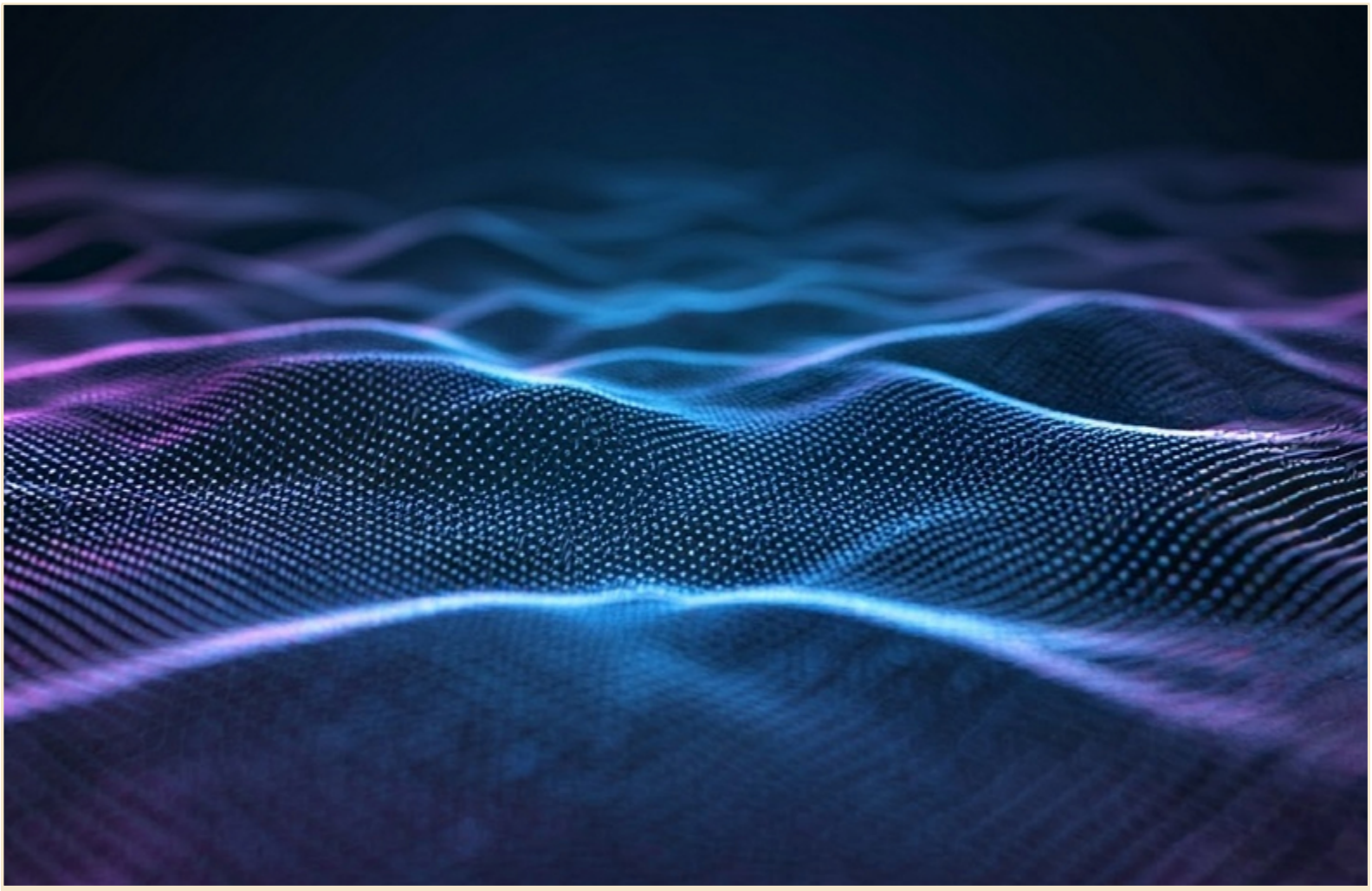

GENERATIVE AI COLLECTION, TRAINING, OUTPUT: WHEN MIGHT COPYRIGHT INFRINGEMENT BE FOUND?

BY YAR CHAIKOVSKY

& JORDAN COYLE

& WOENHO CHUNG

& AMIR JABBARI

Yar Chaikovsky is chair of White & Case's Global Intellectual Property practice group and a partner in the Global Technology Industry Group. Jordan Coyle is a partner in the firm's Washington, D.C. office. Woenho Chung and Amir Jabbari are associates in the firm's Silicon Valley office. All authors are members of the firm's Intellectual Property practice group.

The authors thank Payton Fong, Mick Li, and Elizabeth Oh for their assistance on this article.

Visit **www.competitionpolicyinternational.com** for access to these articles and more!

GENERATIVE AI COLLECTION, TRAINING, OUTPUT: WHEN MIGHT COPYRIGHT INFRINGEMENT BE FOUND?

By Yar Chaikovsky, Jordan Coyle, Woenho Chung, Amir Jabbari

Plaintiffs allege that creators of generative AI models have collected copyrighted works, used those copyrighted works to train their models, and then generated copyrighted output from those models in response to user prompts. As of publication, copyright holders have filed 34 lawsuits accusing providers of generative AI models of infringing their copyrights (sometimes together with non-copyright claims). No court has issued a ruling on the merits of those copyright claims, leaving both litigants and eager spectators of these cases to wonder, among other things: where might copyright infringement liability attach? Could it be at the collection stage where plaintiffs allege that generative AI model creators have ingested and encoded copies of 183,000 books, millions of records of content, and terabytes of other works to train their large language models ("LLMs")? Could it be at the training stage where plaintiffs allege that LLMs are "learning" and "fine-tuning" by copying, digesting, and reproducing datasets comprised of copyrighted material and, in the process, "memorizing" copies of that material? Could it be at the output stage where plaintiffs allege that generative AI models can be prompted to recite copyrighted material? This article examines plaintiffs' allegations and early 12(b) orders addressing these questions.

Scan to Stay Connected!

Scan here to subscribe to CPI's **FREE** daily newsletter.

© 2024 Competition Policy International® All Rights Reserved

01

INTRODUCTION

We live in an era where Artificial Intelligence ("AI") powered applications such as ChatGPT are setting records for the fastest growing user base.[2] ChatGPT and generative AI platforms have become household names, capable of drafting college essays and legal briefs, creating digital art, and composing music. The sudden rise in popularity of applications like ChatGPT has come on the heels of recent developments in the field of AI, which have enabled new paradigms of machine processing, shifting from data-driven, discriminative machine learning tasks toward sophisticated, creative tasks through generative AI.[3] But history has shown such rapid technological developments often clash head-on with existing legal frameworks, which lag in reacting to the rapidly changing technological landscape.[4]

The clash between copyright holders and developers of generative AI models has begun and has spurred debate over the benefits and drawbacks of legislative, judicial, or private resolution.

Some countries have moved quickly to amend their copyright laws to reflect the changing landscape of AI technology. For instance, Japan amended its law to provide broad rights allowing companies to ingest and use copyrighted works for any type of information analysis, including training AI models.[5] In the United States, however, Congress has failed to enact any of at least 41 major bills over the past two years that would regulate or affect AI, including at least eight bills that would require some form of disclosure of AI generation or training data.[6]

Plaintiffs and their lawyers, however, have sprung into action, filing (as of the submission of this article) 34 complaints for copyright infringement (often together with other claims) by AI generators.[7] While there have not been any decisions on the merits of any of these copyright claims, the complaints and early orders on motions to dismiss provide insight into how the involved parties and some courts view the relationship between AI technology and copyright protections. There are many open questions surrounding these claims, but this article will delve into the question of at what stage — collection of data, training, or output — copyright infringement may or may not be found in the AI context.[8]

02

MACHINE LEARNING TO GENERATIVE AI

Artificial intelligence is a broad term with many different definitions (depending on, e.g. the field or jurisdiction) but it is generally understood to describe computational algorithms capable of performing tasks that typically require human intelligence, for example, understanding natural language and learning from experience.[9] Early machine learning algorithms were mostly rule-based and aimed at supporting users and businesses in decision-making based on historical data.[10] These early systems were limited in that they performed tasks based on historical data, as opposed to what we would consider "creating" something new. Later, a more developed form of machine learning algorithms, also known

2 See Krystal Hu, "ChatGPT sets record for fastest-growing user base - analyst note," Reuters (Oct. 29, 2024), https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

3 Banh, Leonardo, et al., "Generative artificial intelligence," *Electronic Markets* (2023) 33:63; Desai, Bhavin, et al., "Large Language Models: A Comprehensive Exploration of Modern AI's Potential and Pitfalls," *Journal of Innovative Technologies* (2023), Vol. 6; Das, Sumit, et al., "Applications of Artificial Intelligence in Machine Learning: Review and Prospect," *International Journal of Computer Applications* (2015), Vol. 115:9.

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

5 Article 30-4 of Japan's Copyright Act.

6 See https://chatgptiseatingtheworld.com/2024/04/18/list-of-ai-bills-before-congress/, ChatGPT is eating the world (Oct. 29, 2024).

7 See https://chatgptiseatingtheworld.com/2024/10/21/status-of-all-33-copyright-lawsuits-v-ai-oct-21-2024/, ChatGPT is eating the world (Oct. 29, 2024).

8 Certainly, one of the most significant questions relating to copyright claims asserted against generative AI generators is whether and to what extent the fair use defense excuses generators' use of copyrighted material at any stage. That question could be the subject of an entirely separate article (perhaps even a book!) and is outside the scope of this article.

9 *Id.*

10 *Id.*

© 2024 Competition Policy International® All Rights Reserved

as deep learning, was developed to utilize neural networks in modeling complex data representations and identifying correlations and patterns in large datasets.[11] With the introduction of deep learning, AI systems could process high-dimensional data, including texts, images, videos, and audio.[12]

Advancements in deep learning techniques led to the development of generative models, a subset of deep learning models capable of creating new content based on existing data. Unlike machine learning systems, generative AI models focus on generating new data rather than merely predicting, based on existing data. Training a generative AI model also differs from training a traditional AI model due to the use of semi-supervised learning.[13] Users interact with generative AI models using natural language or prompts to create the desired output such as text, images, and music, among other things.[14]

03
COPYRIGHT INFRINGEMENT THEORIES

A. Creating the Datasets

AI tools require huge datasets to train their algorithms, which are created by, for example, scraping images, video, and text from the internet. Plaintiffs allege that these massive datasets inevitably contain copyrighted material. Before the rise of AI, plaintiffs initiated lawsuits based on the collection and storage of copyrighted data; for example, authors and book publishers sued newspaper publishers or online databases for digitizing and distributing their copyrighted material.[15]

One question that has arisen in AI copyright cases is whether the collection and storage of copyrighted material alone can form the basis of a copyright infringement claim irrespective of the final output. Intuitively, the answer may not be simple. After all, humans learn by continuously ingesting inputs around us, many of which are copyrighted: books, periodicals, music, graphic arts, and brands to name a handful. From the day we are born, every audio, visual, and tactile input goes into our memory bank and is used (consciously or subconsciously) in every decision that we make for the rest of our lives. As long as we don't recreate one of these inputs and claim it as our own, no one would accuse a human of copyright infringement merely because we learned something from our observation of copyrighted content.

However, plaintiffs may claim that the collection of data to train generative AI platforms is different, because there is an aspect of "intermediate" copying during the initial data collection phase, where copyrighted material is fixed to a medium. Some courts have determined, albeit in a non-AI context, that such "intermediate" copying can form the basis of a copyright infringement claim because it violates the exclusive rights granted to the copyright owner in § 106 of the Copyright Act.[16] Plaintiffs have begun to sue AI companies based on these facts.[17]

One potential defense may lie in the technical details of how modern AI-powered models work. Often, AI models rely on a technique called tokenization. Tokenization transforms natural language text — what can be found online for instance — into a mathematical sequence of arrays filled with integers representing words. AI generators may have a defense against a copyright claim at the collection phase to the extent that they are storing tokens, as opposed to raw copyrightable information.

During the data collection phase, defendants explain that AI models are not storing the literal text of the authors. Similarly, during the training phase, the AI models are not working with the literal text of the authors but rather a transformed numerical representation of the text. While the tokens are eventually reverted into natural languages, one could argue that both the data collection and training phases involved only numerical representations. Although these vectors bear a strong relationship to the original work, defendants will likely argue that the numerical information they communicate with the machine is quite far removed from the protectable expression of the original text.

11 *Id.*

12 *Id.*

13 *Id.*

14 *Id.*

15 See, e.g. *Author's Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).

16 *Sega Enters. v. Accolade, Inc.*, 977 F.2d 1510, 1514 (9th Cir. 1992).

17 *Andersen v. Stability AI Ltd.*, 700 F. Supp. 3d 853, 864 (N.D. Cal. 2023).

© 2024 Competition Policy International® All Rights Reserved

While the original copyrightable work may contain protectable expression, the fact that the AI models learn with vectors, not the original expression in natural language, makes for a complicated conceptual framework and creates an interesting battleground in future copyright cases.

B. Training the AI Model

The data used to train the AI systems are also facing scrutiny. Several class action lawsuits have been filed in the past several years by authors and artists alleging that their copyrighted works were used without permission to train AI models.

Formulating a complaint based solely on AI's use of copyrighted works to train an AI model may be futile. Several motions to dismiss have been granted over the last decade based on ill-pled theories such as vicarious copyright infringement and various Digital Millenium Copyright Act ("DMCA") claims.[18] One of the most common allegations under such a fact pattern is a claim of vicarious copyright infringement. Undoubtedly guided by precedent such as *Napster* (a non-AI case),[19] AI copyright plaintiffs have sought to hold the companies that are using their copyrighted work to train AI models liable for vicarious infringement. In *Napster*, the Second Circuit affirmed the district court's ruling that Napster had knowledge of the infringing activities and had the ability to control and benefit from those activities.[20] However, the flaw in applying *Napster* to cases where the plaintiff has only alleged that its copyrighted work is being used to train an AI model is that in *Napster* it was clear that the music that was being copied was exactly the same as the original work; here, many plaintiffs are not able to allege that their work is similar to an outputted work. Under Ninth Circuit law, without an infringing output there can be no vicarious infringement.[21]

Similarly, a DMCA Section 1202(a)(1) claim fails if the plaintiff is unable to allege an infringing derivative work or output.[22] With respect to a DMCA Section 1202(b) plaintiffs must plausibly allege that the engineers who trained the model intentionally removed the copyright management information from the copyrighted works.[23]

However, all may not be lost for such a plaintiff. At least one court denied a motion to dismiss a claim that plaintiffs' copyrighted work is used to train an AI model without any allegation of similar output.[24] Although we do not yet know how courts will rule on the merits of such claim, the fact that this court allowed plaintiffs to move past the pleading stage based solely on the allegation that their work was used to train an AI model increases the risk of litigation to companies who use data sets with copyrighted works to train their AI models. A successful claim based on training data without a showing of similarity in the output could be a significant challenge to the AI industry, as virtually every AI generator could be liable for infringement just by using copyrighted data to train their AI model.

Plaintiffs that can plead substantial similarity in the output may fare much better. In *Anderson v. Stability AI*, plaintiffs alleged that the AI-generated art produced by defendants was substantially similar to the plaintiffs' original works and attached a 150-page exhibit to the complaint showing exemplary images.[25] This similarity, according to the plaintiffs, indicated that their copyrighted art had been used without permission to train the AI models, leading to outputs that closely resembled their creations. The court denied defendants' motion to dismiss plaintiffs' induced copyright infringement claim. The court held that defendants' argument that plaintiffs induced infringement theory was a mere repacking of the direct infringement claim is better addressed on summary judgment after discovery.

> ***"Plaintiffs that can plead substantial similarity in the output may fare much better"***

18 *Kadrey v. Meta Platforms, Inc.*, No. 23-cv-03417-VC, 2023 U.S. Dist. LEXIS 207683, at *2 (N.D. Cal. Nov. 20, 2023); *Tremblay v. OpenAI, Inc.*, No. 23-cv-03223-AMO, 2024 U.S. Dist. LEXIS 24618, *4 (N.D. Cal. Feb. 12, 2024).

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

20 *Id.*

21 *Kadrey*, 2023 U.S. Dist. LEXIS 207683, at *2; Tremblay, 2024 U.S. Dist. LEXIS 24618, at *4.

22 *Kadrey*, 2023 U.S. Dist. LEXIS 207683, at *2.

23 *Tremblay*, 2024 U.S. Dist. LEXIS 24618, *4.

24 *Kadrey*, 2023 U.S. Dist. LEXIS 207683, at *2.

25 *Anderson v. Stability AI Ltd.*, No. 23-cv-00201-WHO, 2024 U.S. Dist. LEXIS 143204, *33-35 (N.D. Cal. Aug. 18, 2024).

© 2024 Competition Policy International® All Rights Reserved

The degree of similarity of the copyright work to the output of the AI model could be important in affecting the outcome of a case. Whether the output of an AI model is substantially similar, as is alleged in *Anderson*, or less so will factor into infringement analyses and defenses such as fair use. As data sets grow and AI becomes more advanced it may be more difficult for plaintiffs to map their works to the output of the AI models. Plaintiffs will have to become more creative in the ways they find use of their copyrighted work in the output — perhaps turning to AI to assist them.

C. Generative AI Output

Lastly, any ability that consumers have to prompt generative AI to produce copyrighted material as an output will also breed secondary copyright infringement claims. *The New York Times*' complaint against Microsoft and OpenAI is a prime example of such a claim.[26] In this Southern District of New York case, *The New York Times* alleges that a user of the defendants' AI platform could prompt the AI to yield output substantially similar to its copyrighted material.[27] *The New York Times* further alleges that the defendants must have been aware of the potential infringing uses by consumers because they developed the products and understood that the output may contain the copyrighted material.[28] At the time of this publication, Defendants' motion to dismiss is pending.

04
SOLUTIONS AND DEFENSES

A variety of solutions have been proposed to address the floodgates of litigation opened by the use of copyrighted information by generative AI platforms. One possible solution is for companies who create AI models to license the data they use to train their models before they begin training. This is perhaps the simplest approach, albeit at a substantial cost. For example, reports indicate that as of July 2024, OpenAI had spent over $275 million licensing content providers.[29] A similar approach has been applied to music libraries for the past decade: distribution companies license music wholesale from record labels and license these libraries of music for consumers to reproduce or create derivative works.

Another possible solution is to impose a licensing regime after the fact where content contributors get paid royalties when the AI uses their work to create the output. One problem with this approach is that each year AI systems are using exponentially more data to train their models and it will be difficult, perhaps at times impossible, to pinpoint exactly which of millions of data points an AI system is using when generating a specific output. Moreover, even if this identification were possible it could amount to very little compensation for each individual work given the massive quantity of works ingested and challenges with tying output to specific training materials.

26 *New York Times Co. v. Microsoft Corp.*, No. 1:23-cv-11195, Dkt. 170 (S.D.N.Y. Aug. 12, 2024).

27 *Id.*

28 *Id.*

29 See, e.g. https://www.cbinsights.com/research/ai-content-licensing-deals/, CBInsights (Oct. 29, 2024).

© 2024 Competition Policy International® All Rights Reserved

One legal scholar has made the case that AI systems should generally be able to use databases for training whether or not the contents of that database are copyrighted.[30] The underlying basis for such an argument is that AI "isn't competing with authors or artists. Instead, it is using their work in an entirely different manner . . . [AI] systems generally copy works not to get access to their creative expression (the part of the work the law protects), but to get access to the uncopyrightable parts of the work — the ideas, facts, and linguistic structure of the works."[31] Under this framework, AI could learn from data and would be immune to claims of copyright infringement even if its output failed the test for fair use.

However, the authors' premise — that AI isn't competing with authors or artists because it is using their works in an entirely different manner — may be under increased scrutiny in light of recent case law developments. In *Andy Warhol*, the Supreme Court analyzed the transformative nature of the work and held that Warhol infringed a photographer's copyright when he created a series of silk screen images based on the copyrighted photograph. While the district court found Warhol's work to be transformative enough to support his invocation of the fair use defense, the Second Circuit and Supreme Court disagreed. The Supreme Court found that Warhol's use of the copyrighted photograph did not have a purpose and character that is sufficiently distinct from the copyrighted photograph on which it was based: "Goldsmith's original photograph of Prince, and AWF's copying use of that photograph in an image licensed to a special edition magazine devoted to Prince, share substantially the same purpose, and the use is of a commercial nature."[32] While the fair use analysis is very fact-specific, *Andy Warhol* may limit the availability of the fair use defense to AI generators, especially AI generators of visual works.

05
CONCLUSION

As litigants and courts ponder the stage at which AI providers may be liable for copyright infringement, one should remember the parallels between how we as humans learn and how an AI model learns. As AI models become more and more "human-like" the question may become whether we will promulgate a legal framework that treats AI as robots or ultimately more like people.

30 Lemley, Mark, et al., "Fair Learning," 99 Tex. L. Rev. 743.

31 *Id.*, at 772.

32 *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 550 (2023).

© 2024 Competition Policy International® All Rights Reserved

CPI SUBSCRIPTIONS

CPI reaches more than **35,000 readers** in over **150 countries** every day. Our online library houses over **23,000 papers**, articles and interviews.

Visit **competitionpolicyinternational.com** today to see our available plans and join CPI's global community of antitrust experts.

© 2024 Competition Policy International® All Rights Reserved

CPI COMPETITION POLICY® INTERNATIONAL