*Perspectives from FSF Scholars*
*June 3, 2025*
*Vol. 20, No. 24*

# Use of Copyrighted Content to Train AI Models Requires Owners' Consent

**by**

**Seth L. Cooper ***

On May 9, the Copyright Office released an important, and already controversial, pre-publication version of its upcoming report on the use of copyrighted content to train generative artificial intelligence (AI) models. The Office's pre-publication report rightly recognizes that processes for developing AI models give rise to copyright violations when those processes involve the creation of new copies or derivatives of protected content. Copyright law requires AI service providers to obtain an owner's permission to make use of copyrighted content to train their AI models.

Notably, the pre-publication report provides no support for establishing blanket exceptions to copyright protections when content is used for text and data mining (TDM) techniques or for training AI models. Similarly, the report finds that providing copyright owners an "opt out" from such blanket exceptions is antithetical to U.S. copyright law. If AI service providers want to train their models using valuable content that doesn't belong to them, they must pay the owners for it – or put up "fair use" defenses to infringement claims. These conclusions are consistent with the understanding that copyrights are private property rights. As the report recommends, Congress should show restraint and let case law develop to provide clarity on copyright and generative AI.

For its Artificial Intelligence Study, the Copyright Office is releasing its Report on *Copyright and Artificial Intelligence* in multiple parts. The report's first part addresses issues posed by digital replicas, and its second part addresses the copyrightability of material generated using AI systems. Its soon-to-be-finalized next installment – for which a pre-publication version was released on May 9 – is on the use of copyrighted works to train generative AI models. Interest in the report is heightened because the use of copyrighted content to train AI models is the subject of infringement lawsuits pending across the country.

The starting point for analyzing the use of copyrighted content in developing generative AI models is the private property rights basis for copyrights. Owners of copyrighted property have the exclusive right to make determinations regarding its use. These property rights premises are reflected in the Constitution's Article I, Section 8 Copyrights Clause, which grants Congress the power "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." The same premises are embedded in the Copyright Act of 1976. Section 106 of the Act expressly secures copyright owners' exclusive rights to authorize reproductions of the copyrighted works, preparation of derivative works, distribution of copies, as well as displays or public performances of the works. And under Section 501: "Anyone who violates any of the exclusive rights of the copyright owner … is an infringer of the copyright."

In its Notice of Inquiry, the Copyright Office defines generative AI as an application that produces expressive outputs such as text, images, audio, or video in response to "prompts" or instructions by a human user. AI models are developed, in oversimplified terms, by collecting and copying data sets that are curated and fed into the models. During training, the AI model's outputs are evaluated and the model's weights are adjusted to achieve desired performance.

Importantly, the pre-publication version of its report recognizes that "[c]reating and deploying generative AI systems using copyright-protected material involves multiple acts that, absent a license or other defense, may infringe one or more rights." In particular, it finds that copyright owners' exclusive right of reproduction is implicated by acts of acquiring and curating datasets for training materials, including by downloading, transferring, reformatting, and modifying the content that results in new copies or copyrighted works or substantial parts of them.

Additionally, the report recognizes that the right of reproduction is implicated by AI model training processes such as downloading datasets and copying them to high-performance storage before training. Generating new copies of works or substantial portions of works that are reproduced as they are "shown" to the AI model in batches also implicates the right of reproduction if the new copies persist long enough to constitute infringement. The report also acknowledges that the right of reproduction may be implicated by "providing training examples, measuring the model's performance against expected outputs, and iteratively updating weights to improve performance" during AI model training. As the report notes, an AI developer's model weights may "contain copies of works in the training data."

Furthermore, the report recognizes that actions taken to develop AI models that result in adaptations of protected works or transformation of substantial expressive parts of them into new

formats or media also may implicate copyright owners' exclusive rights to prepare derivative works under Section 106 of the 1976 Act.

Consistent with the understanding that the owners of copyrighted content have exclusive rights to make determinations about the use of their intellectual property, the Copyright Office's pre-publication report correctly affirms that when AI developers make new copies or derivatives of protected content, the consent of the owners is required.

However, some parties in the proceeding have advocated for the establishment of blanket exceptions to copyright protections for text and data mining (TDM) techniques or for training AI models – subject to an "opt out" or reservation of rights by copyright owners. The Copyright Office's Notice of Inquiry section on AI model training poses the question: "Should copyright owners have to affirmatively consent (opt in) to the use of their works for training materials, or should they be provided with the means to object (opt out)?" And it asks an immediate follow-up question: "Should consent of the copyright owner be required for all uses of copyrighted works to train AI models or only commercial uses?"

Those questions reflect the 2019 European Union's Directive on Copyright in the Digital Single Market (DSM Directive), which provides copyright exceptions for text and data mining (TDM). TDM techniques are computer-based processes for identifying and analyzing patterns, trends, and relationships derived from substantial volumes of copied text or other data. As the Copyright Office Report observes: "TDM methods predate the current forms of generative AI. They are not necessarily 'generative' in the sense of producing new expressive material but involve some of the same steps, particularly in the creation and curation of datasets." The DSM Directive provides a blanket exception to copyright for the use of TDM for purposes of scientific research. The DSM Directive also provides a blanket exception for the use of TDM for any other purpose, so long as the copyright owner has not expressly reserved rights to the use of his or her works in TDM.

Effectively answering the questions posed in the Notice, the pre-publication report states: "As to the possibility of an opt-out mechanism, the Office agrees that requiring copyright owners to opt out is inconsistent with the basic principle that consent is required for uses within the scope of their statutory rights."

Indeed, providing copyright owners special means to object to third-party uses of their content to develop AI models is unnecessary because such owners already have the right to exclusively determine whether to allow or disallow such copying (or preparation of derivatives) for training materials. Under existing law, copyright owners can license their works for TDM and training AI models, or they may refuse to license them. Copyright owners enjoy those rights under the law, without any requirement to affirmatively object to the use of their works to train AI models.

Moreover, Congress should not grant blanket copyright exceptions for TDM *or* training AI models. Blanket exceptions are contrary to the concept of exclusive rights that are embedded in the Copyright Act of 1976 and that can be traced directly back to the Constitution of 1789. Establishing blanket exceptions for TDM or training AI models would significantly undermine those rights by granting third parties what is effectively a royalty-free license to copy works,

prepare derivatives, or exercise other rights of ownership. Such exceptions would expropriate tremendous economic value from copyright owners and transfer it to third parties. AI service providers interested in training their models using copyrighted content should obtain licenses from the owners just like service providers do in other sectors of our economy.

Although U.S. copyright law contains no "opt-out" or blanket exception to copyright protection for TDM or AI model training, the law does provide two much narrower provisions that may provide legal protections to AI service providers from liability for their unauthorized use of copyrighted content to train their models.

First, Section 1201 of the Digital Millennium Copyright Act (DMCA) provides a triennial rulemaking process conducted by the Copyright Office for granting temporary – and renewable – exemptions from the law's prohibition on circumventing technological protection measures that protect access to copyright owners' works. In 2021, for instance, an anti-circumvention exemption was granted for movie DVDs and Blu-Ray discs to enable TDM for scholarly research projects that the Register of Copyrights found were likely to be non-infringing uses.

Second, the "fair use" doctrine codified in Section 107 of the Copyright Act and expounded through case law provides an affirmative defense to copyright infringement. It is based on four non-exclusive factors applied by courts in specific cases for evaluating whether the use of a copyrighted work is "fair." Those factors include: (1) "the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes"; (2) "the nature of the copyrighted work"; (3) "the amount and substantiality of the portion used in relation to the copyrighted work as a whole"; and (4) "the effect of the use upon the potential market for or value of the copyrighted work."

In several pending copyright infringement cases, AI service providers that allegedly or admittedly trained their models on unauthorized copies of protected content have raised "fair use" defenses. Congress should allow the case law regarding the use of copyrighted content in training AI models to develop in the courts rather than prematurely shortcut the law's development by making legislative changes. Prospective court decisions on what constitutes "fair use" in infringement cases involving generative AI model training will provide greater certainty to copyright owners and AI model providers.

In sum, the Copyright Office pre-publication report straightforwardly applies the law to AI model development processes in recognizing that those processes may rise to the level of copyright violations when protected content is copied or derivatives are made without the owner's consent. Its rejection of blanket exemptions and "opt outs" is also based on a straightforward understanding of the law. Hopefully, the final version of its report on training of AI models using copyrighted content should track with those basic pillars of U.S. copyright law and continue to favor judicial development of our nation’s copyright law over the importation of foreign blanket exemptions.

Importantly, the pre-publication report's recognition that AI model development processes give rise to copyright violations when protected content is used to make copies or derivatives without the owner's consent is consistent with private property rights. Its rejection of blanket exemptions

and "opt outs" also respects the private property rights of copyright owners. When it prepares the final version of its report, the Copyright Office should steadfastly support those pillars of U.S. copyright law and favor the law's further development in the courts.

* Seth L. Cooper is Director of Policy Studies and a Senior Fellow of the Free State Foundation, a free market-oriented think tank in Potomac, MD. The views expressed in this *Perspectives* do not necessarily reflect the views of others on the staff of the Free State Foundation or those affiliated with it.

**Further Readings**

Seth L. Cooper, "Copyright Office Report Confirms Copyrightability of AI-Generated Works," *Perspectives from FSF Scholars*, Vol. 20, No. 9 (February 19, 2025).

Seth L. Cooper, "AI-Generated Copies of Creative Works Can Infringe Copyrights," *Perspectives from FSF Scholars*, Vol. 19, No. 42 (November 22, 2024).

Seth L. Cooper, "The 'NO FAKES Act' Would Protect Americans' Rights Against Harmful Digital Replicas," *Perspectives from FSF Scholars*, Vol. 19, No. 32 (August 27, 2024).

Seth L. Cooper, "It Sounds Like Generative AI Music Services Are Infringing Copyrights," *Perspectives from FSF Scholars*, Vol. 19, No. 24 (July 22, 2024).

Seth L. Cooper, "The 'No AI Fraud Act' Would Secure IP Rights Consistent With the First Amendment," *Perspectives from FSF Scholars*, Vol 19, No. 3 (January 26, 2024).

Seth L. Cooper, "Copyright Case Affirming Human Creativity Sets the Stage for AI Issues," *Perspectives from FSF Scholars*, Vol. 18, No. 49 (November 2, 2023).