AUTHORS ALLIANCE

authorsalliance.org
510.480.8302

2705 Webster St. #5805
Berkeley, CA 94705

October 30, 2023

Suzanne Wilson
General Counsel and Associate Register of Copyrights
United States Copyright Office
101 Independence Ave. S.E.
Washington, D.C.
20559-6000

Via Online Submission

RE: Policy Study on Artificial Intelligence, Docket Number 2023-6

Dear Associate Register Wilson:

Authors Alliance has a mission of advancing the interests of authors who write for the public benefit by sharing their work broadly.[1] Our comments on copyright and AI, in summary, are that:

- Generative AI supports creators and is an important tool for promoting the progress of science and culture;
- Legislation to address copyright issues in generative AI would be premature while the technology and its use are still nascent;
- In a vast majority of cases, the use of copyrighted works as training materials is a non-expressive fair use, consistent with the four factors and recent case law;
- In fair use analysis, factor four should consider the effect of the use of a copyrighted work as training data on the market for that specific work only based on statutory language and case law;
- Neither a compulsory licensing nor a collective licensing scheme for generative AI training data are logistically feasible or sound as a policy matter;
- No new copyright legislation is necessary to clarify the human authorship requirement or provide guidance about authorship issues in AI-generated content;
- The substantial similarity test is adequately equipped to handle claims of AI-generated works infringing on existing works.

We care deeply about access to knowledge because it supports free inquiry and learning, and we are enthusiastic about ways that generative AI can meaningfully further those ideals. In addition

[1] For more about Authors Alliance, our mission, and our leadership, see https://www.authorsalliance.org/about/.

to all the mundane but important efficiency gains generative AI can assist with, we have already seen authors incorporate generative AI systems into their creative processes to produce new works. We have also seen researchers incorporate these tools to help them make new discoveries. There are some clear concerns about how generative AI systems can, for example, make it easier to engage in fraud and deception, as well as perpetuating disinformation. However, in our view those are concerns that largely fall outside the purview of copyright law.

Sincerely,

David Hansen
Executive Director, Authors Alliance

Rachel Brooke
Senior Staff Attorney, Authors Alliance

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?

Copyright is at its core an economic regulation meant to provide incentives for creators to produce and disseminate new expressive works. Ultimately, its goal is to benefit the public by promoting the "progress of science," as the U.S. Constitution puts it. To do that, the Supreme Court has stated, the law's primary objective of copyright is to stimulate artistic creativity "for the general public good."[2] **Because of this, we think new technology like generative AI should typically be judged by what it accomplishes with respect to those goals, and not by the incidental mechanical or technological means that it uses to achieve its ends.**

Based on our experience, we believe these objectives are directly served by the development, deployment, and use of generative AI systems as powerful creative tools for authors, researchers, and creators of all sorts. While these systems are not without risks, such as the creation of deepfakes and the facilitation of the spread of misinformation, these are, by and large, problems outside of the realm of copyright law and can be best addressed by the adoption of best practices, private ordering, and other regulations. We focus our response to this question on how the use of generative AI technology is currently, and is likely to, affect authors and researchers, as these are the types of creators we represent.

Generative AI systems have already shown tremendous potential to benefit authors, including authors of highly creative works. This is accomplished in a variety of ways: by simplifying administrative tasks inherent to being a working author (such as generating pitch letters, marketing materials, and website copy); by aiding in preliminary and editing tasks like ideation and checking grammar and spelling; and, perhaps most importantly, by serving as a powerful creative tool for authors.

Authors of all sorts increasingly rely on generative AI tools to enhance their creative processes, despite the fact that these systems are still nascent and their possible uses for creators are still evolving. In addition to publicly available generative AI systems like ChatGPT and Claude, companies such as Sudowrite have launched subscription-based models specifically for authors, allowing its users to integrate generative AI systems into their writing. In the short time that these tools have existed, they have already evolved to become more useful as the underlying AI models improve. Sudowrite, for example, was founded in 2020, and has altered its generative AI

[2] *Twentieth Century Music Corp. v. Aiken*, 422 U.S. 151, 156 (1975).

system several times in response to feedback from authors about how the tool could be more useful.[3] While some authors have been using AI systems to support their authorship for years,[4] recent advances in machine learning and AI models have enabled new ways of using these tools.

We have seen numerous, compelling examples of text-producing generative AI systems now being used as new creative tools for authors, enabling them to discover and explore new modes of expression. Sasha Stiles, a poet and AI researcher, uses generative AI to create innovative poetry by using a model she created that is trained on her own work. Her recent book of poems, *Technelegy*, contains poems presented as collaborations between the author and generative AI system. The book has received critical acclaim and has challenged the ways in which readers think of both poetry and "human and computer relation."[5]

Tim Boucher, a science fiction writer and artist, has used generative AI to create a series of nearly 100 science fiction books.[6] He has experimented with different forms of "collaboration" with generative AI systems—from using them for ideation to using them to produce first drafts, to using them for late-stage editing. He has also used generative AI systems to produce text he uses as speech for characters in his works which are themselves AI entities. Boucher does not see his works as prototypical novels with a conventional narrative arc, but as nonlinear works with "interlocking pieces," or "slice of life stories,"[7] which lend themselves to the sometimes fragmented and dreamlike nature of generative AI systems' outputs.

Seeing the potential of generative AI to support authorship, some authors are even beginning to write and publish books of prompts that other authors can use as inputs in generative AI systems in support of their own authorship. Mira Gold, an author who has created AI-assisted works under different pen names, recently published a book of prompts for authors who wish to use generative AI as a creative tool to use.[8] Amit Gupta, founder of Sudowrite, noted that books of prompts (or guides to writing prompts) for generative AI systems are becoming more popular among authors who use generative AI systems.[9] Collections of prompts for generative AI is a

[3] *What's New?*, Sudowrite, https://feedback.sudowrite.com/changelog/ (last visited Oct. 19, 2023).

[4] Elizabeth A. Harris, *Peering Into the Future of Novels, With Trained Machines Ready*, N.Y. Times, Apr. 20, 2023, https://www.nytimes.com/2023/04/20/books/ai-novels-stephen-marche.html (noting that author Stephen Marche "has been writing with and about artificial intelligence since 2017").

[5] Charlotte Kent, *Sasha Stile's Technelegy*, Brooklyn Rail, May 2022, https://brooklynrail.org/2022/05/books/Sasha-Stiless-Technelegy.

[6] Tim Boucher, *'I'm Making Thousands Using AI to Write Books'*, Newsweek, May 15, 2023, https://www.newsweek.com/ai-books-art-money-artificial-intelligence-1799923.

[7] Zoom Interview with Tim Boucher (Sept. 20, 2023).

[8] Mira Gold, Romance in the Digital Age: 100 Templates To Craft AI-Powered Serialized Stories (AI Writers Connection Book 1) (2023).

[9] Zoom Interview with Amit Gupta (Sept. 28, 2023); *see, e.g.*, Sean Lowery, 104 Fantasy & Sci-Fi Story Prompts: From Distant Galaxies to Mythical Kingdoms: Your Next Epic Begins Here (2023).

totally new category of textual work that could not have existed until very recently, and its emergence shows both authors' interest in using generative AI systems as a tool and the new opportunities for creators that these systems have brought about.

Generative AI also enables authors to express themselves in new ways: image-generating systems like DALL-E and Midjourney enable authors to create illustrations to accompany their textual works where they otherwise might not be able to. Kristina Kashtanova, whose graphic novel, *Zarya of the Dawn*, ignited much of the debate around copyright and generative AI, was a professional photographer prior to creating the graphic novel.[10] Midjourney allowed Kashtanova to create something they would presumably not otherwise have been able to, enriching the text they wrote as a new means of self-expression. Tim Boucher also uses generative AI systems to produce images that accompany his stories. While Boucher is a graphic artist himself, he has said that the time and cost involved in creating these illustrations by hand would severely limit the amount of time he could spend writing, and would make his project too cost-prohibitive.[11]

For self-published authors and authors that do not use traditional publishers, generative AI systems have also been used to create book covers[12] that may serve to entice readers to read or purchase their works—advancing their goals of reaching readers and profiting from their creative labors. Book cover design is challenging at both a practical, design level and in terms of the art of appealing to readers. Yet despite the common adage, book covers are essential tools for drawing in readers.[13] Generative AI as a book cover design tool therefore makes the business of being a successful author more accessible, working against the near-constant declines in author incomes.[14]

Generative AI is also a powerful tool for academic nonfiction authors. AI models are increasingly being used as part of the research and writing process for nonfiction works, such as scientific research articles and books. Like fiction authors, researchers are turning to generative

[10] The Geek in Review Podcast, *From Pain to Creativity: How AI Helped Kristina Kashtanova Illustrate Their "Zarya of the Dawn" Story*, 3 Geeks and a Law Blog (Apr. 7, 2023), https://www.geeklawblog.com/2023/04/from-pain-to-creativity-how-ai-helped-kristina-kashtanova-illustrate-her-zarya-of-the-dawn-story-featuring-richmond-laws-ashley-dobbs-and-roger-skalbeck-tgir-ep-196.html.

[11] Boucher, *supra* note 7.

[12]Jess Weatherbed, *Not Even NYT Bestsellers Are Safe from AI Cover Art*, The Verge, May 15, 2023, https://www.theverge.com/2023/5/15/23724102/sarah-j-maas-ai-generated-book-cover-bloomsbury-house-of-earth-and-blood.

[13]Jane Friedman, *Don't Crowdsource Your Cover Design*, Pubs. Weekly, May 24, 2019, https://www.publishersweekly.com/pw/by-topic/authors/pw-select/article/80162-don-t-crowdsource-your-cover-design.html.

[14] *Six Takeaways from the Author Guild 2018 Author Income Survey*, Authors Guild, Jan. 5, 2019, https://authorsguild.org/news/six-takeaways-from-the-authors-guild-2018-authors-income-survey/ (demonstrating consistent declines in published author incomes between 2009 and 2018 for both traditionally published and self-published authors).

AI systems like ChatGPT to ideate and to find more concise and accurate ways of writing academic nonfiction works.[15] And tools like elicit.org are increasingly used to assist with literature reviews, summarizing conclusions and highlighting areas of uncertainty or disagreement within the literature.[16]

AI in this context allows researchers to refocus attention on the core of their work—generating new scientific discoveries, innovative new technologies, and new applications in a wide range of fields—rather than on the process of writing and communicating their research. For many of these authors, as explained by Michael Eisen, a computational biologist at the University of California, Berkeley (also former editor-in-chief of the journal *eLife* and a member of the Authors Alliance Advisory Board), "[i]t's never really the goal of anybody to write papers—it's to do science."[17]

Perhaps most significantly, AI has the potential to lower barriers to participation in the scientific community by allowing more researchers to contribute. As explained by a recent review article, there are dozens of efforts underway to apply AI systems in ways that will lower language barriers in scientific communication: "[M]ost scientific papers in the world literature are written in English by non-native English speakers. Non-native English-speaking scientists face many difficulties in writing clearly, succinctly, and without grammatical errors."[18]

Generative AI can also support the progress of science as part of the research process itself. It can achieve this by speeding up data processing, automating data acquisition, making it easier and faster to write code, generating new research hypotheses, and more.[19] Other commenters' submissions, particularly the submission from the UC Berkeley Library, explain in more detail how generative AI (and other AI technology) is already being used to aid researchers by allowing them to more efficiently and effectively explore large corpora of data in fields ranging from computer science to medicine.

These examples only scratch the surface of the many ways that generative AI directly supports authorship and advances the goals of copyright. We are aware that there are many who strenuously object to generative AI on the basis that generative AI systems may decrease

[15] Gemma Conroy, *How ChatGPT and Other AI Tools Could Disrupt Scientific Publishing*, 622 Nature 234, 235, Oct. 12, 2023, https://doi.org/10.1038/d41586-023-03144-w.

[16] Michele Salvagno, Fabio Silvio Taccone & Alberto Giovanni Gerli, *Can Artificial Intelligence Help for Scientific Writing?*, 27 Critical Care 75, 76, Feb 25, 2023, https://doi.org/10.1186/s13054-023-04380-2.

[17] Conroy, *supra* note 15.

[18] Auro Del Giglio & Mateus Uerlei Pereira da Costa, *The Use Of Artificial Intelligence to Improve the Scientific Writing of Non-Native English Speakers*, 69 Revista da Associação Médica Brasileira, Sept. 18 2023, https://doi.org/10.1590/1806-9282.20230560.

[19] Richard Van Noorden & Jeffrey M. Perkel, *AI and Science: What 1,600 researchers Think*, 621 Nature 672, Sept. 17, 2023, https://doi.org/10.1038/d41586-023-02980-0.

incentives for creators by offering consumers cheap and easy substitutes for their work. We understand that like many other new technologies, generative AI may disrupt existing patterns of consumption, but on the whole we believe this new technology has already shown incredible promise not as a substitute but as a tool to foster creativity and innovation in ways that were previously impossible.

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.

No, new legislation to address copyright or related issues with generative AI is not warranted at this time. Both the development of AI models and the ways in which authors and other creators use generative AI systems are still evolving, making copyright legislation to address issues with generative AI premature. Legislating around nascent technologies can pose unexpected problems down the line as the technologies and legal issues they pose continue to evolve.[20] And at a higher level of generality, the purpose of copyright—incentivizing creativity for the public benefit—is served by innovative tools like generative AI systems that benefit not only creators, but the public at large, in a variety of ways. Potential socially beneficial uses include advances in medical imaging,[21] making psychotherapy more accessible,[22] and supporting student learning.[23] While only time will tell which, if any, of these benefits come to fruition, this uncertainty further underscores the appropriateness of a "wait and see" approach in terms of legislative copyright reform.

Time and again we have seen immense value in the U.S. copyright system's flexibility and ability to adapt to innovative new technologies,[24] such as in the context of home recording of

[20] *See, e.g.*, Edward Felten, *The Chilling Effects of the DMCA*, Slate, Mar. 29, 2013, https://slate.com/technology/2013/03/dmca-chilling-effects-how-copyright-law-hurts-security-research.html (explaining how DMCA § 1201 limits security research); Michael Smith & Marshall Van Alstyne, *It's Time to Update Section 230*, Harvard Bus. R., Aug. 12, 2021, https://hbr.org/2021/08/its-time-to-update-section-230 (listing difficulties with applying CDA § 230 to social media platforms on urgent questions like Facebook's role in the January 6th Capitol Riots and Backpage's role in the sexual exploitation of children in 2021).

[21] Steve Lohr, *A.I. May Someday Work Medical Miracles. For Now, It Helps Do Paperwork*, N.Y. Times, June 26, 2023, https://www.nytimes.com/2023/06/26/technology/ai-health-care-documentation.html.

[22] Ashley Andreou, *Generative AI Could Help Solve the U.S. Mental Health Crisis*, Psych. Today, Mar. 9, 2023, https://www.psychologytoday.com/us/blog/the-doctor-of-the-future/202303/generative-ai-could-help-solve-the-us-mental-health-crisis.

[23] Khari Johnson, *Teachers Are Going All In on Generative AI*, Wired, Sept. 15, 2023, https://www.wired.com/story/teachers-are-going-all-in-on-generative-ai/.

[24] See Ian Hargreaves, *Digital Opportunity: A Review of Intellectual Property and Growth* (2011) at 44-46, https://assets.publishing.service.gov.uk/media/5a796832ed915d07d35b53cd/ipreview-finalreport.pdf (reviewing UK intellectual property law and proposing changes to introduce flexibility into its copyright act to achieve the kinds of technological innovation supported in the U.S.).

over the air television,[25] online search engines,[26] and plagiarism detection software,[27] to name a few. Despite the fact that none of these uses were anticipated when the copyright laws governing them were enacted, copyright proved flexible enough to answer the legal questions involved. The development and use of generative AI models and systems should be no different.

Since OpenAI released ChatGPT (based on GPT-3, the most well-known of the text-generating AI models) in November 2022,[28] other generative AI systems have been released and gained in popularity. As more competition emerges in the generative AI space, companies have responded to criticisms of their generative AI systems, revising policies and terms of service accordingly. For example, Anthropic, developer of the "AI assistant," Claude, which many writers use as part of their creative processes, seeks to create products which are "helpful, honest, and harmless" with an emphasis on safety.[29] Perhaps responding to criticisms of ChatGPT and other generative AI systems, Anthropic notes that Claude "is much less likely to produce harmful outputs."[30] By the same token, following the release of MidJourney, DALL-E, and other image-generating systems like it, Adobe released its own image-generating AI system, Firefly. Some of these earlier image-generating AI systems have faced substantial criticism for the extent to which its tool has produced outputs that bear a striking resemblance to copyrighted works, whereas Firefly is trained on only openly-licensed images and images in which Adobe holds rights—which it touts as major selling points.[31] Companies that create generative AI systems have also made alterations in response to public criticism and revelations of unintended legal issues. For example, following news of ChatGPT reproducing long excerpts from books,[32] OpenAI appears to have made alterations to the system to limit the amount of text from copyrighted works it can reproduce.[33]

[25] *Sony Corp. of Am. v. Universal City Studios*, 464 U.S. 417 (1984).

[26] *Kelly v. Arriba Soft Corp.*, 336 F.3d 811 (9th Cir. 2003).

[27] *AV ex rel. Vanderhye v. iParadigms, LLC* 562 F.3d 630 (4th Cir. 2009).

[28] *See* Alyssa Stringer & Kyle Wiggers, *ChatGPT: Everything You Need to Know About the AI-Powered Chatbot*, TechCrunch, Oct. 17, 2023, https://techcrunch.com/2023/10/17/chatgpt-everything-to-know-about-the-ai-chatbot/ (listing timeline including release date).

[29] *Introducing Claude*, Anthropic, Mar. 14, 2023, https://www.anthropic.com/index/introducing-claude.

[30] *Id.*

[31] *Adobe Firely*, Adobe, https://www.adobe.com/sensei/generative-ai/firefly.html (last visited Oct. 20, 2023) ("Trained on Adobe Stock images, openly licensed content, and public domain content, Firefly is designed to be safe for commercial use.")

[32] Chang et. al., *Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4*, Apr. 28, 2023, https://doi.org/10.48550/arXiv.2305.00118 (demonstrating ChatGPT's ability to reproduce lengthy verbatim excerpts from popular fiction).

[33] Compl. at ¶¶ 88-90, *Authors Guild v. OpenAI*, No. 1:23-cv-08292 (S.D.N.Y. Sept. 19, 2023). ("Until very recently, ChatGPT could be prompted to return quotations of text from copyrighted books with a good degree of accuracy . . . . Now, however, ChatGPT generally responds to such prompts with the statement, 'I can't provide verbatim excerpts from copyrighted texts.').

As discussed in question 1, the way that authors and other creators use generative AI to support their work has also changed over time, and will presumably continue to shift as the AI models and generative AI systems continue to develop, and in response to public reactions to the use of these tools. Copyright issues with generative AI can be broken down into three separate questions—whether training data is fair use, the copyrightability of generative AI outputs, and the question of what to do when outputs infringe existing copyrighted works—in which the law is already adequately developed rather than in need of new legislation.

As we explained in a recent joint letter to Congress on this question:

> We owe these value-driving innovations to the broad and flexible framework that Congress wisely created when fashioning our copyright regime. In essence, that framework ensures a fair reward to creators while also enabling technological innovation and follow-on creativity. That dynamic structure is the reason that the United States is not only the most successful creative economy in history, but is also the primary source of the technological innovation that has driven the global economy for over half a century. It is the reason why today's AI leaders have chosen to build their innovative products in the United States, rather than elsewhere.[34]

Training

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.

In the vast majority of cases, the use of copyrighted works to train AI models constitutes fair use. This is based on our understanding of how the four fair use factors apply to the use of copyrighted works as training materials for AI models, as well as holdings in multiple fair use cases including *Authors Guild v. Google* ("*Google Books*"), *Authors Guild v. HathiTrust*, and *A.V. v. iParadigms*. The use of copyrighted works as training data for generative AI models is in most cases a non-expressive use, as it is done as an intermediate step in producing non-infringing content,[35] such as by extracting non-expressive information such as patterns, facts, and data in or about the work.

In addition, fair use of copyrighted works serves the public benefit—an important consideration to be weighed against a copyright owner's interests. This is because the quality and quantity of training data for generative AI has several practical implications for the utility of these tools for

[34] Letter from "broad coalition of public interest organizations, creators, academics, and others" to Members of U.S. Senate and U.S. House of Reps., Sept. 11, 2023, available at https://www.authorsalliance.org/wp-content/uploads/2023/09/AI-Coalition-Letter-9.11.2023-updated.pdf.

[35] Matthew Sag, *Copyright Safety for Generative AI*, Aug. 10, 2023, 61 Houston L. Rev. 2 (forthcoming 2023) at 109-110, available at http://dx.doi.org/10.2139/ssrn.4438593.

authors and other creators. For example, were the use of training data not a fair use, AI models could be trained on only "safe materials," like public domain works or materials specifically authorized for such use. This would be detrimental to the utility and public benefit of these tools: limiting the datasets they are trained on to public domain materials, for example, and omitting virtually all creative content from the past hundred years, would tend to amplify bias and the views of an unrepresentative set of creators.[36]

Fair Use Analysis

At least three of the four fair use factors support the conclusion that the use of copyrighted materials as training materials for generative AI models is a fair use.

First, the purpose and character of the use of copyrighted works as training materials is one entirely different than the purpose for which those works were created. Using in-copyright works to create a system or model with a new and different purpose from the works themselves, which does not compete with those individual works in any meaningful way, is a prototypical fair use. Like using copyrighted works to create a full-text searchable database—the use at issue in *Google Books*—using copyrighted works to create a tool capable of generating new text and images could not be more different than the purpose of the copyrighted works.

At the present, many generative AI models and systems have been developed by non-commercial entities,[37] though some generative AI systems are deployed by commercial entities. But even assuming that the use of copyrighted works as training data is considered commercial, these uses are sufficiently transformative so as to overcome the commercial nature of the entities that use them.[38] In factor one analysis, it is important to weigh any commerciality against both the transformative nature of the use and the degree to which these tools "serve[] the interests of the public."[39] In *iParadigms*, a court also considered whether using copyrighted works to create a tool with substantial public benefit was transformative, and found that it was, despite the fact that the user was a commercial entity.[40] While generative AI tools differ substantially from plagiarism-detection tools, both are new and transformative uses that serve substantial public benefits. Because of the transformative nature of the use of copyrighted works as training materials for AI models and their substantial public benefit, this factor weighs in favor of fair use.

[36] Amanda Levendowski, *How Copyright Can Fix AI's Implicit Bias Problem*, 93 Wash. L. Rev. 579, 615-16 (2018), https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2/.

[37] *Revolutionizing Image Generation by AI: Turning Text into Images*, Ludwig-Maximilians-Universität München, Sept. 1, 2022, https://www.lmu.de/en/newsroom/news-overview/news/revolutionizing-image-generation-by-ai-turning-text-into-images.html (describing professors' work developing Stable Diffusion).

[38] *Authors Guild v. Google, Inc.*, 804 F.3d 202, 219 (2nd Cir. 2015).

[39] *Perfect 10, Inc. v. Amazon.Com, Inc.*, 487 F.3d 701, 722 (9th Cir. 2007).

[40] *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 639 (4th Cir. 2009).

Second, the nature of the copyrighted works used as training materials for AI models does not weigh strongly in either direction. Evidence suggests that these models are trained on a wealth of publicly available data, much of which is factual in nature and subject only to thin copyright protection. There is nothing to indicate that the training datasets in question contain works that are unusually creative or close to the core of copyright, though of course highly creative works like novels and plays are also likely included as training materials.[41] *iParadigms* is also instructive regarding this factor. In the district court hearing, the court noted that, even though some of the works in question were creative and thus close to the core of copyright, "this factor is of lesser import because the allegedly infringing use makes no use of any creative aspect of the student works."[42] Similarly, in the case of generative AI, the copyrighted works are used to extract non-expressive elements despite the fact that creative works themselves are used.

Third, regarding the amount and substantiality of the portion of the works used as training materials, it is not entirely clear whether copyrighted works used to train AI models are used in their entirety. But even assuming that the works are used in their entirety, this is reasonable in light of the purpose of the use. Generative AI systems are effective at generating high-quality images, text, and other outputs *because* of the vast size of the datasets they are trained on,[43] making it reasonable for developers to try to maximize the amount of data these models ingest in order to increase the public benefit of these tools. Because factor three directs courts to consider whether the amount and substantiality of the portion used is "reasonable in relation to the purpose of the copying,"[44] this factor weighs in favor of fair use.

Fourth, the effect of the use on the market for the copyrighted works used to train generative AI systems is unlikely to be significant based on the lack of a substitutional effect between the individual works themselves and the generative AI systems based on AI models that use them as training materials. For example, in the context of textual works, it is highly unlikely that a reader interested in reading a specific book included in GPT-3's training dataset would turn to ChatGPT

[41] *See* Chang, *supra* note 32.

[42] *A.V. v. iParadigms, Limited Liability Company*, 544 F. Supp. 2d 473, 483 (E.D. Va. 2008).

[43] Katherine Lee, A Feder Cooper & James Grimmelmann, *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain*, July 7, 2023, at 27, http://dx.doi.org/10.2139/ssrn.4523551 ("[S]cale also confers new capabilities. Today's generative-AI models are able to produce incredible content, in large part because of their large scale)(citations omitted); Helen Toner, *What Are Generative AI, Large Language Models, and Foundation Models?*, Ctr for Sec. & Info. Tech., May 12, 2023, https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/ ("A key finding of the past several years of language model research has been that using more data and computational power to train models with more parameters consistently results in better performance.").

[44] *Campbell v. Acuff-Rose Music*, 510 U.S. 569, 586 (1994).

to obtain information about that book in lieu of purchasing or checking out that book from the library.[45]

Like the full-text searchable database at issue in *Google Books*, a tool that does not reproduce the expressive elements of the works used in creating it cannot be said to serve as a market competitor for individual underlying works. And just because a tool can be used to facilitate infringement—e.g., by creating an output that bears a strong similarity to an existing copyrighted work—does not mean that the tool itself infringes. For example, a court found in *American Geophysical Union v. Texaco* that the defendant had infringed on the plaintiff's reproduction rights by making unauthorized photocopies of scientific articles.[46] It was the human actors using the tool to produce an infringing copy that were held liable, and not those responsible for developing and deploying the photocopier on which those copies were made.

We discuss factor 4 in greater length in our response to question 8.5, below.

<u>Non-Expressive Use</u>

The use of copyrighted works as training material for generative AI models can also be understood as a non-expressive fair use. The *Google Books* and *HathiTrust* cases are particularly useful in explaining and understanding why the use of copyrighted material as training data for LLMs is a non-expressive fair use. As a factual matter, the creation of HathiTrust itself had as one goal conducting digital humanities research using machine learning—the very use at issue here.[47] But the nature of the uses in *Google Books* and *HathiTrust*—broadly speaking, creating full-text searchable databases of works that did not make substantial expressive content publicly available and thus could not serve as substitutes for the works themselves—are also directly parallel to the use of copyrighted material as training materials for AI models.

Extracting non-expressive information such as metadata from copyrighted material does not replicate the expressive text itself. One potential source of confusion regarding this point lies in the fact that generative AI systems' outputs are themselves "expressive," in that they would be eligible for copyright protection were they created by a human. But except in rare cases of memorization, those outputs do not reproduce the expressive text of the inputs they were trained on. The confusion this can engender also shows why it is important to consider the copyright questions involved in generative AI separately rather than collapsing them into a single inquiry that unduly focuses on the nature of the outputs of these systems.

[45] *See Authors Guild v. Google, Inc.*, 804 F.3d 202, 223 (2nd Cir. 2015) (factor four might weigh against fair use when the use "results in widespread revelation of sufficiently significant portions of the original as to make available a significantly competing substitute").

[46] 37 F.3d 881 (2nd Cir. 1994).

[47] Sag, *supra* note 35 at 119.

The notion that non-expressive (also called "non-consumptive") uses do not infringe copyrights is based in large part on a foundational principle in copyright law: copyright protection does not extend to facts or ideas. Like the Google Books project (as well as text data mining), generative AI systems use data (like copyrighted works) to produce information *about* the works they ingest, including abstractions and metadata, rather than replicating expressive text. The use of copyrighted works as training materials is a fair use "because these models 'learn' latent features and associations within the training datasets, they do not memorize snippets of original expression from individual works."[48]

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?

The effect of the market for the copyrighted works used as training materials should be measured by the effect of the generative AI system upon the market for the original copyrighted work alleged to have been copied, and not the effect of the system's outputs on the broader market. The fair use statute says that the inquiry should focus on "the effect of the use upon the potential market for or value *of **the** copyrighted work*,"[49] not about the effects of the byproducts of that use—in this context, the dissemination of outputs of generative AI systems. Fair use has proven a durable and flexible doctrine in the nearly fifty years since its statutory enactment, and the development of generative AI does not merit changing this sound approach. We can think of no fair use case that has ever assessed market harm by adopting such a broad approach to market harm, but numerous instances in which courts have rejected such an approach.[50]

We hope the Office will take this point seriously. The question under the fourth factor market harm assessment is not merely one of line drawing to determine where a potential market harm might be cognizable. Copyright law is necessarily limited to protecting particular creative expressions, not ideas or concepts, or general classes of works, nor an amorphous "body of works" by a particular author or authors. If fair use were to incorporate these broader

[48] Sag, *supra* note 35 at 101-102.

[49] 17 U.S.C. 107(4) (emphasis added).

[50] *See, e.g.*, *Penelope v. Brown*, 792 F. Supp 132, 138 (D. Mass 1992) (declining to extend market harm analysis to work competing in same general market); *Arica Inst., Inv. v. Palmer*, 970 F.2d 1067, 1078 (2d Cir. 1992) ("[T]he relevant market effect is that which stems from defendant's use of plaintiff's 'expression,' not that which stems from defendant's work as a whole. *Wright,* 953 F.2d at 739. Defendant, after all, is perfectly entitled to create a competing work. The copyright statute simply constrains her from using plaintiff's original expression in doing so. Thus, in cases such as this where a meaningful distinction may be drawn between the infringing portions of defendant's work and the work as a whole, we need only consider the market effect of the infringing portions"); *Ty, Inc. v. Pubs. Int'l, Ltd.*, 292 F.3d 512 (7th Cir 2002).

conceptions of "market harm," it would seriously risk compromising the free expression rights of subsequent users and creators to compete in the broader marketplace of ideas.

This question, and the various arguments it will surely elicit from different stakeholders, both demonstrates the importance of separating the input question (i.e., whether the generative AI systems themselves harm the market for the copyrighted works used as training materials) from the output question (i.e., whether outputs of generative AI systems might compete in the market with the training materials the underlying AI model is trained on) and the temptation to do so. Because generative AI systems generate "quasi-expressive text" in the output stage, creators' fears of job displacement from ripple effects of this phenomenon may not be unfounded. But this does not disturb the conclusion that the use of copyrighted works as training materials remains a fair use. Insofar as legislative or other legal reform is needed to address new social harms posed by generative AI, that reform should be the provenance of other areas of the law, and not copyright.

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?

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

We address questions 10.3 and 10.4 together.

No. Both compulsory licensing and extended collective licensing (ECL) share common, and in our view, serious flaws that make them an extremely unappealing solution to provide permissioned access to copyrighted works as training materials for generative AI models.

As explained in question 8, we believe that the use of copyrighted works as training materials is in most circumstances fair use, obviating the need for a license of any kind. There are currently over a dozen copyright lawsuits pending that raise questions about how fair use may apply to training datasets,[51] and we believe the courts should be given an opportunity to answer before any kind of licensing scheme is seriously considered. If a license system were architected and put in place to operate alongside fair use, it would run the risk of undermining the scope of that right and pressure lawful users into unnecessary licensing arrangements.[52]

[51] *Master List of Lawsuits v. AI, ChatGPT, OpenAI, Microsoft, Meta, Midjourney & Other AI Cos.*, ChatGPT is Eating the World, https://chatgptiseatingtheworld.com/2023/10/19/master-list-of-lawsuits-v-ai-chatgpt-openai-microsoft-meta-midjourney-other-ai-cos/ (last visited Oct. 30, 2023).

[52] James Gibson, *Risk Aversion and Rights Accretion in Intellectual Property*, 116 Yale L.J. 882 (2007) (reviewing the ways that licensing regimes and related litigation pressure can diminish the role of fair use)

Even if the courts in these cases do not find in favor of fair use for AI training datasets, we do not believe compulsory licensing or extended collective licensing are workable solutions to authorize use of copyrighted works as training materials, for several reasons.

First, both compulsory licensing and extended collective licensing schemes are logistically infeasible because of the scale and complexity of the training datasets needed to train AI models. One of the key advantages of some of the largest and most successful AI models is that they are built upon an extremely large and diverse array of training materials. While creators of many of the largest models such as GPT-3 and GPT-4 have not fully disclosed the sources of training datasets, it is known that others have relied on sources such as Common Crawl (millions of works from across the internet), Books3, Arxiv, Stackexchange, and many millions others incorporating materials from every corner of the internet and across content types, including books, newspapers, websites, code, social media posts, legal materials, emails, YouTube subtitles, and more.[53]

These training datasets likely include millions of works owned by copyright holders from around the world. In a legal environment where permission is required to use copyrighted materials as training materials, licenses would need to cover a broad and diverse array of content. Doing otherwise (for example, creating a compulsory licensing or ECL system that applies to only a subset of copyrighted content) would both unfairly privilege certain rightsholders over others, while also exacerbating issues of bias in the models themselves (e.g., by encouraging use of only public domain content or easily licensable corporate-owned assets).[54]

Second, at this scale, training datasets are likely to include massive numbers of orphan works for which licensing would be inappropriate. To effectively license those works under either an ECL or compulsory licensing scheme, adequate information about copyright ownership and licensing arrangements would either need to exist already or be readily obtainable in order to remit license payment to the rightsholders. While some rightsholders choose to make such information easily findable (for example, through registration with the U.S. Copyright Office), those registrations represent a tiny subset of all copyrighted works. For many works, however, no copyright information is available and—particularly for online works such as blog posts, online comments, Wikipedia contributions, and more—identifying and locating rightsholders would be virtually impossible.

As the Office's previous studies on orphan works and mass digitization have highlighted, searches for rightsholders in far more narrowly scoped projects have incurred significant costs,

[53] Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling,* preprint posted Dec. 31, 2020, https://arxiv.org/abs/2101.00027.

[54] Levendowski *supra* note 36 at 610.

with mostly had poor results.[55] If a broad-based compulsory licensing or ECL regime were meant to be anything more than just a tax on generative AI systems, there would need be a way to pay out funds collected to the appropriate rightsholders, and that includes resolving the likely massive orphan works problem embedded within these training datasets. Experience from abroad in jurisdictions with systems that encourage searches for orphan works owners is not encouraging.[56]

Third, even for commercially valuable works that are not orphaned, copyright ownership is far from clear for many works, making remuneration at scale destined for protracted litigation. While existing compulsory licensing regimes focus on areas of practice where licensing and transfer of rights has largely been standardized, most other areas of copyright do not have such clarity. For example, even with commercially published books, publication contracts, i.e., private agreements between authors and publishers, take a wide variety of approaches to allocating ownership rights across geographic locations, formats, and time. Whether, for example, using works as training materials for AI models would constitute an "electronic work" under those licenses would be an important question. This issue has been litigated and remains unresolved and largely unresolvable without close examination of each contract.[57]

Finally, no party exists to administer such licenses, nor is it likely that an appropriate party or parties will emerge. As the Office observed in 2011 and again in 2015, ECL systems have typically been applied only to "limited types of works and uses, such as the use of published works for educational and scientific purposes, or the reproduction of works within an organization solely for internal use. . . . applying ECL to a mass digitization project that provides access to a wide range of works . . . would be a dramatic extension of the concept."[58] While a handful of EU countries have explored wider application (for example, to out of commerce

[55] *See, e.g.*, Cornell University Library, Response to the Notice of Inquiry Concerning Orphan Works (Mar. 23, 2005), https://www.copyright.gov/orphan/comments/OW0569-Thomas.pdf (reporting that for a rights clearance study of 343 in-copyright books, Cornell Library was only able to identify or locate rightsholders for 58% of those books after spending $50,000 in staff time conducting research); Carnegie Mellon University Libraries, Response to Notice of Inquiry Concerning Orphan Works (Mar. 22, 2005), https://www.copyright.gov/orphan/comments/OW0537-CarnegieMellon.pdf.

[56] Bartolomeo Meletti, *Report on the EnDOW Final Conference*, CREATe, May 22, 2018, https://www.create.ac.uk/blog/2018/05/22/report-endow-final-conference/.

[57] *Random House v. Rosetta Books*, 150 F. Supp. 2d 613 (S.D.N.Y. 2001), *judgment aff'd*, 283 F.3d 490 (2d Cir. 2002) (authors likely retained e-book rights given language of the contract); *HarperCollins Pubs. v. Open Road Integrated Media*, 7 F. Supp. 3d 363 (S.D.N.Y. 2014) (publisher owned electronic rights).

[58] U.S. Copyright Office, Legal Issues in Mass Digitization: A Preliminary Analysis and Discussion Document, Oct. 2011, at 36, https://www.copyright.gov/docs/massdigitization/USCOMassDigitization_October2011.pdf; *see also* U.S. Copyright Office, Report on Orphan Works and Mass Digitization (2015), https://www.copyright.gov/orphan/reports/orphan-works2015.pdf.

works), those efforts are still new and are being done within the CMOs that have a history and experience working with ECLs.[59]

A credible system of adequate scale and scope would require an incredible investment of resources, and there are no clear candidates to make this investment. More than ten years ago, in the narrow field of book publishing, Google estimated a cost of $35 million to establish the Books Rights Registry ("BRR") as part of its negotiated settlement in the *Google Books* litigation.[60] The BRR which would have fulfilled many of the functions of a collective management organization administering an ECL-like regime under the proposed settlement, representing the interests of authors. While Google was willing to pay costs associated with such a system at that time, subsequent efforts (e.g., the Office's mass digitization ECL pilot) revealed little appetite among others for such an investment.[61] The costs of an ECL system broad enough to address the scope and scale of materials used for AI training datasets would likely dwarf the costs of any similar system we have seen to date.

More significantly, there are no CMOs nor combinations of existing CMOs that could adequately represent the range of rightsholders who would be implicated in training datasets, as the U.S. lacks the kinds of CMOs present in other jurisdictions that would be needed in order to represent the full range of rightsholders' interests. Even in book publishing—an industry far more mature than industries representing many of the content types necessary to be licensed for AI training—the U.S. does not have an adequate CMO.[62] The development of a whole cadre of new CMOs—representing groups as diverse as illustrators, graphic artists, authors, book publishers, academics, journalists, newspapers, photographers, software developers, bloggers, social media creators, and more—with sufficient infrastructure and wherewithal to represent their various constituencies would require years of effort to develop.

In addition to the need to establish new CMOs, those new CMOs would need to fairly represent the interests of rightsholders in their respective fields. The success of the ECL model in other

[59] European Commission, Report on the Use of Collective Licensing Mechanisms with an Extended Effect Under Article 12(6) of Directive 2019/790/EU on Copyright and Related Rights in the Digital Single Market, Nov. 19, 2021, available at https://ec.europa.eu/newsroom/dae/redirection/document/81236.

[60] *See* Jonathan Band, *The Long and Winding Road to the Google Books Settlement*, 27 John Marshall Rev. Intell. Prop. L. 227, 264, n.358 (2009). The books rights registry is probably the closest the U.S. has gotten to an ECL-like system, which would have been created by application of the negotiated settlement to plaintiff publishers and all members of the proposed authors class in that suit. *See also* Pamela Samuelson, *Legislative Alternatives to the Google Book Settlement*, 34 Colum. J. L. & Arts 697, 708 (2011).

[61] Letter from Kathryn Temple Claggett, Acting Register of Copyrights and Director, U.S. Copyright Office to Sen. Charles Grassley and Rep. Diane Feinstein at 1, *available at* https://www.copyright.gov/policy/massdigitization/senate-letter.pdf.

[62] While groups like the Authors Registry have been suggested as potential candidates, the Authors Registry is not actually a CMO, does not have members, and would need significant investment to represent author interests in book publishing.

jurisdictions (e.g., Nordic countries) has been premised "on the presumption of existence of a representative CMO in the intended field of use."[63] Because ECL allows CMOs to represent both members and non-members, it is important that those CMOs be narrowly scoped such that they can genuinely and fairly represent like groups of rightsholders. This issue of fair representation is in part why Judge Chin rejected the proposed settlement in *Google Books*, in part because it would have allowed one group of authors to represent too many others with disparate interests (e.g., trade book authors and academic authors).[64] Because works represented in AI training datasets implicate the rights of such a diverse group of creators, a large number of CMOs to represent their specific interests would be necessary. Yet this would further multiply the cost of complexity of establishing yet more CMOs to participate in an ECL system for AI training datasets.

Generative AI Outputs

Copyrightability

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. The human authorship requirement in copyright has been enshrined in case law since the late 19th century,[65] and has a constitutional basis in that the intellectual property clause references "authors and inventors," which are necessarily human entities.[66] As discussed further in our response to question 5, revisions to the Copyright Act are infrequent and should be carefully thought out before they are proposed and implemented. Contemporary legislation is not required to establish a proposition firmly rooted in centuries of case law based on a straightforward reading of the Constitution's intellectual property clause. It is a requirement that has seldom been challenged. The advent of generative AI and the sharp increase in AI-generative content in a variety of creative marketplaces has reinvigorated debates in some communities about the issue, but that does not mean that legislation is necessary to address it.

Judicial precedent regarding the human authorship requirement has been clear and consistent, dating back over a century to *Burrow-Giles Lithographic Company v. Sarony*. It is for this reason that the Office and courts considering the issue in the context of generative AI have consistently

[63] Johann Aximm & Lucie Guibault, *Cross-Border Extended Collective Licensing: A Solution to Online Dissemination of Europe's Cultural Heritage*, Amsterdam Law School Research Paper No. 2012-22, Institute for Information Law Research Paper No. 2012-19, at 71, *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2001347.

[64] *Authors Guild v. Google Inc*., 770 F. Supp.2d 666, 679-80 (S.D.N.Y. 2011).

[65] *See Burrow-Giles Lithographic Co. v. Sarony*, 111 U.S. 53 (1884).

[66] *Id.* at 56; *Trade-Mark Cases*, 100 U.S. 82, 94 (1879).

upheld the human authorship requirement,[67] and the rare arguments against it—in this and other contexts— have been rejected.[68] Both the Copyright Office opinion letters and the *Thaler* decision have unequivocally upheld the human authorship requirement, affirmed its applicability to the issue of copyright in AI-generated works, and applied the principle without difficulty.

Instead of proposing revisions to the Copyright Act to enshrine the human authorship requirement in law or clarify the human authorship requirement in the context of AI-generated works, the Office should continue to promulgate guidance for would-be registrants. As case law on the topic continues to unfold, the Office should consider updating its guidance on the human authorship requirement in order to give creators using generative AI as much clarity as possible, which is sufficient to inform copyright registrants about how the doctrine applies to this new type of work.

Moreover, as both generative AI systems and the ways that creators use them change and evolve, the application of the human authorship requirement to content that is AI-generated or AI-assisted may also change. For example, if these tools developed in a way that would give creators more control over the outputs, works created with these tools could potentially be considered works of human authorship. Creating legislation around the human authorship requirement and generative AI could also complicate how creators can use generative AI systems to create copyrightable works, stifling rather than encouraging innovation.

<u>Infringement</u>

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?

Yes, the substantial similarity test is adequately equipped to handle the question of whether a given generative AI output is similar enough to a copyrighted work to be infringing. The substantial similarity doctrine is appropriately focused on protection of creative expression while also providing room for new, creative uses that draw on unprotectable facts or ideas. It is a fact-sensitive, flexible inquiry that federal courts across the country have interpreted differently in different contexts,[69] and there is no reason why it is not suited to cases of infringement regarding

[67] *Thaler v. Perlmutter*, 1:22-cv-1564-BAH (D.D.C. Aug 18, 2023); U.S. Copyright Office Review Board, *Decision Affirming Refusal of Registration of a Recent Entrance to Paradise*, Feb. 14, 2022, https://www.copyright.gov/rulings-filings/review-board/ docs/a-recent-entrance-to-paradise.pdf.

[68] *See id.*; *Naruto v. Slater*, 888 F.3d 418, 426 (9th Cir. 2018).

[69] Christopher T. Zirpoli, Cong. Rsch. Serv., LSB10922, Generative Artificial Intelligence and Copyright Law (Sept. 28, 2023) at 4, https://crsreports.congress.gov/product/pdf/LSB/LSB10922 ("The substantial similarity test is difficult to define and varies across U.S. courts. Courts have variously described the test as requiring, for example, that the works have 'a substantially similar total concept and feel' or 'overall look and feel' or that 'the ordinary reasonable person would fail to differentiate between the two works.' Leading cases have also stated that this determination considers both 'the qualitative and quantitative significance of the copied portion in relation to the plaintiff's work as a whole.'")

AI-generated works. To be sure, issues about who precisely is the infringer in such cases (question 25 of this notice of inquiry), how the derivative work right does or does not apply, and other copyright issues may present new questions for courts applying the substantial similarity test. But substantial similarity is both straightforward and appropriately flexible such that it can be applied without unusual difficulty to allegations of infringement regarding AI-generated works.

The substantial similarity test is also appropriate to address outputs where there is "memorization" in training the AI model such that the output of a generative AI system infringes a copyrighted work. Moreover, because the phenomenon of memorization of training data is relatively rare,[70] it is likely that non-frivolous cases of infringement involving AI-generated works will be infrequent. Law Professor and scholar Matthew Sag has suggested that companies that create and deploy generative AI systems should adopt a series of "best practices" to minimize the chances of memorization of the training materials, and consequently infringing outputs.[71] It is also important to note that companies that create and deploy generative AI tools are incentivized to take measures to minimize the chances of their systems producing infringing outputs based on increased competition in the market.

[70] *See* Sag, *supra* note 35 at 101.

[71] *Id.* at 142-146.