Lowenstein Sandler

Client Alert

IP & Tech Transactions

May 12, 2025

US Copyright Office Releases Long-Awaited Report on Generative AI Training and Copyright Law

By Bryan Sterba, Matt Savare, and Mark P. Kesslen

Overview

On May 9, the U.S. Copyright Office issued a prepublication version of Part 3 of its multipart report titled "Copyright and Artificial Intelligence: Generative AI Training," addressing the use of copyrighted works in the development of generative AI systems. The report responds to congressional inquiries and stakeholder interests, providing a comprehensive analysis of the technical, legal and policy issues raised by the use of copyrighted content in the training of generative AI models. The office's analysis is intended to inform policymakers and stakeholders as they seek to balance technological innovation with the protection of creative works. The analysis spends much time addressing the deficiencies in many of the most heavily advocated arguments in favor of a fair use defense in generative AI training, asserting that at least some (if not most) uses of copyrighted content to train generative models likely do not constitute fair use. Although the report is not legally binding, the position of the Copyright Office can be considered persuasive authority by courts and is often used and cited in legislative proceedings when Congress crafts laws. Moreover, since there is no controlling legislation or case law on the subject, the report may take on additional weight until the issue is resolved through the normal (but slow) legislative or judicial processes.

Technical Analysis of Generative AI Training

The report provides a detailed technical background on how generative AI models are developed, focusing on the use of large datasets of copyrighted works to train machine learning models. It describes the phases of training (pretraining, fine-tuning, etc.); the acquisition and curation of data; and the concept of "memorization," where models may retain and reproduce elements of their training data. The report notes that while AI models are designed to generalize from data, there is evidence that some models can output verbatim or substantially similar copies of training data, raising concerns about potential infringement.

Prima Facie Infringement

The office concludes that multiple acts involved in the development and deployment of generative AI systems, such as data collection, curation, training, and output generation, implicate the exclusive rights of copyright owners, particularly the rights of reproduction and preparation of derivative works. The process of assembling training datasets, making copies for training, and the potential for models to memorize and reproduce protected expression are all identified as acts that, absent a license or defense, generally constitute infringement.

Application of Fair Use

The report provides an extensive analysis of the fair use doctrine as applied to generative AI training, structured around the four statutory factors set forth in the U.S. Copyright Act of 1976 and recently analyzed by the U.S. Supreme Court in the *Warhol* case:

Purpose and Character of the Use:

- The office finds that training a generative AI model on a large and diverse dataset can often be transformative, particularly when the resulting model is used for research, analysis, or non-substitutive tasks. However, the degree of "transformativeness" is context-dependent. Training models to generate outputs that are substantially similar to copyrighted works, or that compete directly in the same market, is less likely to be considered transformative.
- The report emphasizes that commerciality is a relevant consideration, and that the activities of downstream actors (e.g., commercial deployment of models trained on noncommercial datasets) may affect the analysis.

Nature of the Copyrighted Work:

- Use of highly creative or unpublished works weighs against fair use, while use of factual or functional works may weigh in favor. The office notes that generative AI models are often trained on a mix of expressive and functional works.

Amount and Substantiality of the Portion Used:

- Generative AI training typically involves copying entire works, which ordinarily weighs against fair use. However, where the purpose is transformative and the copying is necessary to achieve that purpose (e.g., for certain types of model training), this may be reasonable. The office distinguishes between cases where outputs make only limited use of the training data and those where substantial portions are made available to the public.

Effect on the Market:

- The report identifies several forms of market harm to the copyrighted works, which the office states is the most significant factor when analyzing whether a use is fair. Such harms include damages that courts have traditionally recognized in fair use cases, such as lost sales, lost licensing opportunities, and the impact of AI-generated outputs that serve as substitutes for the original works. The office, acknowledging that it is in "uncharted territory," also introduces the novel concept of market dilution, claiming that AI-generated works (even if they are not substantially similar to specific works) can be problematic because the flood of AI-generated works creates greater competition for sales of the copyrighted works used in training. The office claims: "If thousands of AI-generated romance novels are put on the market, fewer of the human-authored romance novels that the AI was trained on are likely to be sold." Extending this reasoning even further, the office claims that "[e]ven when the output is not substantially similar to a specific underlying work, stylistic imitation made possible by its use in training may impact the creator's market." These statements regarding market dilution and stylistic imitation, which do not appear in controlling case law or the text of the Copyright Act, will likely be a lightning rod for criticism from tech companies and, if adopted by the courts or Congress, will almost certainly weaken any proffered fair use defense.

The office concludes that the fair use analysis is highly fact-specific and that some uses of copyrighted works for generative AI training will qualify as fair use, while others will not. Uses for noncommercial research or analysis that do not enable reproduction of protected content are more likely to be fair, whereas commercial uses that generate competing expressive content, especially through illegal access or uses of copyrighted content from pirate sites, are unlikely to qualify. The office does not directly address the hot-button issue as to whether web-scraping publicly available data weighs in favor or against a fair use defense, but does state in a footnote that "'[p]ublicly available' is not synonymous with 'authorized.'"

Licensing and Policy Considerations

The report surveys the current landscape of voluntary licensing of data for generative AI training, noting the emergence of both individual and collective licensing agreements in sectors such as music, news, and images. Although voluntary licensing is feasible in some contexts, the office acknowledges significant practical challenges in licensing the full range of works needed for large-scale generative AI training.

The office discusses potential statutory approaches such as compulsory licensing, extended collective licensing (ECL), or opt-out regimes. It cautions that compulsory licenses are a significant derogation of copyright owners' rights and should be considered only in cases of clear market failure. ECL may be appropriate in limited circumstances where voluntary licensing is unworkable.

Dismissal of Register of Copyrights

Following the release of this report, the Trump administration has removed the register of copyrights Shira Perlmutter. It remains to be seen whether this change in Copyright Office leadership was related to the report, and whether the change will result in revisions prior to the release of its official version.

Conclusion

The office reaffirms that the existing legal framework, particularly the fair use doctrine, is capable of addressing the challenges posed by generative AI training. The office emphasizes the importance of balancing innovation and the protection of creative works, supporting both the technology and creative sectors. Effective licensing solutions are seen as critical to ensuring that generative AI benefits innovators, creators, and the public alike.

The office recommends allowing the voluntary licensing market for training data to continue developing without immediate government intervention. Should market failures arise in specific sectors, targeted solutions such as ECL may be considered. The office encourages further guidance on antitrust issues related to collective licensing and highlights the need for ongoing monitoring as technology, law, and markets evolve.

Those with additional questions about the legal implications of the report may contact the authors of this alert.

Contacts

Please contact the listed attorneys for further information on the matters discussed herein.

BRYAN STERBA
Partner
T: 973.597.2386
bsterba@lowenstein.com

MATT SAVARE
Partner
Chair, Commercial Contracts
T: 646.414.6911
msavare@lowenstein.com

MARK P. KESSLEN
Partner
Chair, Intellectual Property Group
T: 646.414.6793 / 973.597.2330
mkesslen@lowenstein.com

This Alert has been prepared by Lowenstein Sandler LLP to provide information on recent legal developments of interest to our readers. It is not intended to provide legal advice for a specific situation or to create an attorney-client relationship. Lowenstein Sandler assumes no responsibility to update the Alert based upon events subsequent to the date of its publication, such as new legislation, regulations and judicial decisions. You should consult with counsel to determine applicable legal requirements in a specific fact situation. Attorney Advertising.