Docket No. 2023–6
Internet Archive's Public Comments in Response to the Copyright Office Study on Artificial Intelligence

These comments are provided on behalf of the Internet Archive, a 501(c)(3) non-profit library. Like other libraries, we work to expand access to knowledge by collecting, archiving, and providing public access to a variety of physical and digital collections. The Internet Archive is based in San Francisco, California, but most of our patrons visit our collections online at archive.org.

Overview and General Principles

Copyright law has been adapting to disruptive technologies since its earliest days; our existing copyright law is adequate to meet the disruptions of today. In particular, copyright's flexible fair use provision deals well with the fact-specific nature of new technologies, and has already addressed earlier innovations in machine learning and text-and-data mining. So while Generative AI presents a host of policy challenges that may prompt different kinds of legislative reform, we do not see that new copyright laws are needed to respond to Generative AI today.

Our comments are guided by three core principles.

First, regulation of Artificial Intelligence should be considered holistically–not solely through the isolated lens of copyright law. As explained in the Library Copyright Alliance Principles for Artificial Intelligence and Copyright, "AI has the potential to disrupt many professions, not just individual creators. The response to this disruption (e.g., support for worker retraining through institutions such as community colleges and public libraries) should be developed on an economy-wide basis, and copyright law should not be treated as a means for addressing these broader societal challenges."[1] Going down a typical copyright path of creating new rights and licensing markets could, for AI, serve to worsen social problems like inequality, surveillance and monopolistic behavior of Big Tech and Big Media.

Second, any new copyright regulation of AI should not negatively impact the public's right and ability to access information, knowledge, and culture. A

[1] Library Copyright Alliance Principles for Artificial Intelligence and Copyright. Available at: https://www.librarycopyrightalliance.org/wp-content/uploads/2023/06/AI-principles.pdf.

primary purpose of copyright is to expand access to knowledge. *See Authors Guild v. Google*, 804 F.3d 202, 212 (2d Cir. 2015) ("Thus, while authors are undoubtedly important intended beneficiaries of copyright, the ultimate, primary intended beneficiary is the public, whose access to knowledge copyright seeks to advance . . . ."). Proposals to amend the Copyright Act to address AI should be evaluated by the impact such new regulations would have on the public's access to information, knowledge, and culture. In cases where proposals would have the effect of reducing public access, they should be rejected or balanced out with appropriate exceptions and limitations.

Third, universities, libraries, and other publicly-oriented institutions must be able to continue to ensure the public's access to high quality, verifiable sources of news, scientific research and other information essential to their participation in our democratic society. Strong libraries and educational institutions can help mitigate some of the challenges to our information ecosystem, including those posed by AI. Libraries should be empowered to provide access to educational resources of all sorts– including the powerful Generative AI tools now being developed.

Question 1. 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 any new technology, AI poses both opportunities and risks. In the current landscape, both tend to be exaggerated. In reality, "[n]ew technologies are neither our savior nor our doom. They are, instead, catalysts of change."[2] The direction of that change depends on how those tools are designed, and how they are used.

At the Internet Archive, we are using AI tools to help make our collections more discoverable and useful. For example, when a book is digitized, we use a custom machine learning model to automatically suggest page boundaries of scanned materials—allowing staff to more easily and efficiently work on digitization processes. Internet Archive engineers have also piloted a metadata extractor, a tool that automatically pulls key data elements from digitized books, such as title,

[2] Karph, Dave "Why Can't Our Tech Billionaires Learn Anything New?" (October 18, 2023) Available at: https://davekarpf.substack.com/p/why-cant-our-tech-billionaires-learn

authors, publisher and publication date. This extra information helps our librarians match the digitized book to other cataloged records, helping to resolve the common library problem of working with items that have limited or inaccurate metadata. AI is also being leveraged to assist in writing descriptions of items for our collection description pages—sometimes reducing staff time on such projects from 40 to 10 minutes per item.

In our experience, use of AI tools has had a positive impact on our library, streamlining the workflows of our librarians and data staff, and making our materials easier to discover. While improving library metadata may not grab the headlines in the same way as some other uses of modern AI tools, it is important to recognize that new copyright or regulatory rules would impact them all the same.

Question 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?

It is essential for the Copyright Office and other policymakers to recognize how narrow Generative AI is as a category within overall AI development and more importantly, the field of computational research and text and data mining. Being part of both the library community and the open web community, Internet Archive is deeply concerned that media hype around Generative AI has caused everyone, including policy makers to over-index on the important but very narrow and specific concerns of copyright holders, at the expense of other extremely important policy issues, such as long standing efforts on privacy reform and antitrust rules necessary to mitigate some of the broad social harms caused by concentration in the media and Big Tech platforms.

For the purposes of these comments, we will highlight two areas of particular concern for the Internet Archive: 1. Data Collection and 2. Data Provenance.

Data Collection

The Internet Archive has been collecting digital materials since 1996, well before this was common practice among other libraries. As a result, we have a wide range of unique collections that are of significant interest to researchers wishing

to use our materials for computational research.[3] As a library, our mission includes supporting access to information for research and scientific progress. We care about being able to collect and preserve our digital history for the long term, enabling researchers to understand how ideas are generated and spread, and to research across disciplines and media types to gather insights about all disciplines of human study and curiosity. To further this mission, we have developed a number of mechanisms for researchers to use our collections.

We see part of our role as providing responsible stewardship over our digital collections, in line with established library practices. The Internet Archive has been an active member of the "Collections as Data" community of practice, working in collaboration with cultural heritage institutions around the world.[4] This community of practice has developed guiding principles for heritage institutions, large and small, that guide them in supporting responsible computational use of collections as data. Internet Archive recently worked with this global community to update guidance, culminating in the release of the *Vancouver Statement on Collections as Data,* in light of the accelerating pace of AI development.[5]

The Internet Archive has also been part of efforts to help equip digital humanities researchers with tools necessary to navigate law, policy, ethics, and risk within text data mining research. *Legal Literacies for Text Data Mining – Cross Border* ("*LLTDM-X*") is a project co-led by Internet Archive and UC Berkeley Libraries, funded by the National Endowment for the Humanities to address law and policy issues faced by U.S. digital humanists whose text data mining research seeks to make use of foreign-held or -licensed content in multinational contexts.

Being part of these library community efforts has informed how we support various forms of computational research. Consistent with the Santa Barbara and Vancouver Statements, we have sought where appropriate to lower barriers to the responsible computational use of our collections. While some of our collections are freely available for public download (such as public domain books), many of our collections are access-restricted or otherwise not set up for the kind of "bulk

[3] Our Terms of Service are clear that access to our collections is granted for "scholarship and research purposes only."
[4] Santa Barbara Statement on Collections as Data, May 20, 2019. Available at: https://zenodo.org/records/3066209
[5] Vancouver Statement on Collections as Data, September 13, 2023. Available at: https://zenodo.org/records/8342171

access" necessary for computational research. In some cases, we have facilitated researcher access to limited datasets concluding with the return or deletion of the data upon completion of the project. And for over a decade, we have worked on a case-by-case basis with researchers to provide access to our collections via virtual machines that can be used to do a variety of text and data mining processes. This set up, which allows researchers access to collections without transferring data off premises, has been used in a variety of academically-affiliated and NSF-funded projects.

Earlier this year, we announced the public launch of Archives Research Compute Hub (ARCH), a new research and education service that helps users easily analyze digital collections computationally at scale.[6] ARCH represents a combination of the Internet Archive's experience supporting computational research through Archive-it Research Services, and a collaboration with the Archives Unleashed project of the University of Waterloo and York University.[7] ARCH recently received funding from the Institute of Museum and Library Services to further expand the service.[8]

Access to digital collections and the computational resources needed to utilize modern AI tools is increasingly centralized in a handful of for-profit companies.[9] Yet thousands of libraries, archives, museums, and memory organizations work with Internet Archive to build and make openly accessible digitized and born-digital collections. Making these collections available to as many users in as many ways as possible is critical to preserving and providing access to knowledge. It is also essential for helping to mitigate the significant bias found in many datasets used to train AI.[10] The development of nonprofit, publicly oriented AI

[6] https://webservices.archive.org/pages/arch
[7] https://archivesunleashed.org
[8] https://www.imls.gov/grants/awarded/lg-254878-ols-23
[9] *E.g., Widder, David Gray and West, Sarah and Whittaker, Meredith, Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI (August 17, 2023). Available at SSRN: https://ssrn.com/abstract=4543807 or http://dx.doi.org/10.2139/ssrn.4543807*
[10] E.g., Amanda Levendowski, How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem, 93 Wash. L. Rev. 579 (2018). Available at: https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2, Khan, Mehtab and Hanna, Alex, *The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability* (September 13, 2022). Forthcoming 19 Ohio St. Tech. L.J. (2023), Available at SSRN: https://ssrn.com/abstract=4217148 or http://dx.doi.org/10.2139/ssrn.4217148

tools rests on the ability of such organizations to have access to high quality data sets such as these.[11]

For these reasons, we urge the Copyright Office to be mindful that changes to copyright law which would risk attaching new copyright liability for libraries supporting computational research and text and data mining, in contravention of established legal precedent, would work against the public interest and disrupt the substantial investments that have been made in this area by educational institutions, libraries, and the federal government.

Data Provenance

As a library, the Internet Archive seeks to support information integrity and authenticity. In the digital realm, this is often realized through metadata practices. Our goal is to provide enough context and history for our patrons to understand, in a verifiable manner, who produced or created a particular digital item, how it came to be part of our collections, and whether that item has been changed over time.

Taking the Wayback Machine as an example, when we collect the actual content of a particular web page, we also collect provenance details. As such, for each capture, users can see specific details such as what entity collected the page, the URL and the time and date stamp of when it was collected. Including provenance information allows Wayback links to be used as a trusted citation in academic or journalistic work, by fact-checkers seeking to verify who said what on the web, and even as admissible evidence in court. In this way, provenance helps to support a healthy information ecosystem by supporting the truth-seeking function of core democratic institutions while allowing ordinary readers to verify the source of a particular claim or assertion.

Presently, most AI chatbots and assistants are not built to show their work or give citations for the claims they make. When such chatbots spit out "answers," the source of the answers is not available; usually this is because these systems work through a form of predictive text, not by searching a library for verifiable sources

[11] E.g., Narechania, Tejas and Sitaraman, Ganesh, *An Antimonoploly Approach to Governing Artificial Intelligence* (2023). Available at: https://www.vanderbilt.edu/vanderbilt-policy-accelerator/governing-artificial-intelligence/

of information and returning those as results. Provenance, which is critical to knowledge creation and sharing, is not currently part of the AI ecosystem.[12]

Some have suggested, in response to this problem, that the government mandate transparency into what information AI systems are training on. While that may be a natural suggestion for legal systems with opt-out provisions, such as the European Union, it will not solve the provenance problem given the enormous size of the datasets used to train cutting edge models. These kinds of source-disclosure rules could make the provenance problem worse. Copyright rules regarding transparency must be carefully crafted so as not to discourage AI companies from disclosing data provenance and citing trustworthy sources.

Regardless of where AI innovation leads, we will always need strong libraries and educational institutions to provide access to authoritative sources for future generations. Copyright law must continue to ensure that we are able to play this essential role in society.

Question 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? How important a factor is international consistency in this area across borders?

Our experience with the *Legal Literacies for Text Data Mining – Cross Border* ("*LLTDM-X*") project described above teaches that different legal regimes across borders can have a substantial impact on research.[13] Today, the European Union (along with some other jurisdictions) have a copyright exception for machine learning, subject to an opt-out in certain commercial contexts.[14] If the United States changes its policy with regard to the applicability of fair use to text and

[12] There are some academic efforts to try to change this. *See, e.g.,* Cargnelutti, Matteo. *Did ChatGPT really say that?": Provenance in the age of Generative AI.*)May 22, 2023) Available at: https://lil.law.harvard.edu/blog/2023/05/22/provenance-in-the-age-of-generative-ai/

[13] Padilla, Thomas, *Wrapping up Legal Literacies for Text and Data Mining – Cross-Border (LLTDM-X)*, October 5, 2023. Available at: https://blog.archive.org/2023/10/05/wrapping-up-legal-literacies-for-text-and-data-mining-cross-border-lltdm-x/

[14] See, e.g., COMMUNIA *Policy Paper #15 on using copyrighted works for teaching the machine* (April 26, 2023) Available at: https://communia-association.org/policy-paper/policy-paper-15-on-using-copyrighted-works-for-teaching-the-machine/

data mining, then such research will certainly happen elsewhere. This likely does not support the ability for the US to lead in the "revolution in science and understanding that will change humanity" lead by AI.[15]

Question 5. Is new legislation warranted to address copyright or related issues with generative AI?

No new copyright laws or rights are necessary at this time. The flexibility built in to the fair use doctrine is designed to address in the first instance exactly these kinds of new technological questions, and there are a number of cases making their way through the courts. Policymakers should see how they play out before any additional regulatory interventions are made.

Question 8. Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use?

Others will surely write many pages of responses to this question. For our part, we agree with the Library Copyright Alliance Principles for Copyright and Artificial Intelligence that data collection and nonconsumptive uses for the purposes of training AI models will generally be a fair use under US Copyright Law.[16]

Question 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)?

From a doctrinal copyright standpoint, there has historically been no opting in or out of fair use; a core part of the doctrine is that it requires no permission.[17] On the other hand, fair use, when properly construed, does recognize a distinction between uses that are commercial and those that are not. In this vein, it should be noted that the European Union's machine learning exception for "research

[15] Remarks of Senate Majority Leader Chuck Schumer. June 21, 2023. Available at: https://www.democrats.senate.gov/news/press-releases/majority-leader-schumer-delivers-remarks-to-launch-safe-innovation-framework-for-artificial-intelligence-at-csis
[16] Library Copyright Alliance Principles for Artificial Intelligence and Copyright. Available at: https://www.librarycopyrightalliance.org/wp-content/uploads/2023/06/AI-principles.pdf.
[17] *Campbell v. Acuff-Rose Music, Inc.*, 113 S. Ct. 1642 (1994)

organizations and cultural heritage institutions . . . for the purposes of scientific research" is not subject to an opt out, while its commercial exception is.

As against this existing legal framework, the ability for individuals to voice their desire to opt out of AI training is already being developed on a voluntary basis, and that effort should continue but not be required by law.[18] For AI uses that may not be a fair use, such as an AI generating a painting that is substantially similar to a copied artist's work, then appropriate licensing is likely to be necessary.

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?

When considering recordkeeping and other regulatory requirements for AI, it is important to keep in mind that some of the strictest requirements that have been envisioned would be extremely difficult (or in some cases impossible) to do in practice. Such requirements would shape the technical and competitive landscape around AI development, restricting it to the largest and most well-resourced corporations. Indeed, companies like Google and Microsoft already have a large competitive advantage in the field because they have amassed huge stores of data over the years. In order to ensure AI in the public interest, we cannot allow AI research and development to be limited only to these kinds of companies.

At the time of writing, there is an ongoing debate in the European Union regarding whether and to what extent developers of certain AI models should be subject to these kinds of transparency requirements for copyright or other purposes. On the one hand, it has been noted by some in that debate that "[g]reater openness and transparency in the development of AI models can serve the public interest and facilitate better sharing by building trust among creators and users."[19] On the other hand, it has also rightly been noted that the strictest

---

[18] Heikkila, Melissa, *Artists can now opt out of the next version of Stable Diffusion,* December 16, 2022. Available at: https://www.technologyreview.com/2022/12/16/1065247/artists-can-now-opt-out-of-the-next-version-of-stable-diffusion/

[19] *CC and COMMUNIA Statement on Transparency in the EU AI Act*, October 23, 2023. Availab;e at:

level of recordkeeping "is not practical, nor is it necessary for implementing opt-outs and assessing compliance."[20] But importantly, the European Union has already enshrined in law an exception for AI that would not require research and cultural heritage organizations to implement these costly requirements.

Finally, there are no doubt reasons beyond copyright enforcement to want transparency in AI training. But this policy discussion should happen in that broader context, rather than at the Copyright Office.

Question 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?

We agree with the Copyright Office's position that "copyright can protect only material that is the product of human creativity." Thus, the outputs of a Generative AI tool such as Chat GPT, DALL-E or Midjourney are not, in and of themselves, subject to copyright protection. However, when human authors mix their own original creativity with the AI outputs, they are entitled to protection for the expression they have added.[21]

Question 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?

The traditional policy foundations for extending copyright protection generally do not apply in the case of AI-generated material.[22] There is no evidence that copyright law provides necessary incentives for the creation of AI generated

---

https://creativecommons.org/2023/10/23/cc-and-communia-statement-on-transparency-in-the-eu-ai-act/

[20] *Id.*

[21] E.g., US Copyright Office February 21, 2023 Letter regarding Zarya of the Dawn (Registration # VAu001480196). Available at: https://www.copyright.gov/docs/zarya-of-the-dawn.pdf

[22] See, e.g., Brown-Bramble, Dyllan, *A Case For the AI-Driven Public Domain* (December 18, 2020). Available at SSRN: https://ssrn.com/abstract=3807548 or http://dx.doi.org/10.2139/ssrn.3807548

works, and regardless, the constitutional foundations of copyright make clear that its goal is to incentivize *human* authorship.

Question 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?

We agree with the Library Copyright Alliance: "If an AI produces a work that is substantially similar in protected expression to a work that was ingested by the AI," then it is quite likely that "that new work infringes the copyright in the original work."[23]

Question 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?

Similar to the need for transparency on training, there are significant ethical and policy reasons outside of copyright law to be interested in the question of AI-labeling, e.g., use of AI tools for cognitive manipulation.[24] As such, this question must be considered in a broader policy context, and should not be decided exclusively for the purposes of copyright enforcement.

Question 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?

Deepfakes are a serious problem that needs to be solved, but granting a new right of publicity along the lines of existing state laws would privilege famous people over the general public, and come with serious First Amendment concerns. Regulators should work instead towards strong federal privacy protections for everyone. The public deserves the right to exist in online spaces without being

[23] Library Copyright Alliance Principles for Artificial Intelligence and Copyright. Available at: https://www.librarycopyrightalliance.org/wp-content/uploads/2023/06/AI-principles.pdf.
[24] E.g., Farahany, Nita A., *The Battle for Your Brain: Defending the Right to Think Freely in the Age of Neurotechnology*. St. Martin's Press (March 14, 2023).

constantly surveilled, and the right to not have our likeness used in ways that humiliate, harass, or abuse us. Congress has yet to address these significant challenges of our digital age. It would be unfair to hand out rights to celebrities that would allow them to control how their image is collected, processed and used, but not have any protections for the general public.

Questions 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)?

As a general matter, copyright law covers specific works of art, not styles, methods, or genres. That is not to say that, in certain cases, works that appear to be "in the style of" a specific artist may not infringe a valid copyright of that artist under existing law. But extending copyright protection to styles as such would be a grave mistake; it would take far too much unprotectable material out of the public domain, creating significant uncertainty and liability for anyone choosing to create.

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

Many of the concerns voiced in response to generative AI technology have their roots well beyond the scope of copyright law– in privacy, competition, and other similar areas. Creating a new rights regime and a licensing market for AI training and deployment would not solve these pre-existing problems; it would only serve to worsen them, and further entrench the largest companies who already abuse their vast stores of user data for competitive advantage.[25]

[25] Garcia, Nicholas. *Generative AI is Disruptive, But More Copyright Isn't the Answer. (May 11, 2023). Available at:* https://publicknowledge.org/generative-ai-is-disruptive-but-more-copyright-isnt-the-answer/