AAP | ASSOCIATION OF AMERICAN PUBLISHERS

Electronic submission via *www.regulations.gov*

October 30, 2023

Artificial Intelligence and Copyright (Docket No. 2023-6)
Notice of Inquiry and Request for Comments

The Association of American Publishers (AAP), a not-for-profit organization, represents the leading book, journal, and education publishers in the United States on matters of law and policy, advocating for outcomes that incentivize the publication of creative expression, professional content, and learning solutions. AAP's members range from major commercial book and journal publishers to small, non-profit, university, and scholarly presses, as well as leading publishers of educational materials and digital learning platforms—publishers that curate and bring to market works that educate, entertain, and inform.[1]

Copyright is the engine of free expression, and the foundation of the publishing industry—the original industry of free expression. AAP believes that the exclusive rights accorded by copyright to the author, such as the right of reproduction and the right to authorize derivative works, are essential to the integrity and operation of the Copyright Act. It is this divisible bundle of rights that underpins licensing and allows rights holders to profit from their intellectual property and reinvest in the creation of new works. Professionally edited books, long form prose, and rigorously peer-reviewed academic articles provide rich, high-quality materials essential to developing AI systems capable of generating trustworthy and reliable outputs. Allowing AI systems developers to continue to pillage the creative, intellectual, and rich works of authors that publisher-investment bring to market—robbing the author and publisher of the return on their investment—threatens the continued viability of the cycle of creation and dissemination of such works, potentially leaving AI systems to be trained on low quality, inaccurate materials.

The un-permissioned and uncompensated use of copyrighted works to create the datasets for training generative AI (Gen AI) systems presents a direct assault on the

[1] The U.S. publishing industry supports an extensive network of American businesses and thousands of jobs, with revenue of $28.10 billion for 2022. AAP StatShot Annual Report: Publishing Revenues Totaled $28.10 Billion for 2022 - AAP (publishers.org). The publishing industry is also an integral part of the broader U.S. copyright industries, which collectively added more than $1.8 trillion in annual value to U.S. gross domestic product in 2021. *Copyright Industries in the U.S. Economy: The 2022 Report,* by Robert Stoner and Jéssica Dutra of Economists Incorporated, prepared for the International Intellectual Property Alliance (IIPA), (December 2022), https://www.iipa.org/files/uploads/2022/12/IIPA-Report-2022_Interactive_12-12-2022-1.pdf. Beyond these important economic contributions, an independent and thriving publishing industry supports the nation's political, intellectual, and cultural systems.

livelihoods and professions of authors, publishers, and all those who are integral to the publishing endeavor. Some argue that in order to maximize the benefits AI technology offers, AI systems developers should be allowed free rein to innovate, that they should bear no responsibility to the rights holders whose works are used to train their AI systems—as otherwise they will not successfully compete against foreign counterparts. These arguments not only ignore the impact of such a policy on the creation of high-quality works but also the viability of licensing markets. The protection of exclusive rights, from which licensing derives, is the best way to ensure the continued competitiveness of American industry.

This submission responds to the questions posed in the Copyright Office NOI and RFC, but ultimately rests on three key points:

1. The wholesale reproduction of copyrighted works for purposes of training and developing AI systems is infringement.
2. Should case law develop in a manner that finds wholesale copying to be permissible for training AI systems, legislation should make clear that unlicensed ingestion of copyrighted materials for purposes of training does not qualify as fair use. We also believe the Copyright Office and Congress should monitor legal, business, and technological developments closely to ensure copyright owners' exclusive rights remain vital and robust.
3. Any framework intended to promote AI development must not diminish the copyrights of the authors and publishers whose works are essential to a free, flourishing, and well-informed society.

Because the Office's inquiry is focused on the intersection of AI and copyright, not addressed in our comments are the risks associated with privacy, cybersecurity, discrimination, and bias, and how these risks, if unaddressed, may result in reputational, financial, and other harms to individuals. We also do not address concerns about how the potential misuse of AI technologies could sow misinformation or disinformation among the general public and adversely impact important democratic processes such as a free and fair election. These concerns, along with safeguarding the copyright framework, are important reasons for why AI systems should be regulated.

Finally, we appreciate the Office's recognition of the urgency of this policy study, given the rapid developments involving Gen AI with which policymakers, courts, and stakeholders must grapple. We encourage the Office to consider publishing its factual findings in advance of the complete study—in particular, the sections of the study addressing legal and factual background matters and descriptions of the issues. Having an authoritative overview of these background matters published as soon as they are complete would be of tremendous benefit to all stakeholders.

General Questions:

1. As described above, generative AI systems have the ability to produce material that would be copyrightable if it were created by a human author. What are your views on the potential benefits and risks of this technology? How is the use of this technology currently affecting or likely to affect creators, copyright owners, technology developers, researchers, and the public?

As with all technologies, Gen AI systems present both risks and opportunities for all sectors of the economy. There is, however, already significant risk of copyright infringement of copyrighted works used to train Gen AI systems without consent, credit, or compensation to the rights holders. As Gen AI systems become increasingly sophisticated, there is also the risk that their outputs will become virtually indistinguishable from the works created by authors and artists and, given the speed at which these artificial systems can generate output, will unfairly compete with human creators. Such unfair competition may come in the form of devaluing human creative expression, displacing authors altogether, or flooding the market with low-value competing works.[2] In the field of scientific, technical, and medical research, Gen AI systems trained on materials other than the Version of Record (VoR), which reflect corrections to or retractions of articles reporting on research outcomes, pose a significant risk of generating flawed, biased, or inaccurate outputs, and systems built upon such flawed outputs, in turn, may endanger the health, safety, and security of all Americans.

The publishing industry is keenly aware that Gen AI systems, trained on appropriately licensed works, may prove useful as tools to enhance productivity and efficiency. Some publishing companies are experimenting with the use of AI systems in their businesses. For instance, professional and scholarly publishers are exploring the use of Gen AI systems to enhance research such as identifying relevant papers and providing reliable summaries, potentially allowing the researcher to identify and review prior research in the field as well as identify new avenues of query more expeditiously.[3] Gen AI-assisted research may aid in the discoverability of scholarly works, making otherwise difficult to find scholarship more readily accessible.[4] As the capabilities and risks associated with Gen AI systems are not yet fully understood, publishers are proceeding

---

[2] Robert McMillan, "AI Junk is Starting to Pollute the Internet," Wall Street Journal (July 12, 2023), https://www.wsj.com/articles/chatgpt-already-floods-some-corners-of-the-internet-with-spam-its-just-the-beginning-9c86ea25.

[3] See *Scopus AI: Change the Way you View Knowledge*, ELSEVIER https://beta.elsevier.com/products/scopus/scopus-ai.

[4] Kris Lykke, *The Future of AI in Publishing*, JOHNS HOPKINS UNIVERSITY PRESS (Aug. 3, 2023), https://www.cnn.com/2023/08/10/tech/ai-generated-books-amazon/index.html.

with care as they test and experiment with the different ways these technologies may enhance business processes.

2. Does the increasing use or distribution of AI-generated material raise any unique issues for your sector or industry as compared to other copyright stakeholders?

The works AAP members publish are especially valuable to training Gen AI systems. Professionally edited books, rich in prose, provide high quality expression for "long range context modeling research and cohesive storytelling,"[5] and models trained on these materials produce higher quality outputs than models trained on social media posts, or the limited character musings embodied in what used to be called "tweets." Likewise, well-edited and peer-reviewed textbooks, educational materials, and scientific and scholarly publications are indispensable to training AI models to perform better at tasks in specific knowledge domains, and indispensable to developing trustworthy AI technologies. In recognizing this value, we urge the Copyright Office to appreciate that the source of the value is the well-crafted creative expression of the author. Such works should not be commoditized as mere "data" to be indiscriminately ingested by an AI system, without consent, credit, or compensation to the authors and publishers of high-quality prose.

AI-generated material will raise issues unique to the publishing industry. For instance, an unscrupulous person or organization may use the technology to quickly generate output that mimics the "voice" or style of an author, attach the author's name to the AI-generated output, and falsely market and sell the AI-generated output as that author's work—all without the knowledge of the author, and in the most egregious of circumstances, potentially resulting in reputational damage to the author.[6]

3. Please identify any papers or studies that you believe are relevant to this Notice. These may address, for example, the economic effects of generative AI on the creative industries or how different licensing regimes do or could operate to remunerate copyright owners and/or creators for the use of their works in training AI models. The Office requests that commenters provide a hyperlink to the identified papers.

[5] Leo Gao et al., The Pile: An 800GB Dataset of Diverse Text for Language Modeling, eprint 2101.00027 (arXiv, cs.CL 2020).

[6] Clare Duffy, *An Author Says AI is 'Writing' Unauthorized Books Being Sold Under her Name on Amazon*, CNN (Aug. 10, 2023, 10:03 AM EDT), https://www.cnn.com/2023/08/10/tech/ai-generated-books-amazon/index.html.

Many of the papers that attempt to address the specific economic effects of Gen AI systems have not focused on the creative industries. Some address the potential impact on the economy but focus primarily on potential gains to be made by increasing productivity.[7] While there are as yet no comprehensive economic studies on the value of the U.S. text-and-data mining (TDM) licensing market (which, as we note below is not equivalent to licensing for Gen AI purposes),[8] it is useful to note that scholarly and professional publishers already license their journal databases for such activities.[9]

4. Are there any statutory or regulatory approaches that have been adopted or are under consideration in other countries that relate to copyright and AI that should be considered or avoided in the United States?

While a few jurisdictions have adopted a copyright exception for text-and-data mining, on the premise that it is necessary to AI research and development, AAP does not believe any exceptions to permit the unlicensed use of copyrighted works to develop AI technologies are necessary or desirable. We note that TDM exceptions were adopted before the rise of Gen AI, and the original aim was to facilitate computational analysis of large amounts of text as may be embodied in scholarly and scientific articles for research purposes.[10] TDM uses do not equate with the use of copyrighted works to train Gen AI models. Licensing solutions remain the better tool for facilitating AI development. Through licensing arrangements, AI systems developers can legitimately access the copyrighted works necessary to training trustworthy and reliable AI models while authors, publishers, and other copyright holders and licensees are appropriately compensated for the use of their works.

Although we do not believe a TDM exception is necessary, below we outline the approach taken in four jurisdictions of which the Office should be aware. We note that the TDM exception defined in the EU Directive on copyright in the Digital Single Market (DSM) differentiates TDM for commercial and non-commercial purposes (specifically, for

---

[7]MICHAEL CHUI ET AL., THE ECONOMIC POTENTIAL OF GENERATIVE AI (Stephanie Strom & David DeLallo eds., 2023), The Economic Potential of Generative AI: The Next Productivity Frontier.

[8]Text Mining Market Trends: Growth & Analysis, 2020 – 2027 (Oct. 2023), https://www.reportsanddata.com/report-detail/text-mining-market.

[9]See e.g., *Text & Data Mining: Harness the Power of Big Data and Analytics*, ACS PUBLICATIONS (Nov. 23, 2022), https://solutions.acs.org/solutions/text-and-data-mining/; *Elsevier Text and Data Mining (TDM) License,* ELSEVIER, https://beta.elsevier.com/about/policies-and-standards/text-and-data-mining/license?trial=true; *Sage Journals Text and Data Mining License Agreement*, SAGE JOURNALS, https://journals.sagepub.com/page/policies/text-and-data-mining-license; *Text Data and Mining*, TAYLOR & FRANCIS, Text and Data Mining - Taylor & Francis (taylorandfrancis.com); *Text Data and Mining*, WILEY, https://onlinelibrary.wiley.com/library-info/resources/text-and-datamining.

[10] Diane McDonald and Ursula Kelly, *Value and Benefits of Text Mining* (Jan. 2, 2020), https://beta.jisc.ac.uk/reports/value-and-benefits-of-text-mining.

scientific research). The exception for commercial TDM activity applies unless the rights holder prohibits the use of their copyrighted works by expressly reserving their rights "in an appropriate manner, such as machine-readable means in the case of content made publicly available online."[11] Lawful access to the copyrighted works is required.

The UK Copyright, Designs and Patent Act 1998 in Section 29A provides for a similar exception and allows "copies for text and data analysis for non-commercial research." While this commercial/non-commercial distinction has been a reasonable approach in certain instances, there is concern that some AI systems developers are either acquiring or creating a non-profit entity to undertake data mining activities which are then merged into the for-profit enterprise of the developer.[12]

The formulation of the TDM exceptions in the copyright laws of Japan and Singapore, however, should be avoided. Japan has a broad TDM exception that does not distinguish between commercial and non-commercial purpose for TDM activities. While the exception allows TDM use only where the user's purpose is "not for enjoying or causing another person to enjoy the ideas or emotions expressed in the work," and requires that such use "not unreasonably prejudice the interests of the copyright owner," the exception is nonetheless concerning as it does not expressly require that the TDM user have lawful access to the work.

The Singapore exception is likewise unduly broad, making no distinction between commercial and non-commercial TDM activities, with rights holder opt-out not permitted. While lawful access is required, if the work copied for the TDM exercise happens to be an infringing copy or even if the infringing copy was sourced from a "flagrantly infringing online location," this is excused so long as the user did not know or have reason to believe that the work accessed is an infringing copy.

With respect to statutory approaches under consideration, the transparency provisions introduced by the EU Parliament to the EU AI Act with respect to the use of copyrighted works in the creation of training datasets for AI systems are worth considering. Under the proposed text, developers of Gen AI systems would be required to "document and make publicly available a summary of the use of training data protected

---

[11]Council Directive 2019/790/EU, art. 18, 2019, Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC .

[12] Andy Baio, *AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability*, WAXY (Sept. 30, 2022), https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/.

under copyright law."[13] As Gen AI systems developers have not been forthcoming as to the source of the copyrighted works used to create the training datasets, a robust transparency requirement should be considered as to provide rights holders with the information necessary to protect and enforce their rights.

5. Is new legislation warranted to address copyright or related issues with generative AI? If so, what should it entail? Specific proposals and legislative text are not necessary, but the Office welcomes any proposals or text for review.

Should case law develop in a manner that finds wholesale copying to be permissible for training AI systems, legislation should make clear that unlicensed ingestion of copyrighted materials for purposes of training does not qualify as fair use.

Aside from that, given the long success and adaptability of the U.S. copyright system to new technologies, there is no reason at this time why the current legal framework cannot accommodate and support the continued development of AI. The marketplace is sufficiently flexible to accommodate existing licensing practices, while also driving new licensing schemes based on the exclusive rights afforded by copyright law. For the same reasons, the industry strongly cautions against creating new exceptions, such as for text-and-data mining (TDM) to purportedly facilitate the use of data embodied in copyrighted works for purposes of AI training.

Market-based licensing solutions are the superior tool for facilitating development of AI systems while respecting and protecting the exclusive rights of authors, publishers, and other copyright owners and licensees. Publishers are opposed to the introduction of a compulsory licensing regime to redress unauthorized uses of copyrighted works to train Gen AI systems. Fundamentally, authors and publishers should remain free to exercise their exclusive rights, to determine how and in what ways their works are to be used, and by whom.

Given the lack of visibility into how and which copyrighted works Gen AI systems developers are using to create training datasets, a limited inquiry by Congress into whether the Copyright Office might engage in a rulemaking exercise to outline a transparency requirement with respect to how a Gen AI systems developer sourced the copyrighted content used to create their training datasets, whether such use was licensed, and from whom may be of use.

[13] Luca Bertuzzi, *AI Act: MEPs close in on rules for general purpose AI, foundation models*, EURACTIV (Apr. 20, 2023), AI Act: MEPs close in on rules for general purpose AI, foundation models – EURACTIV.com (The EU AI Act is currently in the trilogue process, and the outcome remains uncertain).

Finally, we recommend that the Office continue to monitor judicial, international, and industry developments, to determine if a compelling need for action arises.

Training

6. What kinds of copyright-protected training materials are used to train AI models, and how are those materials collected and curated?

Every type of copyright-protected work — text, images, audio, video, and software code — have been used to train AI models, with much of it scraped from the Internet, including from piracy sites. OpenAI's GPT model, for example, was trained on content and material from the Common Crawl dataset,[14] as well as Wikipedia and books (Books1 and Books2 datasets, and more recently Books3).[15] With respect to image generators, some text-to-image models (such as Stability AI's Stable Diffusion model) were trained on large datasets of images created by the Large-scale Artificial Intelligence Open Network (LAION).[16]

6.1. How or where do developers of AI models acquire the materials or datasets that their models are trained on? To what extent is training material first collected by third-party entities (such as academic researchers or private companies)?

Several Gen AI systems have included pirated books in their training datasets.[17] For example, a *Washington Post* analysis of Google's C4 dataset, which has been used to train high-profile Large Language models including Google's T5 and Meta's LLaMA, found that "b-ok.org, a notorious market for pirated e-books that has since been seized

---

[14] *Common Crawl: Overview*, COMMON CRAWL BY AMAZON, https://commoncrawl.org/overview (described as a corpus containing "petabytes of data, regularly collected since 2008," consisting of raw web page data, metadata extracts, and text extracts).

[15] Dave Ver Meer, *Number of ChatGPT Users and Key Stats (2023)*, NAMEPEPPER (Oct. 17, 2023), https://www.namepepper.com/chatgpt-users; Dennis Layton, *ChatGPT – Show me the Data Sources*, MEDIUM (Jan. 30, 2023), ChatGPT — Show me the Data Sources | by Dennis Layton | Medium; *See also*, Kate Knibbs, *The Battle Over Books3 Could Change AI Forever*, WIRED (Sept. 4, 2023), https://www.wired.com/story/battle-over-books3/; Alex Reisner, *Revealed: The Authors Whose Pirated Books Are Powering Generative* AI, THE ATLANTIC (Aug. 19, 2023), https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/.

[16] See, *LAION Roars: The Story of LAION, the dataset behind Stable Diffusion*, THE BATCH (Jun. 7, 2023), The Story of LAION, the Dataset Behind Stable Diffusion (deeplearning.ai).

[17] Reisner, *supra* note 15; Kevin Schaul et al., *Inside the Secret List of Websites that Make AI like ChatGPT Sound Smart*, WASH. POST (Apr. 19, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/; Bernhard Warner, *Extreme Heat Shows the Need for Another Kind of Climate* Investment, N.Y. TIMES (July 22, 2023), https://www.nytimes.com/2023/07/22/business/dealbook/extreme-heat-climate-investment.html.

by the U.S. Justice Department," was among the largest data sources in the dataset.[18] Additionally, "[a]t least 27 other sites identified by the U.S. government as markets for piracy and counterfeits were present in the data set."[19]

6.2. To what extent are copyrighted works licensed from copyright owners for use as training materials? To your knowledge, what licensing models are currently being offered and used?

AAP members were not approached for authorization or license to use their works in the training of Gen AI systems that publicly launched in 2022.[20]

Professional and scholarly publishers already license their databases for TDM use.[21] Where the TDM research is for non-commercial purposes and the researcher already has access to a publisher's database through an institutional license, a licensing fee is typically not required. However, publishers indicate on their websites that where the TDM exercise is for a commercial purpose, licensing fees are required, and the user is advised to contact the publisher. Trade and education publishers are likewise exploring licensing options and have been approached by AI systems developers to discuss such arrangements.

6.3. To what extent is non-copyrighted material (such as public domain works) used for AI training? Alternatively, to what extent is training material created or commissioned by developers of AI models?

Given the lack of transparency with respect to how AI systems developers source the materials used to create training datasets—copyrighted or otherwise—the extent to which non-copyrighted material is used for AI training is not currently accurately ascertainable. Presumably, along with in-copyright works scraped from piracy websites, public domain works are also collected and included in training datasets.

Gen AI systems developers have funded the creation of training datasets by non-profit entities, which datasets are then used for commercial purposes. According to news accounts, LAION, a non-profit entity that created the image-text caption datasets used to train Stable Diffusion and Google's Imagen, was funded by Stability AI.[22]

---

[18]Schaul et al., *supra* note 17.

[19]Schaul et al., *supra* note 17.

[20] *See*, Leah Asmelash, *These Books are Being Used to Train AI. No One Told Authors*, CNN (Oct. 8, 2023), These books are being used to train AI. No one told the authors | CNN.

[21] See e.g., *supra* note 9 (Examples of scholarly publishing TDM licenses).

[22] Baio, *supra* note 12.

6.4. Are some or all training materials retained by developers of AI models after training is complete, and for what purpose(s)? Please describe any relevant storage and retention practices.

It is possible AI systems developers retain the training datasets after training is completed.[23] Assuming that there is some retention, such practice would be useful for conducting assessments and audits on the AI system and its training dataset. For instance, to identify potential risks in the application of the AI system. AI systems developers should be required to keep accurate records regarding the copyrighted works used to create the training datasets, including how the copyrighted works were sourced, whether the use was licensed, and to allow rights holders to interrogate the models to determine whether their copyright protected works were used in creating the training dataset.

7. To the extent that it informs your views, please briefly describe your personal knowledge of the process by which AI models are trained. The Office is particularly interested in:

7.1. How are training materials used and/or reproduced when training an AI model? Please include your understanding of the nature and duration of any reproduction of works that occur during the training process, as well as your views on the extent to which these activities implicate the exclusive rights of copyright owners.

Generally, any AI training process that uses copyrighted works will implicate at least one of the copyright owner's exclusive rights at least once, and often will implicate them multiple times. Even a temporary, incidental copy of a copyright protected work implicates the reproduction right if it is "sufficiently permanent or stable to permit it to be perceived [or] reproduced . . . for a period of more than transitory duration."[24] Such

[23] As noted in our response to Question 7.1, retention of copyrighted works is not necessary to establish unauthorized copying.

[24] 17 U.S.C. § 101. In considering the question of temporary reproductions in depth, this Office concluded that "RAM reproductions are generally 'fixed' and thus constitute 'copies' that are within the scope of the copyright owner's reproduction right" and recommended against "the adoption of a general exception from the reproduction right to render noninfringing all temporary copies that are incidental to lawful uses." DMCA Section 104 Report at 110, 141 (2001). The Office more recently observed, "It appears that the RAM copy doctrine is today firmly established as a matter of case law." Software-Enabled Consumer Products at 19, n.101 (2016).

intermediate copying may be infringing regardless of the final output,[25] and such copying is not nullified if the copies of the works are subsequently deleted.[26]

It is our understanding that training materials may be reproduced when training AI models during the following steps.[27] Not every training process uses all of the following steps, and some training procedures may include additional steps that involve reproduction:

1. **Dataset Acquisition:** A dataset[28] acquired from a third party will be reproduced, and possibly distributed, when it is digitally transmitted from the third party to the operator of the AI model.

2. **Initial Data Loading:** The dataset is loaded from disk storage into memory before the training process begins. This step involves making a temporary reproduction of the dataset in memory.

3. **Tokenization:** The text data from the dataset is tokenized into smaller units (words or subwords) to prepare it for training. Tokenization may be considered either a reproduction or the creation of a derivative work because the Copyright Act defines "copies" to include formats which can only be "perceived, reproduced, or otherwise communicated" with the aid of a machine. During this process, the dataset is reproduced in a tokenized form in memory.

4. **Training:** During training, the dataset may be reproduced, in whole or in part, multiple times. The training process may involve a number of optimization processes such as mini-batching, shuffling, or caching, each of which may involve a temporary reproduction of the dataset.

5. **Validation and Testing:** During the training process, the model's performance is evaluated on separate validation and testing datasets. These datasets are also loaded into memory, resulting in temporary reproductions.

---

[25] Sega Enters. v. Accolade, Inc., 977 F.2d 1510, 1519 (9th Cir. 1992).

[26] Capitol Records, LLC v. ReDigi Inc., 910 F.3d 649, 658 (2d Cir. 2018).

[27] See generally, Nitin Kushwaha, A Guide to Build Your Own Large Language Models from Scratch, (Aug. 4, 2023), https://python.plainenglish.io/a-guide-to-build-your-own-large-language-models-from-scratch-533a1fd55de7; *Demystifying the Training Process of ChatGPT and Other Language Models*, NINE TWO THREE (May 29, 2023), https://www.ninetwothree.co/blog/demystifying-the-training-process-of-chatgpt-and-other-language-models.

[28] The term "dataset" is used here and elsewhere as shorthand for the rich content AI developers seek from our publishing sector to improve their AI system's ability to understand complex language, syntax and content, and to sound more human. The datasets which include books and scientific and medical articles include much more than simply numbers or coding.

6. **Transfer Learning or Fine-tuning:** If the model undergoes transfer learning or fine-tuning using the dataset, the dataset might be reproduced in some form during this process as well.

Some or all of these steps will involve a reproduction of sufficient duration to constitute prima facie infringement, depending on the specific facts of the training process.

7.2. How are inferences gained from the training process stored or represented within an AI model?

7.3. Is it possible for an AI model to "unlearn" inferences it gained from training on a particular piece of training material? If so, is it economically feasible? In addition to retraining a model, are there other ways to "unlearn" inferences from training?

It may be possible that "training" an AI model to "selectively forget" certain elements of its training data is economically feasible. New research presents an early indication that there may be an effective technique for unlearning in generative language models.[29] The researchers do, however, note that while their "technique offers a promising start, its applicability across various content types remains to be thoroughly tested," and "further research is needed to refine and extend the methodology for broader unlearning tasks in LLMs."[30] In the area of privacy, there is also research into machine "unlearning," in which scientists are looking into how "selective amnesia" may be induced in AI systems, allowing individuals to withdraw their information and thereby potentially prevent an AI systems developer from profiting from its unauthorized use of such information.[31] Selective "unlearning" would likely be less economically prohibitive than algorithmic disgorgement, which remedy would require the destruction of the illegally sourced training dataset and any models built with it.[32] How the remedy (i.e., algorithmic disgorgement) would be effectively implemented, however, remains in question.

---

[29] See Ronen Eldan & Mark Russinovich, *Who's Harry Potter? Approximate Unlearning in LLMs* (Oct. 4, 2023), 2310.02238.pdf (arxiv.org).

[30] Id.

[31] See, Tom Simonite, *Now that Machines Can Learn, Can They Unlearn?*, WIRED (Aug. 19, 2021), https://www.wired.com/story/machines-can-learn-can-they-unlearn/; See also, *Deep Unlearning: AI Researchers Teach Models to Unlearn Dat*a, THE BATCH (Sep. 1, 2021), https://www.deeplearning.ai/the-batch/deep-unlearning/.

[32] See, Kate Kaye, *The FTC's New Enforcement Weapon Spells Death for Algorithms*, PROTOCOL (Mar. 14, 2022), https://www.protocol.com/policy/ftc-algorithm-destroy-data-privacy.

7.4. Absent access to the underlying dataset, is it possible to identify whether an AI model was trained on a particular piece of training material?

To our knowledge, the technology to identify particular copyrighted works that were used to train an AI model does not yet exist. Several of the current lawsuits involving the unauthorized use of copyrighted works to train AI models include allegations that the models are able to reproduce verbatim portions of works—and even the entirety of works.[33] Virtual identity or striking similarity between an output generated by an AI model and a copyrighted work may support an inference that the model was trained on that work.[34]

8. Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use? Please discuss any case law you believe relevant to this question.

Fair use is a fact-dependent defense to the charge of infringement, decided on a case-by-case basis, and its nature resists categorical conclusions. That said, there are very limited circumstances in which the use of copyrighted works to train AI models would constitute fair use, given the lack of a transformative purpose combined with the harm to the market for and value of copyrighted works.

We address both points in more detail in response to the Office's questions below, but several broad points are worth noting here.

First, a rule that unauthorized use of copyrighted works to train AI models constitutes fair use essentially compels authors and publishers to subsidize the development of AI models. This subsidization would be unlike what authors implicitly consent to under traditional forms of fair use—authors and publishers recognize that when they disseminate their works to the public, they invite criticism and commentary, they invite scholars to build on the works and historians to document the works. Such uses advance the purpose of copyright law and effectuate the First Amendment.[35]

---

[33] Emilia David, *Universal Music sues AI company Anthropic for distributing song lyrics*, THE VERGE (Oct 19, 2023, 1:19 PM EDT), https://www.theverge.com/2023/10/19/23924100/universal-music-sue-anthropic-lyrics-copyright-katy-perry; Complaint, J.L. v. Alphabet Inc., ¶¶ 15, 111, No. 3:23-cv-3440, N.D. Cali., July 11, 2023.

[34] 4 Nimmer on Copyright § 13D.07 (2023).

[35] See Alan Latman, Fair Use of Copyrighted Works at 7, Committee Print (1958) (noting that one theory supporting the doctrine of fair use was premised on an author's deemed "consent to certain reasonable

At this point in time, there is only speculation as to the net effect Gen AI models will have on copyright and free speech and how its benefits and harms will be distributed. Given this, it would be unfair and inequitable to compel authors and publishers to subsidize the development of specific AI models.

Second, fair use should not facilitate wholesale piracy. As has been widely reported, several Gen AI systems have included pirated books in their training datasets.[36] For example, a *Washington Post* analysis of Google's C4 dataset, which has been used to train high-profile Large Language models including Google's T5 and Meta's LLaMA, found that "b-ok.org, a notorious market for pirated e-books that has since been seized by the U.S. Justice Department," was among the largest data sources in the dataset.[37] Additionally, "[a]t least 27 other sites identified by the U.S. government as markets for piracy and counterfeits were present in the data set."[38]

This fact should weigh against a finding of fair use. The Supreme Court has held that "Fair use presupposes 'good faith' and 'fair dealing,'" and in *Harper and Row*, the fact "that The Nation knowingly exploited a purloined manuscript" weighed against fair use.[39] The Court has never questioned this point, and while some lower courts have declined to follow the lead of *Harper and Row*,[40] others have not.[41]

8.1. In light of the Supreme Court's recent decisions in Google v. Oracle America and Andy Warhol Foundation v. Goldsmith, how should the "purpose and character" of the use of copyrighted works to train an AI model be evaluated? What is the relevant use to be analyzed? Do different stages of training, such as pre-training and fine-tuning, raise different considerations under the first fair use factor?

---

uses of his copyrighted work to promote the ends of public welfare for which he was granted copyright."); Eldred v. Ashcroft, 537 US 186, 219 (2003) (describing fair use as one of copyright law's "built-in First Amendment accommodations").

[36] Reisner, *supra* note 15; Schaul et al., *supra* note 17; Warner, *supra* note 17.

[37] Schaul et al., *supra* note 17.

[38] Schaul et al., *supra* note 17.

[39] Harper & Row, Publrs. V. Nation Enters., 471 U.S. 539, 562–63, 105 S. Ct. 2218, 85 L. Ed. 2d 588 (1985).

[40] Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P., 742 F.3d 17, 27 (2d Cir. 2014), superseded, 756 F.3d 73 (2d Cir. 2014); Nxivm Corp. v. Ross Institute, 364 F.3d 471, 479 (2d Cir. 2004), cert. denied, 543 U.S. 1000 (2004).

[41] Nunez v. Caribbean Int'l News Corp., 235 F.3d 18, 23 (1st Cir. 2000) ("An unlawful acquisition of the copyrighted work generally weighs against a finding of fair use; no such theft occurred here."); Atari Games Corp. v. Nintendo Am., Inc., 975 F.2d 832, 843 (Fed. Cir. 1992).

There are numerous ways the Copyright Office and courts could evaluate the "purpose and character" of the use of copyrighted works to train an AI model, and the specific facts of a given case will play a key role in shaping this analysis. We offer one potential analysis below.

In *Warhol*, the Supreme Court highlighted the role of justification in analyzing the "purpose and character" under the first fair use factor. The Court relied on its earlier discussion of parody and satire to illustrate justification. In *Campbell*, the Court distinguished parody (which targets an author or work for humor or ridicule) from satire (which ridicules society but does not necessarily target an author or work), and explained "[p]arody needs to mimic an original to make its point, and so has some claim to use the creation of its victim's (or collective victims') imagination, whereas satire can stand on its own two feet and so requires justification for the very act of borrowing."[42] Thus, justification is greater when a specific work is necessary to the secondary purpose, and it is lesser when a specific work is fungible to the purpose—when there are alternatives or substitutes that could achieve the same purpose.

Even prior to *Warhol*, Circuit Courts have relied on justification to guide fair use decisions. For example, in denying fair use to a defendant who had used plaintiff's photographs of a secret celebrity wedding ceremony to illustrate a magazine article covering the event, the Ninth Circuit explained that "the controversy here has little to do with photos," nor were the photos "even necessary to prove that controverted fact"—rather they merely portrayed the subject of the article. This was insufficient to find transformation.[43]

Similarly, when dismissing a fair use claim regarding the incorporation of a plaintiff's copyrighted comedy routine into the defendant's play, the Second Circuit characterized the purpose served by the use of the routine as a "McGuffin," explaining that defendant's play needed *something* to advance its narrative, but the specific content of that something "appears irrelevant to this purpose."[44] The court further emphasized that "Such unaltered utilization of a purportedly copyrighted work, which bears no relevance to the original work, necessitates justification to qualify for a fair use defense" and concluded that nothing in the record demonstrated such justification.

---

[42] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1275-76 (2023) (quoting Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 580-81 (1994).)

[43] Monge v. Maya Magazines, Inc., 688 F.3d 1164, 1175 (9th Cir. 2012).

[44] TCA TV Corp. v. McCollum, 839 F.3d 168, 182 (2d Cir. 2016).

A comparison of AI training to existing computational analysis decisions like *Google Books*[45] and *HathiTrust*[46] reveals a similar lack of justification for the use of copyrighted works in training AI models. *Google Books* and *HathiTrust* both involved the digitization of copyrighted books in order to create full-text search of those books.[47] Writing for the majority in *Google Books*, Judge Leval explained,

> "As with *HathiTrust* (and *iParadigms*), the purpose of Google's copying of the original copyrighted books is to make available significant information about those books, permitting a searcher to identify those that contain a word or term of interest, as well as those that do not include reference to it. In addition, through the ngrams tool, Google allows readers to learn the frequency of usage of selected words in the aggregate corpus of published books in different historical periods. We have no doubt that the purpose of this copying is the sort of transformative purpose described in Campbell as strongly favoring satisfaction of the first factor."[48]

The necessity of copying a work to enable full-text search of the work is apparent: one could not accomplish the purpose by copying a different work. For example, to search for a phrase that appears in *The Hemingses of Monticello* by Annette Gordon-Reed, one needs to copy the full text of *The Hemingses of Monticello* and not, say, Ron Chernow's *Alexander Hamilton*.

By contrast, the purpose of using copyrighted works to train AI models is to extrapolate patterns and statistical inferences about a domain broader than the works themselves—such as expression in the English language as a whole. Additionally, the purpose of using an AI model is not, as with the HathiTrust and Google search engines, to learn information about any specific book that was copied, but to generate responses that draw upon broad-based knowledge.[49]

---

[45] Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).

[46] Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014).

[47] The full-text search function was just one of three uses at issue in *HathiTrust*; the court also considered fair use of HathiTrust's provision of accessible format copies for the print-disabled and provision of replacement copies for preservation purposes.

[48] Authors Guild, 804 F.3d at 217. Notably, the Second Circuit described the dispute in Google Books as one that "tests the boundaries of fair use." Id at 206.

[49] That's not to say a user couldn't prompt a Large Language Model to answer a question about one of the underlying books that was used for training (e.g., "What is the plot of Catcher in the Rye?"), but given the indeterminately broad range of prompts that users can and do provide to LLMs, this exception does not prove the rule.

No one specific work will be necessary for this purpose, and developers will generally have a range of substitutes or alternatives that achieve the same purpose (including creating their own training materials). Thus, the justification for using copyrighted works in training AI models is negligible, and the "purpose and character" less likely to be considered transformative under the first fair use factor.

8.2. How should the analysis apply to entities that collect and distribute copyrighted material for training but may not themselves engage in the training?

The distribution of copyrighted material to third parties for training is infringing and not protected by fair use. None of the four factors would weigh in favor of fair use: there is no transformative purpose with the wholesale, unchanged reproduction and distribution of complete works; the works are at the core of copyright protection; the complete works are being reproduced; and the market for the works is undoubtedly harmed by the unfettered distribution of free, complete copies of works.

It is axiomatic that an infringer cannot justify its infringement by arguing that an end user might ultimately make fair use of the infringing copies. Lower courts have been near unanimous in rejecting the fair use defense based on the purpose of a third party.[50]

8.3. The use of copyrighted materials in a training dataset or to train generative AI models may be done for noncommercial or research purposes. How should the fair use analysis apply if AI models or datasets are later adapted for use of a commercial nature? Does it make a

[50] De Fontbrune v. Wofsy, 39 F.4th 1214, 1224 (9th Cir. 2022) ("The end-user's utilization of the product is largely irrelevant"); Zomba Enters. v. Panorama Records, Inc., 491 F.3d 574, 582 (6th Cir. 2007) ("the end-user's utilization of the product is largely irrelevant; instead, the focus is on whether alleged infringer's use is transformative and/or commercial."); Los Angeles News Serv. v. Reuters Television Intl Ltd., 149 F.3d 987, 994 (9th Cir. 1998) (rejecting argument that use of works for fair use immunizes from infringement the transmission of the works to the user for that purpose and holding "the question of whether defendants copying and transmission of the works constitutes fair use is distinct from whether their subscribers broadcasts of the works are fair use."); Infinity Broad. Corp. v. Kirkwood, 150 F.3d 104, 108 (2d Cir. 1998) ("it is [defendant's] own retransmission of the broadcasts, not the acts of his end-users, that is at issue here and all [defendant] does is sell access to unaltered radio broadcasts"); Princeton Univ. Press v. Mich. Doc. Servs., 99 F.3d 1381, 1389 (6th Cir. 1996) (en banc); Los Angeles News Serv. v. Tullo, 973 F.2d 791, 797 (9th Cir. 1992) ("the ultimate use to which the customer puts the tape is irrelevant"); Fox Broad. Co. v. Dish Network, 905 F. Supp. 2d 1088, 1106 (C.D. Cal. 2012), affd, 723 F.3d 1067 (9th Cir. 2013) ("[t]he fact that consumers ultimately use AutoHop for private home use, a fair use does not render [Dishs] intermediate copies themselves a fair use as well"); Blackwell Publg, Inc. v. Excel Research Grp., LLC, 661 F. Supp. 2d 786, 793 (E.D. Mich. 2009); UMG Recordings, Inc. v. MP3.com, Inc., 92 F. Supp. 2d 349 (S.D.N.Y. 2000); Basic Books Inc. v. Kinko's Graphics Corp., 758 F. Supp. 1522, 1531-32 (SDNY 1991).

difference if funding for these noncommercial or research uses is provided by for-profit developers of AI systems?

As a general matter, the fact that training is undertaken by a noncommercial entity or for research purposes should accord little, if any, weight to a fair use analysis.[51] Such uses still provide benefits to the users and encroach upon copyright owners' interests in much the same way as commercial uses.[52] Once trained, an AI model can be deployed by multiple third parties, including commercial entities, so the noncommercial nature of the entity developing the model does not mitigate the likely harm to copyright owners that will result. Finally, OpenAI provides precedence for a nonprofit research organization engaged in developing Gen AI models itself transitioning to a for-profit company[53]—one currently valued at $80 billion.[54]

The drafters of the Copyright Act and this Office recognized that for decades, the line between nonprofit and commercial activities has increasingly blurred.[55] In the context of AI, the line between noncommercial and for-profit organizations is even more difficult to draw. The prevalence of "data laundering"—the practice of a commercial entity funding a non-profit or research institution's creation of training dataset, which it then uses for its own activities—compounds this issue. For example, Stability AI funded the development of generative AI model Stable Diffusion by a research group at the Ludwig Maximilian University of Munich, and likewise funded the creation of the training datasets by LAION, a non-profit organization registered in Germany.[56]

---

[51] Harper & Row, Publrs., 471 U.S. at 562 ("The crux of the profit/nonprofit distinction is not whether the sole motive of the use is monetary gain but whether the user stands to profit from exploitation of the copyrighted material without paying the customary price"); Weissmann v. Freeman, 868 F.2d 1313, 1324 (2d Cir. 1989) (non-profit use is not dispositive).

[52] Soc'y of Holy Transfiguration Monastery, Inc. v. Gregory, 689 F.3d 29, 61 (1st Cir. 2012); Worldwide Church of God v. Phila. Church of God, Inc., 227 F.3d 1110, 1118 (9th Cir. 2000); Weissmann v. Freeman, 868 F.2d 1313, 1324 (2d Cir. 1989).

[53] Chloe Xiang, *OpenAI Is Now Everything It Promised Not to Be: Corporate, Closed-Source, and For-Profit*, MOTHERBOARD (Feb. 28, 2023, 11:35 am), https://www.vice.com/en/article/5d3naz/openai-is-now-everything-it-promised-not-to-be-corporate-closed-source-and-for-profit.

[54] Cade Metz, *OpenAI in Talks for Deal That Would Value Company at $80 Billion*, N.Y. TIMES (Oct. 20, 2023), https://www.nytimes.com/2023/10/20/technology/openai-artifical-intelligence-value.html.

[55] H.R. 94-1476 (1976) ("[t]he line between commercial and 'nonprofit' organizations is increasingly difficult to draw"); Supplementary Report of the Register of Copyrights on the General Revision of the U.S. Copyright Laws: 1965 Revision Bill" (House Committee Print, 1965) at 21 (blanket exemption for nonprofit performances "would involve serious dangers to the author's rights" in light of technological advancements that blurred the line between nonprofit and commercial activities).

[56] Baio, *supra* note 12.

8.4. What quantity of training materials do developers of generative AI models use for training? Does the volume of material used to train an AI model affect the fair use analysis? If so, how?

The volume of material used to train an AI model certainly affects the fair use analysis. The more systematic the unauthorized copying of protected works by a Gen AI developer, the greater the harm to copyright owners and the Copyright Act as a whole. A higher volume by definition results in greater market harm, not to mention increased security risks, and devaluation of the copyright system. Fair use should not provide a "volume discount" for unauthorized copying, and AI developers should not be given a perverse incentive to expropriate greater quantities of copyrighted works. The Copyright Office can consider whether legislation that clarifies this point is appropriate.

8.5. Under the fourth factor of the fair use analysis, how should the effect on the potential market for or value of a copyrighted work used to train an AI model be measured? Should the inquiry be whether the outputs of the AI system incorporating the model compete with a particular copyrighted work, the body of works of the same author, or the market for that general class of works?

There are at least three potential market harms that would be relevant under the fourth factor of fair use. First, the unauthorized reproduction of copyrighted works to train AI models harms the market for licensing copyrighted works for training AI models. As noted above, such licensing is already occurring in the market. In the parlance of fair use, such licensing is reasonable and represents a market that copyright owners are likely to continue to develop.[57] The unauthorized reproduction of copyrighted works to train AI models supplants and usurps a market that properly belongs to copyright owners, weighing against fair use.[58]

Second, to the extent a Gen AI model can reproduce copies of a work or derivatives of a work, it can cause harm to those markets—i.e., the market for digital

[57] Accord Am. Geophysical Union v. Texaco Inc., 60 F.3d 913, 930 (2d Cir. 1994) (courts should consider "traditional, reasonable, or likely to be developed markets" when determining a use's effect on the potential licensing revenues of a copyright owner); Seltzer v. Green Day, Inc., 725 F.3d 1170, 1179 (9th Cir. 2013) (same). Furthermore, the fact that some copyright owners have not yet licensed their works for training AI models, or have chosen not to license in this market, does not weigh in favor of fair use under the fourth factor. Balsley v. LFP, Inc., 691 F.3d 747, 761 (6th Cir. 2012); Castle Rock Entm't, Inc. v. Carol Publ'g Grp., Inc., 150 F.3d 132, 145-46 (2d Cir. 1998); Pacific and Southern Co., Inc. v. Duncan, 744 F.2d 1490, 1496-97 (11th Cir. 1984).

[58] Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169, 180 (2d Cir. 2018) (citing Infinity Broad. Corp. v. Kirkwood, 150 F.3d 104, 110 (2d Cir. 1998)).

copies or online displays of a work and the market(s) for derivative works. As noted in our response to Question 7.4, several lawsuits against AI developers include allegations that the models are able to reproduce verbatim portions of works—and even entire works, such as song lyrics.[59] Others allege the models readily generate summaries of works, which in some instances may constitute unauthorized derivative works.[60] Such copies and derivative works are paradigmatic superseding uses under the fourth factor.[61]

Third, by generating outputs of the same type of work as the inputs, generative AI causes harm to the market for a general class of works that is cognizable under fair use. Section 107 directs a court to consider the "value of a copyrighted work." If a copyrighted work is reproduced to train a Gen AI model that will generate works that compete in the market with the copyrighted work, it will clearly reduce the value of that copyrighted work.

9. 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)?

The default rule under copyright law is that copyright owners have the exclusive right to authorize any use of their works.[62] Whereas the use of copyrighted works for training AI models implicates a copyright owner's exclusive rights and would generally not be permitted by fair use, that use could only occur if copyright owners opt in; an opt out regime (such as that imposed by Article 4 of the EU DSM Directive) would be inconsistent with exclusive rights. In short, an opt-out system would turn copyright on its head.

9.1. Should consent of the copyright owner be required for all uses of copyrighted works to train AI models or only commercial uses?

---

[59] David, *supra* note 33; Complaint, J.L. v. Alphabet Inc., ¶¶ 15, 111, No. 3:23-cv-3440, N.D. Cali., July 11, 2023.

[60] Winston Cho, *Top Authors Join Lawsuit Against OpenAI Over "Mass-Scale Copyright Infringement" of Novels,* THE HOLLYWOOD REPORTER (Sept. 20, 2023), https://www.hollywoodreporter.com/business/business-news/top-authors-join-lawsuit-against-openai-over-mass-scale-copyright-infringement-of-novels-1235595123/; Winston Cho, *Authors Sue Meta, OpenAI in Lawsuits Alleging Infringement of Hundreds of Thousands of Novels*, THE HOLLYWOOD REPORTER (Sept. 12, 2023), https://www.hollywoodreporter.com/business/business-news/authors-sue-meta-openai-class-action-1235588711/.

[61] 4 Nimmer on Copyright § 13F.08 (2023).

[62] 17 U.S.C. § 106.

Yes, authorization is required for all uses of copyrighted works to train AI models, not just commercial ones. Copyright grants the copyright owner exclusive rights over their works, and this includes the right to control how and by whom their work is used.

As regards the Office's footnote 47 to Question 9.1, which references Articles 3 and 4 of the EU DSM Directive, it should be noted that the Directive preceded the rapid growth and significant attention to Gen AI systems over the last year. The Directive does not refer to AI, much less Gen AI, nor does it suggest a scientific research context would be permissible where text, sounds, or images would be used to generate new text, sounds, or images.

9.2. If an "opt out" approach were adopted, how would that process work for a copyright owner who objected to the use of their works for training? Are there technical tools that might facilitate this process, such as a technical flag or metadata indicating that an automated service should not collect and store a work for AI training uses?

An "opt out" approach turns the copyright framework on its head. Copyright secures exclusive rights to the copyright owner and determinations regarding how and by whom the rights holder's works are used lie solely with the copyright holder (see also Response to Question 9).

In addition, there are practical issues with an opt out approach. As Gen AI developers have routinely sourced books, scientific articles and other copyrighted materials from piracy sites, relying on technical tools to prevent scraping and collection of copyrighted works as training material under an opt out approach would be ineffective. Piracy site operators reproduce and distribute copyrighted works without the consent of authors and publishers, and publishers do not have an opportunity to embed the technological tools that would enable an opt out.

Furthermore, there is as yet no technology that can detect or flag when an AI-bot disregards or ignores opt out instructions or "do not train/scrape" tags. Thus, absent a transparency requirement that identifies and discloses the copyrighted works used for training purposes, rights holders would be unable to verify compliance with the opt-out instructions.

There is also a risk that adequate tools for an "opt out" approach would not be developed in a timely fashion, would not be accessible to large and small copyright holders alike, would not be adequately implemented, or would impair online access/discoverability for purposes unrelated to AI training. The inability to develop

"standard technical measures" under 17 U.S.C. § 512(i)(1)(B) in the 25 years after the DMCA went into effect serves as a cautionary tale.[63]

As regards DSM Article 4 TDM reservation, there have been efforts to develop a non-normative TDM Reservation Protocol for web content under the auspices of the World Wide Web Consortium (W3C) to "allow a rightsholder to declare his choice regarding text & data mining of Web resources he controls, thereby allowing recipients of that declaration to adjust their scraping behavior, or to reach a separate agreement with the rightsholder that satisfies all parties."[64] A critical mass of early adopters has yet to be reached.

In addition, because the W3C TDM Reservation Protocol is limited to web content and applicable at the server level, professional and scholarly publishers are also exploring the development of a TDM reservation protocol that can be embedded at the item level, (i.e., into the metadata information of the PDF or ePub version of an article, for example). This effort is still in development and whether effective opt-out mechanisms can be developed and effectively implemented remains undetermined.[65]

9.3. What legal, technical, or practical obstacles are there to establishing or using such a process? Given the volume of works used in training, is it feasible to get consent in advance from copyright owners?

As noted in the preceding section, the progress for developing the technical tools to implement an opt-out mechanism has been slow, and the tools may prove to have limited scope/effectiveness. As a practical matter, the absence of technologies that would alert a rights holder when its opt-out instructions or "do not train/do not scrape" tags are disregarded, coupled with the lack of transparency as to which copyrighted works have been acquired by a Gen AI systems developer or dataset creator without a license, would result in rights holders having little recourse when their exclusive rights have been violated.

---

[63] USCO Letter to Congress, Standard Technical Measures and Section 512 (Dec. 20, 2022) ("not a single technology has been designated a 'standard technical measure' under section 512(i)") (quoting U.S. COPYRIGHT OFFICE, SECTION 512 OF TITLE 17, at 67 (2020)).

[64] See, W3C COMMUNITY GROUP, TDM RESERVATION PROTOCOL (TDMREP): FINAL COMMUNITY GROUP REPORT (Laurent Le Meur ed., 2022). TDM Reservation Protocol (TDMRep) (w3.org).

[65] A group of U.S. creator groups noted their concerns with DSM Articles 3 and 4, primarily that the articles are not in compliance with the Berne Convention. See, Appeal from Creators of Copyrighted Works, Appeal for Action on Violations of the Berne Convention by the Application to Copying of Creative Works for AI Development of the TDM exception (July 2023), https://nwu.org/wp-content/uploads/2023/07/creators-coalition-AI-exceptions.pdf.

At the same time, it is feasible to obtain in advance rights holder consent to use their copyrighted works for training, and this is what the law requires. Licensing is occurring across all publishing sectors. The claim that the volume of works used for training makes it burdensome for a Gen AI systems developer to seek permission is not an excuse for infringing on the copyrights and livelihoods of the thousands of authors, publishers, and other artists. Copyrighted works are as integral to Gen AI technologies as they are to services like Netflix and Spotify, both of which license the works made available through their platforms. Developers of Gen AI are capable of doing the same and rights holders are willing to work with them to effect such licensing. Given the valuations of AI developers — for example, it has been reported that OpenAI could see its valuation at an estimated $80 billion[66] — licensing deals with the publishing industry would represent a fraction of the cost. To give a sense of scale, AAP has approximately 130 members across trade, education, and professional and scholarly publishing. For training purposes in any given sector, negotiating licenses with some percentage of the industry is quite feasible. Voluntary collective licensing approaches may also emerge. (See Response to Question 10.2.)

Publishers invest significant resources in bringing authors' works to market — taking risks on new works from new and existing authors. These investments in professional writing, peer-reviewed reporting of scientific research, and separating fact from fiction will only be more important as Gen AI technologies indiscriminately trained on unvetted material accelerate the spread of misinformation and bias. Requiring Gen AI developers to obtain licenses is essential to ensuring that publishers continue to have proper incentives to invest in new works that benefit society as a whole.

9.4. If an objection is not honored, what remedies should be available? Are existing remedies for infringement appropriate or should there be a separate cause of action?

Our current view is that for copyright infringement actions, the existing remedies under Title 17, chapter 5, of the Copyright Act, are available and appropriate. The rights holder may seek injunctive relief to enjoin the AI systems developer from including or using the copyrighted works collected through the un-permissioned scraping and from further acts of unauthorized uses, as well as compensation for damages sustained. The failure to adhere to opt-out instructions could be used as evidence of willfulness for purposes of statutory damages calculations pursuant to 17 U.S.C. § 504(c).

Of greater concern is the rightsholder's ability to prove a copyright infringement claim in an opt-out system. We note that the lack of technical measures that can detect

[66] See, Metz, *supra* 54.; See also, *OpenAI in talks to sell shares at $86 billion valuation – Bloomberg News*, REUTERS (Oct. 18, 2023), https://www.reuters.com/technology/openai-talks-sell-shares-86-billion-valuation-bloomberg-news-2023-10-18/.

when an AI bot disregards or ignores opt-out or do-not-train/do-not-scrape instructions (much less send alerts to the copyright owner for such incidences) may make the ability to collect evidence of copyright infringement highly uncertain or unduly burdensome.

9.5. In cases where the human creator does not own the copyright—for example, because they have assigned it or because the work was made for hire—should they have a right to object to an AI model being trained on their work? If so, how would such a system work?

Only the copyright owner exercises control over whether and how the copyrighted work should be reproduced, distributed, or otherwise used, including for training AI models. Where the human creator has relinquished the copyright in their creation or never had copyright in cases where they contracted away, sold, assigned or otherwise disposed of their copyright, they would not have the right to object to an AI model being trained on their work if the person or entity to whom the rights were assigned has authorized such use.

10. If copyright owners' consent is required to train generative AI models, how can or should licenses be obtained?

Licenses can and should be obtained through direct negotiations with the copyright owners. However, there may be instances where voluntary collective licensing arrangements, through the appropriate intermediaries, can be a viable option.

10.1. Is direct voluntary licensing feasible in some or all creative sectors?

Yes, direct voluntary licensing is feasible and is certainly the case for the publishing industry. Professional and scholarly publishers already employ licensing arrangements to facilitate access to their databases, whether for non-commercial research purposes or for commercial use. Other sectors of the publishing industry are exploring how they may facilitate access to their copyrighted works, consistent with the rights their authors have assigned to them.

10.2. Is a voluntary collective licensing scheme a feasible or desirable approach? Are there existing collective management organizations that are well-suited to provide those licenses, and are there legal or other impediments that would prevent those organizations from performing this role? Should Congress consider statutory or other changes, such as an antitrust exception, to facilitate negotiation of collective licenses?

Voluntary collective licensing is consistent with the exclusive rights of copyright owners and may prove to be a feasible approach alongside direct licensing. It is possible

to engage in voluntary collective licensing without legislative support. We believe it is currently premature to consider any statutory or other changes to facilitate negotiation of collective licenses.

10.3. Should Congress consider establishing a compulsory licensing regime? If so, what should such a regime look like? What activities should the license cover, what works would be subject to the license, and would copyright owners have the ability to opt out? How should royalty rates and terms be set, allocated, reported and distributed?

No, Congress should not consider a compulsory licensing regime. The Copyright Office has correctly observed, "Compulsory licensing has been, and should be, regarded as an extreme last resort in copyright law,"[67] and it has held this view firmly for decades.[68] Congress shares this view, recognizing that compulsory licenses fly in the face of the exclusive rights of authors, and it has enacted compulsory licenses reluctantly, sparingly, and narrowly—and only after clear and demonstrated need.[69]

---

[67] Second Supplementary Report of Register of Copyrights (Oct. 1975).

[68] U.S. Copyright Office, Satellite Television Extension and Localism Act, pg. 1 (August 29, 2011) ("[B]y their nature, statutory licenses are exceptions under copyright law and a limitation on the fundamental principle that authors should enjoy exclusive rights to their creative works, including for the purpose of controlling the terms of public dissemination. Historically, the Copyright Office has supported statutory licenses only when warranted by special circumstances and only for as long as necessary to achieve a specific goal. And Congress has enacted such provisions sparingly."); *Competition and Commerce in Digital Books: The Proposed Google Book Settlement: Hearing Before the Comm. On the Judiciary*, 111th Cong. (2009) (statement of Marybeth Peters, Register of Copyrights), *available at* https://www.copyright.gov/docs/regstat091009.html ("Congress generally adopts compulsory licenses only reluctantly in the face of a failure of the marketplace, after open and public deliberations that involve all affected stakeholders, and after ensuring that they are appropriately tailored"); Copyright/Cable Television: Hearings on H.R. 1805, H.R. 2007, H.R. 2108, H.R. 3528, H.R. 3530, H.R. 3560, H.R. 3940, H.R. 5870, and H.R. 5949 Before the Subcomm. On Courts, Civil Liberties, and the Admin. of Justice of the Comm. on the Judiciary, 97th Cong. 959-60 (1981) (statement of David Ladd, Register of Copyrights) ("[A] compulsory license mechanism is in derogation of the rights of authors and copyright owners. It should be utilized only if compelling reasons support its existence").

[69] See, e.g., S. Rep. No. 106-42, at 10 (1999) ("[T]he Committee is aware that in creating compulsory licenses, it is acting in derogation of the exclusive property rights granted by the Copyright Act to copyright holders, and that it therefore needs to act as narrowly as possible to minimize the effects of the Government's intrusion on the broader market in which the affected property rights and industries operate."); H.R. Rep. No. 94-1476, 94th Cong., 2d Sess. (1976) at 119 ("The Committee * * * concluded that the performance of nondramatic literary works should not be subject to Commission determination. It was particularly concerned that a compulsory license for literary works would result in loss of control by authors over the use of their work in violation of the basic principles of artistic and creative freedom. It is recognized that copyright not only provides compensation to authors, but also protection as to how and where their works are used.").

There is no such need here. Direct licensing of copyrighted works for training purposes is successfully occurring, AI developers have a wealth of open and authorized sources of training data (such as public domain and open access materials), and AI developers can generate their own training data in many cases.

10.4. Is an extended collective licensing scheme a feasible or desirable approach?

Because extended collective licensing also acts in derogation of the exclusive rights of copyright owners, it raises many of the same concerns as compulsory licensing. As with compulsory licensing, extended collective licensing is not currently needed or justified.

10.5. Should licensing regimes vary based on the type of work at issue?

Licensing regimes may necessarily vary depending on the type of work at issue. Some works may be more amenable to voluntary collective licensing regimes than others.

11. What legal, technical or practical issues might there be with respect to obtaining appropriate licenses for training? Who, if anyone, should be responsible for securing them (for example when the curator of a training dataset, the developer who trains an AI model, and the company employing that model in an AI system are different entities and may have different commercial or noncommercial roles)?

As noted in earlier responses, copyrighted works are as integral to Gen AI technologies as they are to services like Netflix and Spotify. Like Netflix and Spotify, AI systems developers should invest in licensing departments to secure the appropriate rights from the copyright owners of the works they wish to use for training purposes. It is feasible for AI developers to negotiate licenses for training. (See Response to Question 9.3.)

Each of the above actors would be responsible for securing rights, whether directly (e.g., from the publisher or publisher's agent) or indirectly (e.g., the publisher's contract with the AI developer could address downstream users).

The requirement for securing rights should also apply to dataset creators or curators. However, we note that the role of the independent curator may evolve in the future. In many cases, so-called "curators" only assemble unlicensed copyrighted works (in some circumstances incorrectly claiming such a right under DSM Directive Article 3 when the actual downstream uses are commercial; and others, without a justification at

all). In many cases this conduct is essentially piracy. It is hoped this function will become less prevalent when licensing becomes the norm.

12. Is it possible or feasible to identify the degree to which a particular work contributes to a particular output from a generative AI system? Please explain.

As regards literary works, there are circumstances where comparing a particular output to a particular input can indicate infringement of the input, including output in the form of verbatim text, summaries, or a derivative work, amongst others.

However, we consider this issue to be an unhelpful distraction for the following reasons: (i) licensing is required for training regardless of whether the output infringes a particular work and (ii) the test for whether an output is infringing is substantial similarity and access, not degree of contribution.

13. What would be the economic impacts of a licensing requirement on the development and adoption of generative AI systems?

A licensing requirement — i.e., the default rule under the copyright framework of exclusive rights — would have a positive economic impact on the development and adoption of Gen AI systems, as well as the continued creation and distribution of high-quality works by the creative sector. The ability of copyright owners to decide when, where, to whom, and for how much they will authorize the use of their works is fundamental to their ability to achieve copyright's purpose.

Licensing fees are an important source of income for U.S. creators and rightsholders and support the continued investment in new human-created works. The importance of sustaining the U.S. publishing industry cannot be understated. AAP members publish high-quality literary works, including works that present novel ideas and new facts unearthed by authors; hold governments, businesses, and citizens accountable; contribute to a vibrant culture; educate, and inspire Americans of all ages; and report on scientific progress. Trustworthy Gen AI systems require high-quality new publications to remain state-of-the-art, and a flourishing publishing industry is best positioned to increase the value of Gen AI systems. A flourishing publishing industry will also help protect against some of the potential ills of Gen AI systems, including misinformation and bias.

In addition, AI systems developed or trained on works derived or created from authorized sources are more likely to yield reliable outputs than works obtained from illegal or pirate sources. It is essential to trustworthy and reliable AI that developers utilize high quality, curated content to create training corpora for their models. For example, in the case of AI training based on professional and scholarly communications,

we note the importance of AI developers using the Version of Record (VoR), under appropriate licenses. The VoR is the final, publisher-maintained article, updated and archived continually in consultation with the author. Accepted manuscripts, pre-prints, or illegally uploaded text versions of the article may be subject to post publication modification or retraction, which if used as training material in their uncorrected state, could create serious and cascading scientific or medical errors in AI generated outputs.

Beyond industry economic impacts, given that AI technologies will be (and are being) integrated into applications that will impact the lives and well-being of individuals, whether financially, physically, mentally, or professionally, it is critically important that licensing requirements be implemented to ensure that high quality, peer reviewed, vetted material is used to create the training corpora of AI systems and to refresh that training corpora going forward.

14. Please describe any other factors you believe are relevant with respect to potential copyright liability for training AI models.

Recordkeeping and disclosure requirements on AI developers concerning the identity of copyrighted works used for training purposes, how they were acquired (i.e., identification of the authorized licensor(s) and/or other source(s)), and how the foundation model was trained, are essential for protecting creators and rightsholders. Deficiencies could be evidence of willfulness for purposes of statutory damages calculations pursuant to 17 U.S.C. § 504(c).

Since licensing should be required for works protected by copyright, the Library of Congress should not grant exemptions to section 1201's prohibition against circumvention of technological measures that control access to copyrighted works for the purpose of training AI models.

Transparency & Recordkeeping

15. In order to allow copyright owners to determine whether their works have been used, should developers of AI models be required to collect, retain, and disclose records regarding the materials used to train their models? Should creators of training datasets have a similar obligation?

Both developers of AI foundation models and creators of publicly accessible training datasets should be required to maintain accurate records as to (a) the copyrighted works used to create the training datasets ingested by AI foundation models; and (b) whether the material is copyright protected, and if so, whether the use of this content is licensed and from whom.

15.1. What level of specificity should be required?

As to a general disclosure obligation, an AI systems developer or the training dataset creator should be obliged to provide an accurate record of the copyright protected and other materials used to create the training datasets, whether the use of the copyrighted works in the dataset is licensed, and if so, the licensor(s) from whom the copyrighted material is sourced, as well as to identify potential risks in the AI system given the nature of the data set on which it was trained, and whether and how the risks have been mitigated.

15.2. To whom should disclosures be made?

Disclosures regarding the nature of the training data set, how sourced (i.e., identification of the authorized licensor(s) and/or other source(s)) and whether the copyrighted materials used are licensed should, at a minimum, be disclosed to the owners of the copyrighted material used to create datasets (if not already notified) and to the users of the AI system, i.e., the parties that license the AI system, with the intention of integrating the same into their business processes. We note that disclosure requirements for copyright owners are consistent with broader disclosure requirements necessary to building the public's trust in these AI systems.[70]

15.3. What obligations, if any, should be placed on developers of AI systems that incorporate models from third parties?

Developers of AI systems that incorporate models from third parties should be required to disclose information sufficient to determine the use of copyrighted works in training materials, such as the identity of the third-party provider or developer of the foundation model and iteration of the model.

15.4. What would be the cost or other impact of such a recordkeeping system for developers of AI models or systems, creators, consumers, or other relevant parties?

Given that appropriately crafted recordkeeping systems would largely align with best practices for responsible and ethical dataset development,[71] the cost and impact on developers of AI models should already be factored in. The benefits of such requirements for creators and consumers would exceed any costs.

[70] See, e.g., Timnit Gebru et al., *Datasheets for Datasets* (2021), https://arxiv.org/abs/1803.09010].
[71] See, e.g., Timnit Gebru et al., *Datasheets for Datasets* (2021), https://arxiv.org/abs/1803.09010.

16. What obligations, if any, should there be to notify copyright owners that their works have been used to train an AI model?

AI systems developers should, at the outset, have sought licenses from the copyright owners of the works they sought to use to train their AI systems. That licensing negotiation is the notice that rights holders should have with respect to the potential use of their works. However, at a minimum, developers should have an obligation to provide a readily accessible and easily searchable tool for copyright owners to be able to determine if their works have been used to train an AI model. Such a tool should include all metadata associated with the content and be able to support emerging standard identifiers such as the International Standard Content Code, which are designed to allow content owners to identify and match content in decentralized environments.[72]

17. Outside of copyright law, are there existing U.S. laws that could require developers of AI models or systems to retain or disclose records about the materials they used for training?

Data privacy laws, in particular, if the training materials included personal data or information whether related to health, financial, education, or employment records.

Generative AI Outputs | Copyrightability

18. Under copyright law, are there circumstances when a human using a generative AI system should be considered the "author" of material produced by the system? If so, what factors are relevant to that determination? For example, is selecting what material an AI model is trained on and/or providing an iterative series of text commands or prompts sufficient to claim authorship of the resulting output?

A human using a Gen AI system to create material may be considered the author under certain circumstances. Whether there is sufficient human authorship is a fact-specific inquiry that the Office and courts will be required to undertake.

The Office stated as much when it noted that "a work containing AI-generated material will also contain sufficient human authorship to support a copyright claim," when a human selects or arranges "AI-generated material in a sufficiently creative way that "the resulting work as a whole constitutes an original work of authorship."[73] Where

[72] ISCC – Content Codes: A Proposal for a Modern and Open Content-Based Identifier, https://iscc.codes/ (Last Updated: Apr. 25, 2022, 11:36 AM).

[73] 37 CFR § 202 (2023), Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence.

human authorship does not exist for material produced by Gen AI systems, there should be no protection under the Copyright Act.

19. Are any revisions to the Copyright Act necessary to clarify the human authorship requirement or to provide additional standards to determine when content including AI-generated material is subject to copyright protection?

No, revisions to the Copyright Act are unnecessary to clarify the human authorship requirement. In *Thaler v. Perlmutter*,[74] the court made clear that "United States copyright law protects only works of human creation," and that "(h)uman authorship is a bedrock requirement. In upholding the Copyright Office's interpretation that human authorship is required for copyright protection, the court found that "a work generated entirely by an artificial system absent human involvement" is ineligible for copyright protection.

The court, however, noted that there remain unanswered questions with respect to whether AI-generated output can be or should be subject to copyright protection, stating that "(t)he increased attenuation of human creativity from the actual generation of the final work will prompt challenging questions regarding how much human input is necessary to qualify the user of an AI system as an "author" of a generated work, the scope of the protection obtained over the resultant image, how to assess the originality of AI-generated works where the systems may have been trained on unknown pre-existing works, how copyright might best be used to incentivize creative works involving AI, and more."[75]

20. Is legal protection for AI-generated material desirable as a policy matter? Is legal protection for AI-generated material necessary to encourage development of generative AI technologies and systems? Does existing copyright protection for computer code that operates a generative AI system provide sufficient incentives?

Legal protection for AI-generated material would not serve the purpose of copyright law, and the Copyright Act should not be compromised. Copyright affords authors and publishers legal protections for the intellectual property—the works—they produce, which encourages the creation, commercialization, and dissemination of new works to the public, thereby promoting the "progress of science and the useful arts." Legal protection assures the artist, the author, the rights holder that their investment, whether intellectual or otherwise, in the creation and dissemination of the creative work will see a return that allows them to continue in their profession. Gen AI systems require

[74] Thaler v. Perlmutter, No. 22-1564 (BAH), 2023 U.S. Dist. LEXIS 145823 (2023).

[75] Id.

no such incentive as the output of these systems, on their own, lack the type of human creative expression on which copyright is premised. As such, material or output solely generated by a Gen AI system, completely absent human involvement accomplishes no policy imperative to grant copyright protection to the AI algorithm.

AI generated output enjoys no copyright protection, and yet, over the last several years, developers of Gen AI systems have invested and continue to invest in further developing this technology. Their investment has not been and is not tied to the grant of legal protections such as copyright to the output of Gen AI systems. So, in this case, the market itself provides sufficient incentives.

20.1. If you believe protection is desirable, should it be a form of copyright or a separate sui generis right? If the latter, in what respects should protection for AI-generated material differ from copyright?

Not answered.

21. Does the Copyright Clause in the U.S. Constitution permit copyright protection for AI-generated material? Would such protection "promote the progress of science and useful arts"? If so, how?

Regardless of where the Office stands on copyright protection for AI-generated material, we caution the Copyright Office against finding the text of the Copyright Clause dictates such a policy. As the Supreme Court has noted, "it is generally for Congress, not the courts, to decide how best to pursue the Copyright Clause's objectives."[76]

Infringement

22. Can AI-generated outputs implicate the exclusive rights of preexisting copyrighted works, such as the right of reproduction or the derivative work right? If so, in what circumstances?

Yes, AI-generated outputs may implicate several exclusive rights granted to copyright owners. To show that an AI-generated output infringed upon a prior work, the copyright owner would have to show that the Gen AI system (1) had access to, and (2) generated output that is "substantially similar" to the original works. As noted previously, research into the underlying LLMs that power Gen AI systems have shown that the training datasets too often include copyrighted books and images scraped from the Internet, including from piracy sites, establishing access. From such access, the Gen AI system is likely to have "learned" from the work's expression—not just the ideas and facts

[76] Eldred v. Ashcroft, 537 US 186, 212 (2003).

included in the expression—thus rendering the Gen AI system capable of producing output that reproduces the protectable elements of the original work.

23. Is the substantial similarity test adequate to address claims of infringement based on outputs from a generative AI system, or is some other standard appropriate or necessary?

The substantial similarity test may be adequate to address whether the output of a Gen AI system has infringed an existing copyrighted work. However, the current lack of transparency with respect to which copyrighted works are included in training datasets makes it difficult for copyright owners to prove that the Gen AI system developer (or the user of the Gen AI system) had access to the work so as to allow the developer/user to actually copy the work.

24. How can copyright owners prove the element of copying (such as by demonstrating access to a copyrighted work) if the developer of the AI model does not maintain or make available records of what training material it used? Are existing civil discovery rules sufficient to address this situation?

There are tools that may be useful for interrogating an AI model as to whether images, books, magazine articles, etc. have been included in AI training datasets.[77] However, such tools may not accurately account for all copyrighted works that may have been included in an AI training dataset. Where this is the case, a copyright owner may circumstantially prove that the copyrighted work was included in the AI model's training dataset.[78] For instance, a copyright owner may look to research showing that certain LLMs underlying particular AI models used works scraped from piracy databases on which are likely hosted millions of infringing copies of copyrighted works.[79]

Existing civil discovery rules are likely insufficient to address the situation. They must be supplemented by a transparency requirement whereby Gen AI systems developers are required to (1) keep records of the copyrighted works included in the training datasets, the entities that created the training datasets (if not the Gen AI systems developer), and (2) disclose this information to the copyright owners, and potentially the end users of the Gen AI system.

---

[77] *Have I Been Trained?*, HIBT, https://haveibeentrained.com/.

[78] See, CHRISTOPER T. ZIRPOLI, CONG. RSCH. SERV., LSB10922, GENERATIVE ARTIFICIAL INTELLIGENCE AND COPYRIGHT LAW (Sept. 29, 2023).

[79] Schaul, *supra* note 17; Ian Bogost, *My Books Were Used to Train Meta's Generative AI. Good.*, THE ATLANTIC (Sept. 27, 2023), My Books Were Used to Train Meta's Generative AI. Good. - The Atlantic.

25. If AI-generated material is found to infringe a copyrighted work, who should be directly or secondarily liable—the developer of a generative AI model, the developer of the system incorporating that model, end users of the system, or other parties?

Exclusive rights are illusory without appropriate liability and enforcement regimes. As copyright liability is fact dependent, and joint and severable, the parties that may be held directly or secondarily liable will depend on the circumstances of the infringing use. There may be instances where the end user of the system — the party that crafted and refined the descriptive text prompts that directed the AI system to generate the output may be held directly liable for infringement. There may be cases where both the developer of the Gen AI model and the developer of the system incorporating that model into another product or service, depending on the facts, can also be held either directly liable or secondarily liable under the three doctrines of secondary liability: vicarious, contributory, and inducement. For example, as discussed in our responses to questions 6.1 and 8 above, the fact that a developer may have used copyrighted works acquired from illicit or pirate sources should strengthen the grounds for liability based on outputs. Secondary liability doctrines in particular will become increasingly important in tracing the acts that prompt infringement with Gen AI systems to hold developers appropriately accountable for the harms that result.

25.1. Do "open-source" AI models raise unique considerations with respect to infringement based on their outputs?

Not answered.

26. If a generative AI system is trained on copyrighted works containing copyright management information, how does 17 U.S.C. 1202(b) apply to the treatment of that information in outputs of the system?

Section 1202(b) appropriately creates liability where a Gen AI system developer, in creating or using a training dataset that includes copyrighted works, intentionally removes or alters the copyright management information embedded in the work and where a Gen AI systems developer knowingly distributes the model's output with altered copyright management information (or with the copyright management information removed). But Section 1202(b) as written may not be sufficient in the context of Gen AI systems given the knowledge and intent requirements needed to establish liability. Copyright management information (CMI) is necessary to rights holders' ability to track and monetize their works. The removal of CMI by an AI systems developer either to conceal or enable infringement makes it more difficult for a rights holder to identify when their works have been ingested for training and thus deprives them of their ability to control how their works are used and whether they are compensated for such use,

regardless of whether there is additional evidence of the AI developer's state of mind or intent.

27. Please describe any other issues that you believe policymakers should consider with respect to potential copyright liability based on AI-generated output.

Not answered.

Labeling or Identification

28. Should the law require AI-generated material to be labeled or otherwise publicly identified as being generated by AI? If so, in what context should the requirement apply and how should it work?

It may be useful for AI-generated material to be disclosed or identified as such for the public's benefit. However, questions with respect to labeling and identification are tied to policy issues not necessarily addressable by copyright law.

28.1. Who should be responsible for identifying a work as AI-generated?

Not answered.

28.2. Are there technical or practical barriers to labeling or identification requirements?

Not answered.

28.3. If a notification or labeling requirement is adopted, what should be the consequences of the failure to label a particular work or the removal of a label?

Not answered.

29. What tools exist or are in development to identify AI-generated material, including by standard-setting bodies? How accurate are these tools? What are their limitations?

A number of companies are developing tools intended to detect and identify text or images partially or entirely generated by Gen AI systems, although many are still

experimental and not yet widely deployed.[80] Companies are also developing "watermarking" systems,[81] where an invisible watermark is affixed to Gen AI generated text or images to allow the output to be identified as such. While progress is being made, these tools are not yet without weaknesses.

Additional Questions About Issues Related to Copyright

30. What legal rights, if any, currently apply to AI-generated material that features the name or likeness, including vocal likeness, of a particular person?

Not answered.

31. Should Congress establish a new federal right, similar to state law rights of publicity, that would apply to AI-generated material? If so, should it preempt state laws or set a ceiling or floor for state law protections? What should be the contours of such a right?

Not answered.

32. Are there or should there be protections against an AI system generating outputs that imitate the artistic style of a human creator (such as an AI system producing visual works "in the style of" a specific artist)? Who should be eligible for such protection? What form should it take?

We oppose protections related to artistic style at this time. The notion of "style" is somewhat vague, and creating a standard around this rather nebulous concept may bleed into non-copyrightable uses. Copyright law only protects the specific expression of an idea. While certain style elements may be reminiscent of another author or artist, unless the purportedly infringing subsequent work is also substantially similar to the prior "specific expression" there is no infringement.[82] Style alone does not afford copyright protection to an artist's or author's body of work. It is the specific expression embodied in the work that is protected.

---

[80] See, *The AI Detector: Checks for ChatGPT, GPT4, Bard, Clause & More*, CONTENT AT SCALE, AI Detector Checks ChatGPT & GPT-4 | Paraphrasing Tool (contentatscale.ai).

[81] See Pete Syme, *Google is Launching a Tool that Helps Users to Identify Whether a Picture is AI-Generated*, BUSINESS INSIDER (May 11, 2023), Google's Search Tool Helps Users to Identify AI-Generated Fakes (businessinsider.com); Gerrit de Vynck, *AI Images are Getting Harder to Spot. Google Thinks it has a Solution*, WASH. POST (Aug. 29, 2023), AI images are getting harder to spot. Google thinks it has a solution. - The Washington Post; Svewn Gowal & Pushmeet Kohli, *Identifying AI-generated Images with SynthID*, GOOGLE (Aug. 29, 2023), Identifying AI-generated images with SynthID (deepmind.com).

[82] See Steinberg v. Columbia Pictures Industries, Inc., 663 F. Supp. 706 (S.D.N.Y. 1987) (where the court stated that "style is one ingredient of expression.").

33. With respect to sound recordings, how does section 114(b) of the Copyright Act relate to state law, such as state right of publicity laws? Does this issue require legislative attention in the context of generative AI?

Not answered.

34. Please identify any issues not mentioned above that the Copyright Office should consider in conducting this study.

Not answered.