Break It Till You Make It

Limitations of Copyright Liability Under A Pre-training Paradigm of AI Development

Rui-Jie Yew [1 2] **Dylan Hadfield-Menell** [1]

1. Article Overview

Recent progress in the development of generative AI systems has brought questions of copyright liability to the attention of litigators (*Anderson et al. v. Stability AI*, (N.D. Cal. 2023) (No. 3:23-cv-00201)), scholars (Khan & Hanna, 2023; Vyas et al., 2023), artists, technologists, and the general public. Whereas predictive or cluster-based AI models produce quantitative insights about data, generative systems produce outputs drawn from training data–including artworks and literary pieces. This technology has created possibilities for new artistic expression at the same time as they threaten the livelihood of artists whose works are used in training.

The legal literature has proposed tests to analyze liability for direct infringement when copyrighted works are used in model training. These tests typically rely on a tight-knit relationship between model training and model deployment (Sobel, 2017; Lemley & Casey, 2020; Quang, 2021). Where model outputs are non-infringing, the process of model training on copyrighted works is also characterized as non-infringing. However, the development and deployment of AI systems increasingly rely on multiple learning processes and actors. This creates opportunities for AI developers to identify technical and organizational workarounds that subvert legal analyses. These loopholes arise from the mismatch between the (assumed) integration between training and deployment and the *pre-training* process that has emerged as the dominant paradigm for model development (Devlin et al., 2018; Bommasani et al., 2021; Ramesh et al., 2022). In practice, large models are usually trained with generic tasks (e.g., next-word prediction) on broad datasets and *fine-tuned* (i.e., re-trained) on narrower and, typically, smaller task-specific datasets. Pre-training is expensive, creating a split in the market for these systems between a small group of high-resource corporate actors (OpenAI, 2019; Microsoft, 2023) that produce pre-trained models and a large group of smaller downstream actors that *fine-tune* pre-trained models (Bender et al., 2021).

Based on this split in processes and actors, we explore copyright's secondary liability doctrine in the practical effect of copyright regulation on the development and deployment of AI systems. From this insight, we draw on similarities and dissimilarities between generative AI and the regulation of peer-to-peer file sharing through secondary copyright liability to understand how companies may manage their copyright liability in practice. We discuss how developers of pre-trained models can, through a similar combination of technical and developmental strategies, also subvert regulatory goals. Effectively combating these subversion strategies reveals the importance of a systems-level analysis and understanding to regulate AI systems. We conclude with a discussion of regulatory strategies to close these loopholes and propose duties of care for developers of ML models to evaluate and mitigate their models' present and downstream effects on the authors of copyrighted works that are used in training.

1.1. Illustrative Example: Music Generation

We use an extended example of music generation to illustrate the point. We first consider a candidate infringing use of data to copy Ariana Grande's music with a generative model. Then, we show how a development pipeline can split pre-training and fine-tuning across different entities for this application. This split complicates the determination of liability and, potentially, shields both developers from direct, primary liability.

In order to overview the relevant technical and legal dimensions of this issue, we will use a running example of music generation. Consider a music generation system that uses a pre-trained music model, such as in (Agostinelli et al., 2023) or (Zeng et al., 2021). This pre-training step uses a large dataset of recordings that includes, among others, Ariana Grande's catalog. Zeng et al. (2021) show how such a model can be repurposed effectively for both expressive (melody completion) and non-expressive (accompaniment suggestion, genre classification, and style classification) tasks.

The model is fine-tuned for melody completion with exam-

[1]MIT CSAIL [2]Brown University. Correspondence to: Rui-Jie Yew <ryew@cs.brown.edu>.

Accepted to the *1st Workshop on Generative AI and Law*, co-located with the *International Conference on Machine Learning*, Honolulu, Hawaii, USA. 2023.

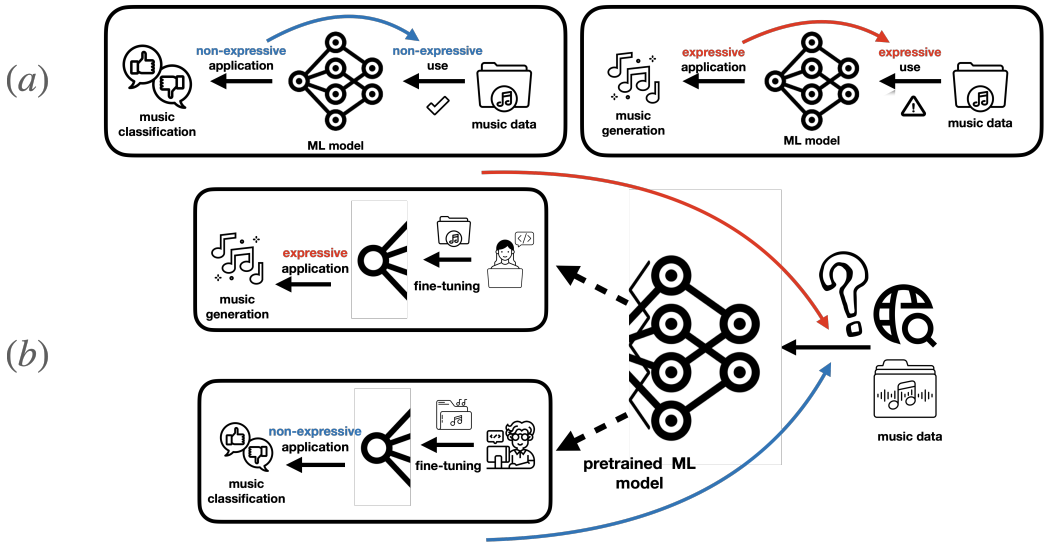

Figure 1. **(a)**: An illustration of the non-expressive application of an ML model toward music genre classification and an expressive application of an ML model toward music generation. From proposed judicial tests, the application and outputs of the model have had a strong bearing on whether the use of copyrighted works to train the model would be considered non-infringing. **(b)**: Fine-tuning involves training the last layer of the pre-trained network toward a specialized task. Pre-trained models produced from training on copyrighted works can be applied in ways that have been characterized as both expressive and non-expressive. Whether the use of copyrighted works in the process of training a model is characterized as non-expressive (non-infringing) or expressive (potentially infringing) has been highly dependent on the outputs or application of the model. pre-trained models do not embed any eventual applications or outputs, breaking the relationship that is centered in proposed judicial tests to determine whether the use of copyrighted works in training models is expressive or non-expressive.

ples that are labeled as good and bad completions of the melody, as described in Zeng et al. (2021). We will suppose that this step does not use any of Ariana Grande's catalog, as the datasets for these tasks are typically much smaller[1]. Finally, the model is deployed and produces a viral melody completion that is highly similar to Ariana Grande's song "Thank U, Next" and a legal battle ensues.

The primary question is whether the use of Ariana Grande's music to train the model means that it effectively copied her music to complete a melody. For example, the model may have memorized protected aspects of Ariana Grande's recordings and used them to perform well in the downstream task of melody completion. While it is unclear how courts will rule in similar cases, it is quite plausible that the developers of the melody completion model will be judged to have infringed on Ariana Grande's copyright because the ultimate use of the model is expressive and the output could interfere with the market for Ariana Grande's music.

The key issue that our paper highlights is that this answer becomes much more complicated if separate organizations conduct pre-training and fine-tuning. Suppose that Company A pre-trains an "embedding" model[2] that represents songs as high-dimensional vectors. Separately, Company B fine-tunes a generative layer on top of these embeddings based on a small set of public domain examples of good and bad melody completions. The final system is the same as before, as is the potentially infringing output.

In this case, both companies appear to have a good defense against direct copyright liability. Company A produced a numerical representation of music by analyzing building blocks of musical expression from a generic dataset. It is hard to argue that this is copying an artist's expression. Similarly, Company B only used public domain data that conveyed generic information about good and bad melody completion. It is even possible that none of Company B's employees have listened to Ariana Grande's music. Because the purpose of use changes at different steps of model training, the attribution of liability could become less obvious. Furthermore, applying proposed judicial tests and copyright law for direct liability as they stand may make the most defensible phase of training (pre-training) also the only phase that actually uses copyrighted data.

[1] We could also suppose that the company lawyers have required that engineers use music in the public domain in this step to reduce liability.

[2] see, e.g., https://platform.openai.com/docs/models/embeddings

2. Duties of Care

Simply scraping an image or having a technical system that creates an output similar to a piece of art may not automatically trigger copyright infringement.[3] Rather, it is the causal process of copying between these events that gives rise to potential liability. By introducing multiple training processes, intermediary developers, and additional network layers, a pre-training paradigm of machine learning challenges the legal conceptualization of this causal process. It renders both the responsibility of those involved as well as the technical link between data and output tenuous. In light of this, the responsibility of developers (Khan & Hanna, 2023; Cobbe et al., 2023; Cen et al., 2023) for model outcomes could be articulated through a duty of care. A variant of this effort could be in advancing a duty of care that developers of general-purpose models would have toward authors to: (1) monitor training processes and (2) conduct evaluations on downstream use cases. We consider two important doctrines in copyright law and how such a duty could be advanced through these doctrines. Finally, we discuss the limits of copyright liability as a regulatory measure.

Fair Use Doctrine Case-by-case fair use determinations could pose an opportunity to assess new technologies and harms. Copyright liability is considered a strict liability tort by lawmakers and legal scholars (Grimmelmann, 2017; Balganesh, 2012). However, legal scholars have also argued that the extensive judicial considerations baked into copyright law and the fair use doctrine make findings of liability for infringement more similar to a fault liability, or negligence, tort (Goold, 2015; Hetcher, 2013). Duties and standards of care are foundational to negligence law because they provide an expectation of responsibility and a bedrock for liability determinations (Wheeler, 2023).

Duty of care considerations could be embedded into fair use to, e.g., encourage upstream developers evaluate which aspects of an author's expression are potentially captured from the training process (as part of (1)) and to monitor how the model can be applied in ways that cause market harm to authors (as part of (2)). Fair use is also a point in litigation to contextualize standards of care, particularly in public interest, non-commercial use cases.[4]

Secondary Liability Secondary copyright liability is a doctrine of copyright liability that handles the liability of those who facilitate direct infringements. Because generative AI systems have been used to produce works that infringe copyright, they introduce questions of secondary copyright liability. However, generative AI systems also differ other from other technologies that have triggered secondary liability litigation (A&M Records, Inc. v. Napster, Inc., 239 F.3d 1004 (9th Cir. 2001); *MGM Studios, Inc. v. Grokster, Ltd.*, 545 U.S. 913 (2005)) in the past because copyrighted works are also *directly used* in the development of generative AI systems.

Establishing standards of care for the oversight of both upstream (as part of (1)) and downstream activities (as part of (2)) can facilitate liability attribution in light of an increasingly complex algorithmic supply chain (Cobbe et al., 2023). Gargantuan datasets can be difficult to manage. On top of this, model developers may not have an incentive to look into and document whose works are in these datasets, leading to documentation debt (Bender et al., 2021). Evidence of access could facilitate a finding of copying. Instead, upstream dataset collection and labelling could be assigned to third-parties, which could be leveraged strategically to limit condemning information.

Downstream oversight is also important to liability attribution. Treated as a single developmental pipeline, using Ariana Grande's music to create a Grande music generation machine very clearly has implications for copyright liability. Broken into separate training processes, the production of such a system may no longer directly embed the goal of learning and generating from Ariana Grande's expression–making liability difficult to disentangle. In peer-to-peer network litigation under secondary liability, the splitting of technological supply chains similarly challenged secondary liability copyright doctrines (Choi, 2005).

2.1. Limitations of Copyright Liability

Without high-quality training data, typically in the form of copyrighted works, AI systems would simply fail to function. However, because copyright law focuses on authorial expression and is utilitarian at heart, it may not stand as an avenue for recourse and compensation where copyrighted works still clearly contribute to system functionality–such as for autocomplete systems or for image classification. This does not mean that harms that could arise from these uses are unimportant, but that copyright law may be mismatched to address them. Additionally, given the burdensome process of case-by-case litigation, other areas of law such as competition law may be better suited to address these harms (Choi, 2023).

However, copyright law is among the first laws being wielded against generative AI in the United States. As Levendowski (2018) points out, until standards and laws that address these harms are moved forward, the application of copyright law has enormous potential to shape the incen-

[3] As Balganesh (2012) notes, for copyright law: "Neither the bare act of reproduction nor the mere production of a substantially similar work is sufficient to trigger liability without the other."

[4] For example, public interest research technologists should not face the same standards as for-profit AI developers. Duties or standards of care should not limit their ability to assess these models.

tives of AI development and deployment as a whole–making it that much more important that we work to get it right.

Acknowledgements

This work was supported by a gift from Effective Giving. The authors are grateful to Peter Henderson for insights on copyright, as well as to Julian Manyika, Stephen Casper, and Dora Zhao for discussions on pre-training for language and vision models.

References

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Balganesh, S. The obligatory structure of copyright law: Unbundling the wrong of copying. *Harvard Law Review*, 125(7):1664–1690, 2012.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Cen, S., Hopkins, A., Ilyas, A., and Madry, A. On ai deployment: Ai supply chains (and why they matter). https://aipolicy.substack.com/p/supply-chains-2, 2023.

Choi, B. H. The grokster dead-end. *Harv. JL & Tech.*, 19: 393, 2005.

Choi, E. Protecting visual artists from generative ai: An interdisciplinary perspective. In *1st ICML Workshop on Generative AI and Law*, 2023.

Cobbe, J., Veale, M., and Singh, J. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1186–1197, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Goold, P. R. Is copyright infringement a strict liability tort. *Berkeley Tech. LJ*, 30:305, 2015.

Grimmelmann, J. *Internet law: Cases and problems*. Semaphore Press, 2017.

Hetcher, S. The immorality of strict liability in copyright. *Marq. Intell. Prop. L. Rev.*, 17:1, 2013.

Anderson et al. v. Stability AI, (N.D. Cal. 2023) (No. 3:23-cv-00201).

Khan, M. and Hanna, A. The subjects and stages of ai dataset development: A framework for dataset accountability. *Forthcoming 19 Ohio St. Tech. L.J.*, 2023.

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

Levendowski, A. How copyright law can fix artificial intelligence's implicit bias problem. *Wash. L. Rev.*, 93:579, 2018.

MGM Studios, Inc. v. Grokster, Ltd., 545 U.S. 913 (2005).

Microsoft. General availability of azure openai service expands access to large, advanced ai models with added enterprise benefits. https://perma.cc/YP7H-GJQN, 2023.

OpenAI. Better language models and their implications. https://perma.cc/T3E3-34RJ, 2019.

Quang, J. Does training ai violate copyright law? *Berkeley Tech. LJ*, 36:1407, 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*, 2022.

Sobel, B. L. Artificial intelligence's fair use crisis. *Colum. JL & Arts*, 41:45, 2017.

A&M Records, Inc. v. Napster, Inc., 239 F.3d 1004 (9th Cir. 2001).

Vyas, N., Kakade, S., and Barak, B. Provable copyright protection for generative models. *arXiv preprint arXiv:2302.10870*, 2023.

Wheeler, T. The three challenges of ai regulation. *Brookings*, 2023.

Zeng, M., Tan, X., Wang, R., Ju, Z., Qin, T., and Liu, T.-Y. Musicbert: Symbolic music understanding with large-scale pre-training. *arXiv preprint arXiv:2106.05630*, 2021.