



May 15, 2025

If you have any questions regarding the matters discussed in this memorandum, please contact the following attorneys or call your regular Skadden contact.

Stuart D. Levi

Partner / New York
212.735.2750
stuart.levi@skadden.com

Mana Ghaemmaghami

Associate / New York
212.735.2594
mana.ghaemmaghami@skadden.com

MacKinzie M. Neal

Associate / New York
212.735.2856
mackinzie.neal@skadden.com

This memorandum is provided by Skadden, Arps, Slate, Meagher & Flom LLP and its affiliates for educational and informational purposes only and is not intended and should not be construed as legal advice. This memorandum is considered advertising under applicable state laws.

One Manhattan West
New York, NY 10001
212.735.3000

Copyright Office Weighs In on AI Training and Fair Use

On May 9, 2025, the United States Copyright Office (the USCO) released a 108-page report on whether the unauthorized use of copyrighted materials to train generative artificial intelligence (AI) systems is defensible as a fair use.¹ While both sides of this debate will find aspects of the long-awaited report that support their positions, the clear view of the USCO is that certain uses cannot be defended as fair use. The nonbinding report was released against an important political backdrop that we discuss below, which may ultimately impact whether this is the USCO's final say on this matter.

Summary of the Report's Key Findings

Infringement

- Using copyrighted works to train AI models may constitute *prima facie* infringement of the right to reproduce such works.
- Where AI-generated outputs are substantially similar to the training data inputs, there is a "strong argument" that the models' weights themselves infringe the reproduction and derivative work rights of the original works. *This point has been sharply debated, with different courts taking opposing views.*² *The fact that the USCO believes that in certain cases the models themselves may be infringing could lend support to this argument.*

Fair Use

- Whether training a generative AI foundation model on copyrighted works is transformative for fair-use purposes is a "matter of degree," and where a model is trained to produce content that "shares the purpose of [the original work of] appealing to a particular audience," the use is "at best, modestly transformative." *While the report acknowledges that the fair-use defense is available in certain scenarios, the USCO seems less inclined to support this defense where models generate outputs that infringe copyrighted works. Whether the USCO would have a different view if only de minimis amounts of infringing content were generated is not clear.*
- The arguments that using copyrighted works to train AI models is inherently transformative because it is not for expressive purposes or because it can be analogized to human learning are each "mistaken." AI models absorb "the essence of linguistic expression," and even the objective of human learning does not wholesale justify acts that would otherwise constitute copyright infringement, such as allowing one to make unlimited copies of copyrighted materials. *The report also advances an interesting*

¹ The [first USCO report addressed](#), issued in July 2024, addressed digital replicas. The [second report, issued in January 2025](#) addressed the copyrightability of AI-generated works.

² Compare *Kadrey v. Meta Platforms* No. 23-cv-3417 (N.D. Cal., March 24, 2025) (allegations that a model can be infringing is "nonsensical") and *Andersen v. Stability AI* 744 F. Supp. 3d 956, 982–84 (N.D. Cal. 2024) (copies or protected elements of the original work remained, in some format, within the model).

Copyright Office Weighs In on AI Training and Fair Use

policy argument that the exclusive rights of copyright law are premised on the fact that humans retain only “imperfect impressions of the works they have experienced, filtered through their own unique [views],” while AI allows for “the creation of perfect copies with the ability to analyze works nearly instantaneously.” This suggests that, in the view of the USCO, the Copyright Act’s inherent balance between encouraging creativity and encouraging innovation may not operate as intended in the generative AI context if works can be used as training data without permission. Nonetheless, the report concludes that government intervention into this issue is not currently needed.

- The use of guardrails to prevent or minimize the creation of infringing outputs (e.g., blocking certain prompts and training protocols designed to make infringing outputs less likely) weighs in favor of a fair-use argument. *While many AI models already take such steps (e.g., refusing to generate images of famous personalities), developers may adopt more guardrails going forward given that the USCO considers these techniques to effect a fair-use analysis.*
- The knowing use of pirated or illegally accessed works as training data weighs against a fair-use defense, but is not determinative. *A number of training data cases allege that the model developer knew that the data it was using was obtained illegally. We expect that the plaintiffs in those cases will cite the USCO position to support their claims.*
- When assessing the “effect of the use on the potential market” (the fourth fair-use factor), courts should consider whether the AI model is generating works in a similar style or category as the original work, thereby diluting the market for that work. *Typically, courts focus only on the market impact where the infringing work is the same or substantially similar to the original work (i.e., focusing on lost sales as the harm). By including market dilution as a potential harm, the USCO is taking a broad view of this fair-use factor, especially given that the report acknowledges that copyright does not protect “style” as a separate element. Concern about the ability of AI models to generate works in the style of original text and images, and the speed and scale at which AI systems can generate such works, likely shaped the USCO’s view.*
- Making commercial use of “vast troves” of copyrighted works to produce expressive content that competes with the original works in existing markets, especially where access to the original work was accomplished through illegal access, “goes beyond established fair use boundaries.”
- Government intervention is not necessary at this time. The fair-use doctrine can address many cases, and where fair use is not available, a voluntary licensing market should continue to develop. *The USCO often comments on whether additional*

legislation is needed. Its position here is particularly important given the numerous statements by AI model developers in recent weeks that the Trump administration’s “AI Action Plan” should include an affirmative statement that the use of copyrighted material as training data is fair use.

Political Context

The report was issued amid an unusual political context. The day before the USCO released the report, President Trump fired Dr. Carla Hayden, the longtime Librarian of Congress, and the day after the report was issued, the president fired the Register of Copyrights, Shira Perlmutter. This might explain why the report was couched as a “pre-publication” version and released Friday evening. These are unusual steps for the USCO, perhaps reflecting a rushed release. Whether the firings related to the contents of the report or whether the USCO rushed out the report knowing these terminations were coming remains unknown, but the events are likely closely linked. What that ultimately means for the report and whether it is truly the USCO’s final report on this issue remains to be seen.

Background to the Report

Many generative AI systems are trained on massive troves of data. Through exposure to these datasets, AI models “learn” patterns, structures and relationships within the data, ultimately enabling the model to generate new content that can range from text and images to music and video. This training data is obtained from a variety of sources, including by web scraping. While a number of copyright holders have licensed their data for such usage, most of the content is used by AI models without authorization. This activity has spurred numerous lawsuits, which typically include claims for copyright infringement. The crux of these cases is whether this activity is protected as fair use.

The Views of the USCO

Do AI Models Infringe Training Data?

The report takes the position that **the development and deployment of generative AI systems implicate several of the exclusive rights granted to copyright owners under the Copyright Act**, including the rights to create copies and derivative works. According to the report, multiple acts in the AI development pipeline — such as data collection, curation, training and output generation — can constitute *prima facie* infringement, particularly of the right of reproduction. This can occur, the report explains, when developers initially download and store works to use in training and when they create intermediate copies during training.

Copyright Office Weighs In on AI Training and Fair Use

More significantly, the report addresses whether a model's weights are themselves infringing copies of the underlying works (often called "memorization"). The report acknowledges that this is a highly disputed concept, with model developers and others asserting that models are merely comprised of strings of numbers and therefore weightings cannot be infringing copies, and others asserting that the fact that models sometimes generate outputs substantially similar to training data is evidence to the contrary. The report concludes that **where generated outputs are substantially similar to inputs, there is a "strong argument" that copying the model's weights implicates the reproduction and derivative work rights of the original works.** The report analogizes this to digital files that encode or compress content using mathematical representations, which are copies of the underlying content even though such content is not directly perceivable.

Application of the Fair-Use Defense

The report then addresses whether the unauthorized use of training data, if found to be infringing, is nonetheless protected by the fair-use defense. **The report addresses each of the four statutory fair-use factors:**

- i. **Purpose and character of the use:** Over the last several years, the key issue when analyzing this fair-use factor has been whether the use at issue is "transformative." The report states that **with respect to AI models, transformativeness "is a matter of degree."** On one end of the spectrum is training a generative AI foundation model on a large and diverse dataset to generate a wide range of outputs "across a diverse array of new situations," which is likely to be transformative. On the other end of the spectrum is training an AI model to generate outputs that are substantially similar to copyrighted works in the training dataset, which is unlikely to be transformative. The report acknowledges that most cases fall between these two extremes, and **where a model is trained to produce content that "shares the purpose of [the original work of] appealing to a particular audience," the use is "at best, modestly transformative."**

Most importantly, the report rejects two arguments that are often made regarding the issue of transformative use:

- The use of copyrighted works to train AI models is inherently transformative because it is not for expressive purposes. The report states that this argument is "mistaken," since models absorb "the essence of linguistic expression" (*i.e.*, how words "are selected and arranged at the sentence, paragraph, and document level.")
- AI training is inherently transformative because it can be analogized to human learning. The report notes that this analogy rests on the faulty premise that fair use is a defense for

all acts if those acts are performed for the purpose of learning. The report cites as an example that a student could not rely on fair use to copy all of the books at a library. The report also comments that the analogy fails because while humans only retain imperfect impressions of the works they have experienced, "filtered through their own unique personalities, histories, memories, and worldviews," generative AI training involves the creation of perfect copies "with the ability to analyze works nearly instantaneously." According to the report, this is significant since **the structure of exclusive copyright rights is premised on certain human limitations.**

The report states that knowingly using a dataset that consists of such pirated or illegally accessed works should weigh against fair use without being determinative.

- ii. **Nature of the copyrighted work:** Not surprisingly, the report notes that where the works involved for training data are more expressive, or previously unpublished, this factor will disfavor fair use.
- iii. **Amount and substantiality of the portion used:** The report acknowledges that for certain forms of AI training, copying entire works may be necessary. While this will ordinarily weigh against a finding of fair use, the report acknowledges this is not dispositive if the use is transformative.³

The report also considers the amount of the original work made available to the public through AI-generated outputs, noting the disagreement among commenters regarding how often original works are materially replicated and how difficult it would be to implement guardrails to prevent that. Nonetheless, the report concludes that **where a model developer or deployer implements guardrails to block copyrighted content from being generated, the third factor will weigh less heavily against fair use.** Guardrail efforts include, among other techniques: blocking prompts likely to reproduce copyrighted content; training protocols designed to make infringing outputs less likely; and internal system prompts that instruct a model not to generate names of copyrighted characters or create images in the style of living artists.

- iv. **Effect on the market:** The report takes a broad view on the market impact factor, identifying three potential market effects to consider in a fair-use analysis:
 - **Lost sales and direct substitution.** According to the report, the most straightforward form of market harm occurs when AI-generated outputs substitute for the original works, leading to lost sales. This is particularly clear

³ The report cites the *Google books case* (where Google copied entire books to create a searchable database) as an example of such a situation. *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).

Copyright Office Weighs In on AI Training and Fair Use

when a model can generate outputs that are verbatim or substantially similar copies of works from its training data, and those outputs are readily accessible to end users.

- **Market dilution and competition in the same class of works.** The USCO takes a broad view of market impacts and includes the risk of market dilution, namely where **AI-generated outputs — though not identical to any specific work — compete in the same market as the original works.** This can occur when AI models are trained on a body of works and then generate new works in the same style, genre or category, thereby increasing competition and potentially reducing the market value of the originals. In this analysis, the report highlights as a factor the USCO's view that the speed and scale at which AI systems can generate content mean that AI has the unique potential to flood a market.
- **Lost licensing opportunities.** Finally, the report identifies the market harm arising from a potential lost revenue opportunity for rights holders who might have licensed their works for training purposes. The report points to the nascent market for licensing content for training data, and concludes that where licensing options exist or are likely to be feasible, this consideration will disfavor a finding of fair use.

Some courts have considered the public's benefit from the infringing activity when assessing this fourth fair-use factor. The report acknowledges that AI can produce many societal benefits, but that it can also harm the public by impeding the growth of the creative economy. Overall, the report concludes that there are no copyright-related benefits that weigh in favor of fair use.

The report concludes that **there will not be a single answer regarding whether the unauthorized use of copyright materials to train AI models is fair use.** As the USCO does in considering transformative use, the report provides two ends of the spectrum: On one end are uses for noncommercial research that do not enable portions of the original works to be

reproduced in the outputs, which is likely to be a fair use. On the other end is the copying of expressive works from pirated sources to generate unrestricted content that competes in the marketplace, when licensing is reasonably available, which may not be a fair use.

Licensing

The report summarizes the comments the USCO received on the feasibility of a licensing scheme for AI training data. In general, AI developers commented that such a scheme would be too expensive and cumbersome, while rights holders countered that even if this were true, the expense should be seen as the price of doing business, and the burden of licensing should not excuse unauthorized use of copyrighted material. The report acknowledges that while there have been a number of one-off AI training data licensing agreements, this solution may not be scalable and that other approaches — such as collective licensing — may be required. The report noted the difficulty in assessing the current licensing market because it may be distorted by the unsettled legal questions about fair use. Finally, the USCO agreed with the general consensus among commenters that a compulsory licensing scheme with fixed royalty fees should be avoided.

The report recommends allowing the licensing market to continue to develop without government intervention.

Looking Forward

The USCO issues the report at a time when more than 40 cases are pending relating to the issues raised by using copyrighted material without authorization to train AI models. While the report merely reflects the USCO's views, it will likely shape the ongoing fair-use debate, with each side finding points in the report to support its position.

As noted, whether this report is indeed the final USCO position or whether the Trump administration pressures a new register of copyrights to adopt a different approach and issue a revised report remains to be seen.