whittaker, M. Open (for business): Big tech, concentrated power, and the political economy of open AI. *Concentrated Power, and the Political Economy of Open AI (August 17, 2023)*, 2023.

Yoon, T., Choi, J. Y., Kwon, S., and Ryu, E. K. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.