Rebecca Chambers
Project LEND
University of California, Berkeley
Berkeley Center for Law & Technology

By Electronic Submission

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

Re: Notice of Inquiry (NOI) and request for comments, Artificial Intelligence and Copyright, Docket No. 2023-6

October 30, 2023

Dear Associate Register Wilson,

The undersigned researchers of Project LEND respectfully submit these comments in response to the Copyright Office's Artificial Intelligence Notice of Inquiry, published on August 30, 2023.[1]

Project LEND is a grant-funded project based out of the University of California (UC) whose mission is to investigate the potential for expanded lawful uses of digital materials held by academic and research libraries. The project seeks to analyze all aspects of a digital access program — including user needs, legal frameworks, technical requirements, and collection scope — to design an expanded service or set of services for UC faculty, staff, and students, and, through its partner HathiTrust, to extend these services to over 200 more academic institutions.[2] As part of this mission, Project LEND seeks to clarify how UC faculty, researchers, and students can use artificial intelligence (AI) in conjunction with UC's digital collections to further their academic ambitions. Given our interest in this new technology, we appreciate the Copyright Office's attention to AI.

Introduction

The University of California spans ten campuses, serves 294,309 undergraduate and graduate students, and employs 24,800 faculty members and 49,000 other academics.[3] The University's mission is "to serve society as a center of higher learning, providing long-term societal benefits

---

[1] Artificial Intelligence and Copyright Notice of Inquiry and Request for Comments, 88 Fed. Reg. 167,59942 (Aug. 30, 2023).

[2] For a more complete description of the Project, see Project LEND, UC Libraries, https://libraries.universityofcalifornia.edu/project-lend/ (last visited Sept. 25, 2023). Principal investigators for the project are Rice Majors, the associate university librarian for scholarly resources at UC Davis, Erik Mitchell, the Audrey Geisel University Librarian at UC San Diego, and Günter Waibel, the associate vice provost and executive director of the California Digital Library. Rebecca Chambers is the Project's full-time legal fellow, supervised by Molly Van Houweling, Harold C. Hohback Distinguished Professor of Patent Law and Intellectual Property, University of California, Berkeley School of Law, Co-Director, Berkeley Center for Law & Technology.

[3] The University of California at a Glance (August 2023) https://ucop.edu/institutional-research-academic-planning/_files/uc-facts-at-a-glance.pdf (last visited Oct. 13, 2023).

through transmitting advanced knowledge, discovering new knowledge, and functioning as an active working repository of organized knowledge."[4]

Working together, the libraries of the ten UC campuses and the California Digital Library serve the research, teaching, learning, and clinical care needs of the community.[5] In service to this mission, UC libraries have developed a collection of over 40 million print books, 1.7 million ebooks, and 248,000 collectively licensed electronic journals, a collection surpassed only by the Library of Congress.[6] UC libraries collectively lead the University's Open Access (OA) program and on each campus the libraries make use of the collective resources of UC while also supporting the local needs of campus students and faculty. In 2008, UC was a founding member of HathiTrust, a not-for-profit collaborative of academic and research libraries now preserving more than 18 million digitized items in the HathiTrust Digital Library. HathiTrust offers researchers reading access to the fullest extent allowable by U.S. and international copyright law, as well as text and data mining tools for the entire corpus. In total, UC has digitized over 48 million items.[7]

The librarians and staff who run these libraries acquire relevant materials, assist students and researchers in locating the best sources, curate special exhibits, and advise on the many implications of using works in the collection in research and scholarship. Library staff are instrumental to UC's prodigious production of scholarly works.[8]

Many UC faculty and students are also authors who contribute to the copyright system. Their authorship relies upon extracting and augmenting knowledge from the UC libraries' collections. These students, faculty, and researchers should be able to use all the tools at their disposal to further UC's mission of discovering and transmitting knowledge and the Progress Clause's mandate that copyright "promote the Progress of Science and useful Arts." [9] Scholars create new knowledge by building on existing knowledge, and copyright should not be a hurdle or a hindrance to technologies that facilitate research on a new scale.

It is important to distinguish between the kinds of computational methods researchers use. Text and data mining (TDM) is a broad term for the use of "automated techniques and statistical methods to identify patterns in a given database."[10] For example, TDM algorithms can be built to

---

[4] University of California, Office of the President, *UC's Mission*, https://www.ucop.edu/uc-mission/#:~:text=%22The%20distinctive%20mission%20of%20the,working%20repository%20of%20organized%20knowledge (last visited Oct. 4, 2023).

[5] The California Digital Library is part of the UC Office of the President and aims to take advantage of emerging technologies in order to create one of the world's largest digital research libraries. *See* University of California, California Digital Library, *About CDL*, https://cdlib.org/about/ (last visited Oct. 16, 2023).

[6] University of California, *Budget for Current Operations, 2023-2024*, https://www.ucop.edu/operating-budget/_files/rbudget/2023-24-budget-detail.pdf (last visited Oct. 13, 2023).

[7] University of California, *Budget for Current Operations, 2023-2024*, https://www.ucop.edu/operating-budget/_files/rbudget/2023-24-budget-detail.pdf (last visited Oct. 13, 2023).

[8] UC produces over 8% of all scholarly publishing in the U.S. and performs 10% of all U.S. academic research and development. *See* University of California, *Accountability Report 2019: Research* https://accountability.universityofcalifornia.edu/2019/chapters/chapter-9.html (last visited Oct. 13, 2023).

[9] U.S. Const. art. 1, § 8, cl. 8.

[10] Rachael Samberg, Cody Hennesy, *Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis*, https://escholarship.org/uc/item/55j0h74g (2017).

detect the frequency of words.[11] The Copyright Office has previously held that TDM, with limitations, is likely a fair use.[12] Some, but not all, TDM uses artificial intelligence. AI itself is an umbrella term for a wide variety of systems. Non-generative AI, or traditional AI, performs specific tasks based on pre-programmed rules and parameters. Generative AI technologies, on the other hand, are trained on massive amounts of data, learn connections between the information presented, and can produce new expression.[13] The main difference between non-generative and generative AI models lies in their outputs: non-generative AIs analyze data and establish inferences or patterns, while generative AIs take those predictions to create something new.[14]

AI models can make connections, unpack themes, and spur creativity and innovation at a scale heretofore unknown. For academic and research libraries and their patrons, AI can increase efficiency, enhance research capabilities, and open new arenas of discovery. However, AI technology is only at the beginning of its implementation by scholars; there are still many unknowns about its ultimate ramifications, benefits, and effects upon markets, humans, and creativity. The answers to some of the questions posed by the Copyright Office in this NOI may change as AI implementation becomes more widespread. The questions posed by the Copyright Office affect all these technologies and the outcome of this inquiry will have effects on all uses of AI, including those in academic and research library environments.

How AI Can Benefit Students, Researchers, and Libraries

The Copyright Office has the chance to support the Constitution's mandate that copyright law "promote the Progress of Science and useful Arts"[15] by recognizing the rights nonprofit educational institutions, including academic and research libraries, have to use the tools that will help drive important research forward.

Academic and research libraries, such as the UC libraries, could train AI models on the library collections for multiple purposes.[16] For example, the collections could be used to train a generative AI model to write a short abstract or summary of every book and article in the UC collection, which would create a more searchable index of works and increase the discoverability of research materials immensely.[17] The UC corpus could also be used to train a non-generative AI model to pull metadata about UC's collections so librarians can more easily identify gaps in literature,

---

[11] *See Google Books Ngram Viewer*, https://books.google.com/ngrams/info (last visited Oct. 13, 2023).
[12] *See* United States Copyright Office, *Section 1201 Rulemaking: Eighth Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention - Recommendation of the Register of Copyrights* (Oct. 2021).
[13] Forbes, Bernard Marr, *The Difference Between Generative AI and Traditional AI: An Easy Explanation For Anyone*, https://www.forbes.com/sites/bernardmarr/2023/07/24/the-difference-between-generative-ai-and-traditional-ai-an-easy-explanation-for-anyone/?sh=7826a7a3508a (last visited Sept. 29, 2023).
[14] There are multiple ways in which a generative AI system can be built. *See* Katherine Lee, A. Feder Cooper, James Grimmelman, *Talking 'Bout AI Generation: Copyright and the Generative AI Supply Chain* (2023).
[15] U.S. Const. art. 1, § 8, cl. 8.
[16] UC patrons using AI tools on the UC library corpus would have lawful access to the materials, which is more than can be said about many of the commercial AI tools out there.
[17] Using copyrighted works to create an index that increases discoverability was found to be a fair use. *See Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014), *Authors Guild v. Google*, 721 F.3d 132 (2d Cir. 2015).
Additionally, a summary is unlikely to be a derivative work because an author's derivative rights do not include an exclusive right to supply information about her works. *See Google LLC v. Oracle America, Inc.*, 593 U.S. _ (2021).

address any lack of representation by certain marginalized groups, and more accurately curate their collections to meet the needs of students, researchers, and scholars.

UC scholars across campuses and disciplines are already using AI technology to further their scholarship. Some scholars are using AI to replicate the improvisational jazz stylings of Toots Thieleman,[18] while others are using it to understand the impact that natural disasters, like Hurricane Katrina, have on urban planning and development.[19] Some are using AI to create summaries of articles to determine whether a given article meets their research needs and to garner research and book suggestions.[20] UC researchers could further employ AI for a large variety of research and writing projects. Scholars could build and train their own AI model on assets that have been lawfully obtained, which may include some obtained from the UC libraries. For example, a scholar could train a generative AI model to create an annotated bibliography, write an abstract, and generate ideas for scholarly articles. AI algorithms can also be used for non-generative purposes such as identification, data gathering, and connection drawing, making AI an essential research tool. More examples of how UC faculty are already using AI in their scholarship can be found in Appendix A.[21]

This comment makes four recommendations, laid out below, each focusing on a question from the NOI. To address the Copyright Office's questions in as tangible and succinct a manner as possible, the recommendations will answer each question through the lens of a use case scenario that represents one possible way researchers may make use of AI. In this use case, a digital humanities scholar wants to study the Pulitzer Prize for Fiction[22] and use a non-generative AI model to help her discern the underlying themes, moods, and attitudes of the winning books. As the Pulitzer Prize for Fiction is the United States of America's biggest literary prize, a study of what themes and subjects are rewarded can help the scholar understand what story the United States wants to tell about itself (and what types of stories have historically been excluded). In order to do this, the scholar needs to train an AI model on the Pulitzer Prize for Fiction corpus, which currently includes 103 books published between 1918 and 2023.[23] Nine of these books are in the public domain,[24] but the rest remain presumably under copyright. Some of these books, especially the more recent publications, are commercial successes with ebook versions readily available. Others, like the 1936 winner *Honey in the Horn*, may be found only in physical form and are often not available as

---

[18] Katherine Connor, Josh Baxt, *Creating Computing: Teaching Computers to Think Like a Human*, https://today.ucsd.edu/story/jazz (last visited Oct. 17, 2023).
[19] Justin Hosbey, Research UC Berkeley, https://vcresearch.berkeley.edu/faculty/justin-hosbey (last visited Oct. 4, 2023), see also J.T. Roane and Justin Hosbey, Mapping Black Ecologies, Current Research in Digital humanities (Aug. 23, 2019) https://crdh.rrchnm.org/essays/v02-05-mapping-black-ecologies/ (last visited Oct. 4, 2023).
[20] Project LEND, Preliminary Results from Focus Groups (2023).
[21] For more examples of AI model-reliant research, please see the NOI response being submitted by the Authors Alliance organization which incorporates perspectives from researchers.
[22] The Pulitzer Prize for Novel was awarded from 1918-1947 and the Pulitzer Prize for Fiction has been awarded since 1948. The scenario lumps these prizes together.
[23] Scholars already use statistical modeling of narrative based on sentiment analysis and other modes of topical analysis. AI technologies are a further evolution in how these established methodologies can be applied. *See* Richard Jean So, *Redlining Culture: A Data History of Racial Inequality and Postwar Fiction* (2020).
[24] Based upon the assumption that anything published in the United States before 1928 is now in the public domain. No Prize was awarded in 1920.

licensed ebooks, although some libraries may be able to provide a digitized version.[25] Therefore, in order to create her corpus, the scholar will have to engage in a variety of reproductive activities, from scanning to scraping, on the relevant copyrighted works to transform them all into a form that can be fed to the AI model. Then the AI can help the scholar make the connections and identify what binds the books that have won the Pulitzer Prize for Fiction together. All these actions, including reproducing the works, training the AI model, and using the results in a scholarly article should be fair uses.[26]

To ensure that important research and scholarship of the type exemplified by this hypothetical flourishes, Project LEND proposes four recommendations, summarized below:

1. The Copyright Office should recognize that the unauthorized reproduction and use of copyrighted works to train AI models for nonprofit educational and research purposes can be a fair use.
2. The source of funding for these nonprofit educational and research purposes should have little bearing on the fair use analysis, even if the funder is a for-profit AI developer.
3. Further, in order for scholars, students, and researchers to make full use of the copyrighted works provided by their institutions and to avoid gaps in their research results, the Copyright Office should not adopt an opt in/out regime or a voluntary collective licensing scheme.
4. Finally, the Copyright Office should take into consideration how AI reproduces the biases found in its training data and acknowledge that training AI on copyrighted works may be one way to reduce such bias.

All of these recommendations are designed to support the important research and scholarship done by UC researchers, students, and scholars and to further the Copyright Office's mandate of "promot[ing] the Progress of science and the Useful Arts."[27] The understandings furthered by scholars delving into diaries and other writings to understand the correlations between culture and stress[28] or to understand the ways in which language can reveal a lot about historical notions of religion[29] are extremely important work that should be fostered.

Project LEND does not weigh in on the potentially infringing nature of generative AI outputs as those cases will be very fact dependent and hard to predict in advance.[30] Courts will need to decide

---

[25] The University of California, Berkeley provides digital access to *Honey in the Horn* through Project Muse, a platform UC Berkeley subscribes to. Each chapter has to be downloaded separately. Some books may be available for computation through corpora like HathiTrust.

[26] Referred hereafter to as the "Pulitzer AI Project."

[27] U.S. Const. art. 1, § 8, cl. 8.

[28] UC Irvine, Faculty Profile System, Benjamin Colby, https://faculty.uci.edu/profile/?facultyId=2607 (last visited Oct. 17, 2023).

[29] Bryan Wagner, Research UC Berkeley, https://vcresearch.berkeley.edu/faculty/bryan-wagner (last visited Oct. 4, 2023).

[30] Infringement could occur, as Matthew Sag wrote, in limited circumstances, such as the Snoopy problem: an AI is trained on many different photos of Snoopy (which are labeled as such) and when the input is "Snoopy," the output will likely be "Snoopy." Matt Sag, *Copyright Safety for Generative AI* (2023).

on a case-by-case basis if an output is substantially similar to an input.[31] Project LEND takes no position as to whether generative AI output infringe or are fair use.

Recommendation 1: In response to Question 8 of the Notice of Inquiry, the Copyright Office should recognize the unauthorized reproduction and use of copyrighted works to train AI models for nonprofit educational or research purposes can be fair use.

Copyright law protects any "original works of authorship fixed in a tangible medium of expression" and provides the copyright holder with only a certain scope of protection over their work.[32] Copyright owners can control the reproduction, distribution, creation of derivative works, public performance, and public display of the *expression* contained within their copyrighted work, but not the proliferation of the underlying ideas.[33] Copyright holders cannot stop follow-on users from extracting the unprotectable ideas and facts from the protected work. The Supreme Court in *Baker v. Selden* reinforced this idea-expression distinction when it held that a copyright did not give exclusive rights to the bookkeeping system described by the copyrighted work; in other words, while the book was copyrightable, the method it described was not, as "ideas, procedures, processes, methods of operation, concepts, principles, and discoveries" are excluded from copyrightable subject matter.[34]

Extracting the unprotectable elements from a copyrighted work often requires unauthorized actions that do nominally fall within copyright's protections, especially when using a mechanical extraction process like crawling. However, courts have held, in a variety of contexts, that it is a fair use to extract the facts and ideas from copyrighted works, even when it requires unauthorized reproduction.[35] The use of AI models to extract and draw connections between these unprotectable facts and ideas is what legal scholar Professor Matthew Sag has termed a "non-expressive use."[36] However, just claiming a work is being used only for its non-expressive elements is not enough to be deemed a fair use; as Sag puts it, a non-expressive use is not a free-standing defense to infringement.[37] The use must also pass the fair use analysis, a flexible four factor test that weighs the benefits of the use with the harms to the original work.

---

[31] *See* Katherine Lee, A. Feder Cooper, James Grimmelman, *Talking 'Bout AI Generation: Copyright and the Generative AI Supply Chain* (2023).

[32] 17 U.S.C. § 102(a).

[33] 17 U.S.C. § 106(1)-(5), 17 U.S.C. § 102(b).

[34] *Baker v. Selden*, 101 U.S. 99 (1879); 17 U.S.C. § 102(b).

[35] *Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1992) (reverse engineering for interoperability is fair use), *Kelly v. Arriba Soft Corp.*, 336 F.3d 811 (9th Cir. 2003) (the creation and use of thumbnail images for a search engine is fair use), *Field v. Google Inc.*, 412 F.Supp. 2d 1106 (D. Nev. 2006) (storing a website in cache is fair use), *Perfect 10 Inc. v. Amazon.com, Inc.*, 508 F.3d 1146 (9th Cir. 2007) (the creation and use of thumbnail images for a search engine is a fair use), *A.V. ex rel. Vanderhye v. iParadigms, L.L.C.*, 562 F.3d 630 (4th Cir. 2009) (reproducing copyrighted works for purposes of identifying plagiarism is fair use), *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014) (a full text searchable database is fair use), *Authors Guild v. Google, Inc.*, 721 F.3d 132 (2d Cir. 2015) (a search function which shows snippets around the searched-for term is fair use), *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. _, 143 S. Ct. 1258 (2023) ("provid[ing] unavailable information about the original" is a fair use.

[36] Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 Northwestern University Law Review, (2009). Non-consumptive use is also often used; however, Professor Sag prefers non-expressive as non-consumptive was already a term of art in water rights literature, was used in the failed *Google Books* settlement, and points readers more firmly towards the idea-expression dichotomy.

[37] Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 Northwestern University Law Review (2009).

The remainder of this section will apply the fair use analysis to the Pulitzer AI Project as a hypothetical example of the types of activities Project LEND aims to facilitate. As this analysis demonstrates, nonprofit educational uses of AI should be considered fair uses.

Preamble. The fair use provision of the Copyright Act privileges certain uses as laid out in the preamble, including those for "teaching, scholarship [and] research."[38] Part of Project LEND's mission is to enable as many uses of UC libraries' digital materials for teaching, scholarship, and research as possible, and that includes using UC library resources to train AI models for the purposes of extracting and using the non-expressive elements of the included copyrighted works for use in research and scholarship, aligning with the preamble.

First Factor: Purpose & Character of the Use. The first factor of the fair use analysis looks at whether the secondary use has a different purpose and character than the original use, including whether it was transformative and commercial or noncommercial.[39] Training an AI model on copyrighted works is a transformative use.[40] A transformative use "adds something new, with a further purpose or different character, altering the first with new expression, meaning or message."[41] Reproducing a copyrighted work to extract the underlying unprotectable ideas, themes, and is just such a "further purpose." Courts have routinely found that uses that extract unprotectable ideas, themes, and facts serve valid transformative purposes. As the Supreme Court recently said in its *Andy Warhol Foundation v. Goldsmith* opinion, using a copyrighted work for the purpose of "provid[ing] otherwise unavailable information about the original [is] a valid fair use purpose under the first factor."[42] In *Authors Guild v. HathiTrust*, HathiTrust's full-text searchable database served an entirely different purpose than the original works: that of identifying books which contained the desired search term.[43] *Authors Guild v. Google* expanded on *HathiTrust* and held that the snippets (2-3 lines of text surrounding the searched-for term) provided important context and transformed the book's text into data that could further important research goals, a "quintessentially transformative use."[44] Using AI to generate valuable information about a set of works has an entirely different purpose from that of the original author, whose purpose was to express ideas in a particular manner.[45] The purpose of the copyrighted works which won the Pulitzer Prize for Fiction, for example, was to be read for their expressive content. Training an AI on those copyrighted works to identify the underlying unprotectable elements of ideas, themes, and facts is a quintessentially transformative use.

The first factor also considers the commerciality of the follow-on use and privileges nonprofit and educational uses. The statute itself instructs courts to look at "whether such use is of a commercial

---

[38] 17 U.S.C. § 107.
[39] 17 U.S.C. § 107(1).
[40] Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 Northwestern University Law Review (2009).
[41] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 579 (1994).
[42] Responding to the Notice of Inquiries' Question 8.1. *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. _, 143 S. Ct. 1258 (2023). (pg. 32).
[43] *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014).
[44] *Authors Guild v. Google*, 721 F.3d 132 (2d Cir. 2015).
[45] Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 66 Journal of the Copyright Society of the USA 291 (2019).

nature or is for *nonprofit educational purposes*."[46] Nonprofit educational purposes are thus favored; however, a commercial nature has not barred many non-expressive uses from being deemed a fair use.[47] For example, in *Perfect 10 v. Amazon*, Google's purpose in using the thumbnails of copyrighted works in its image search was a commercial endeavor. However, the court found that "the significantly transformative nature of Google's search engine, particularly in light of its public benefit, outweigh[ed] Google's superseding and commercial uses of the thumbnail."[48] If a non-expressive use with a commercial purpose can be a fair use, then a non-expressive use in a noncommercial, educational setting with a public benefit, such as the Pulitzer AI Project, is likely privileged under the first factor and helps to fulfill copyright law's mandate to "promote the Progress of Science and the useful Arts."[49] This factor weighs in favor of fair use due to the transformative and noncommercial nature of the use.

Second Factor: Nature of the Copyrighted Works. The second fair use factor looks at the nature of the copyrighted work.[50] The more creative and expressive a work is, the more protection it is afforded, while the more factual, the less. The second factor is not very influential, however, when dealing with transformative works, especially when those works target unprotected elements of the works they copy.[51] For example, in *A.V. ex rel Vanderhye v. iParadigms*, the court found the second factor weighed in favor of fair use because the use of the works was unrelated to any of their creative components.[52] Courts have also considered how difficult it is to cleanly extract unprotectable elements from a copyrighted work and held that the second factor is neutral or leans toward fair use when there is something about the work that makes extracting those unprotectable elements hard.[53] Additionally, the second factor also looks at whether the copyrighted works have been published. A use is less likely to be a fair use if it uses unpublished materials.[54] In the case of the Pulitzer AI Project, all the works have been published and are only being used for their non-expressive components, which has been established as a likely transformative use. Therefore, this factor is neutral or weighs slightly in favor of fair use.

Third Factor: Amount and Substantiality. The third fair use factor looks at the amount and substantiality of the portion used in relation to the whole of the copyrighted work, specifically whether the amount of copying was reasonable in relation to the purpose.[55] Often, the more

---

[46] 17 U.S.C. § 107(1) (emphasis added).

[47] 17 U.S.C. § 107.

[48] *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146 (9th Cir. 2007).

[49] U.S. Const. art. 1, § 8, cl. 8.

[50] 17 U.S.C. § 107(2).

[51] See, e.g., *Field v. Google, Inc.*, 412 F.Supp. 2d 1106 (D. Nev. 2006).

[52] *A.V. ex rel. Vanderhye v. iParadigms, L.L.C.*, 562 F.3d 630 (4th Cir. 2009), *see also Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605 (2d Cir. 2006) ("using creative, artistic images as historical artifacts, aiming to extract documentary significance did not weigh in the plaintiff's favor"), *Bouchat v. Balt. Ravens Ltd. P'ship.*, 737 F.3d 932 (4th Cir. 2013) (if the disputed use of the copyrighted work is not related to its mode of expression but rather to its historical facts, then the creative nature of the work matters much less than it would otherwise").

[53] *See Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1992) (second factor favored the defendant where video game programs contained unprotected aspects that could not be examined without copying"), *Sony Computer Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596 (9th Cir. 2000) (second factor strongly favored the defendant where copying of the work was necessary to access its unprotected elements), *Atari Games Corp. v. Nintendo of Am. Inc.*, 975 F. 2d 832 (Fed. Cir. 1992) (when the nature of the work requires intermediate copying to understand the ideas and processes in a copyrighted work, that nature supports a fair use for intermediate copying).

[54] *Harper & Row Publishers, Inc. v. Nation Enterprises*, 471 US 539 (1985).

[55] 17 U.S.C. § 107(3).

transformative the use, the less of a role the third factor plays.[56] However, transformative use is not a free pass to take as much as one pleases; the appropriation must still be reasonable.[57] Courts have readily identified those cases where the output contained too much expression. For example, the *Harry Potter Lexicon* and the *Seinfeld Quiz Book* both claimed to be taking facts from the copyrighted works they were based on. The courts clearly saw that these uses fell short of their intended transformative purpose and contained too much of the copyrighted expression.[58] By contrast, courts have held that copying an entire work is *not* too much where the output is only unprotected information or a very small amount of expression.[59] For the transformative purpose of the Pulitzer AI Project, each work needs to be copied in its entirety, but because the Pulitzer AI Project appropriates and generates only the underlying facts and data, it has little chance of appropriating too much of the creative expression in its output.[60] This factor is neutral in the fair use analysis.

Fourth Factor: Effect on the Potential Market or Value of the Copyrighted Work. The fourth and final factor in the fair use analysis is the secondary use's effect on the potential market or value of the original work.[61] This factor, along with the "transformative use" inquiry of the first factor are the two most influential factors.[62] Market harm is limited to "whether the use provides a meaningful substitute."[63] The more transformative a use, the less likely it will be able to act as a market substitute.[64] A non-expressive use like the Pulitzer AI Project does not affect the market for the original expression as it reveals little of the underlying copyrighted works used to generate the unprotectable insights and is hence unable to function as a market substitute. Additionally, these works will have been legally acquired by the UC libraries through a lawful sale or license. The market for the book, periodical, or journal has not been usurped; rather, it has worked as intended.

---

[56] *Authors Guild Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014), *Authors Guild, Inc. v. Google*, 721 F.3d 132 (2nd Cir. 2015) (cases which held that taking the entirety of a copyrighted work was justified in light of the transformative use).

[57] Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 66 Journal of the Copyright Society of the USA, 291 (2019).

[58] *Warner Bros. Entertainment, Inc. v. RDR Books*, 575 F.Supp.2d 513 (SDNY 2008), *Castle Rock Entertainment Inc. v. Carol Publishing Group*, 150 F.3d 132 (2d Cir. 1998), *see also Associated Press v. Meltwater Holdings* (S.D.N.Y 2013) (held that Meltwater's use of every AP lede, which are meant to convey the heart of the story, was too much copying, despite that Meltwater was only seeking to copy news and facts.)

[59] *Authors Guild Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014), *Authors Guild, Inc. v. Google*, 721 F.3d 132 (2nd Cir. 2015).

[60] Matthew Sage, *The New Legal Landscape for Text Mining and Machine Learning*, 66 Journal of the Copyright Society of the USA, 291 (2019).

[61] 17 U.S.C. § 107(4).

[62] Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Options Updated 1978–2019*, 10 NYU Journal of Intellectual Property and Entertainment Law 1, 22-23 (2020).

[63] Benjamin Sobel, *Artificial intelligence's Fair Use Crisis*, 41 Columbia Journal of Law & the Arts 45 (2017) citing *Authors Guild v. Google*'s discussion of the fourth factor, which waffled between examining the fourth factor through the lens of "whether a use *replicates protected expression in a manner that* provides a meaningful substitute for the original" or "whether the use provides a meaningful substitute for the original *with no mention of whether or not the use is expressive*" (emphasis added.)

[64] Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 66 Journal of the Copyright Society of the USA, 291 (2019).

Some argue that markets already exist for the purposes of licensing works to train AI models.[65] However, the mere existence of a market for licensing works for AI shouldn't automatically make an unlicensed use unfair.[66] Courts have limited their consideration of market effects to those that interfere with a cognizable copyright interest.[67] That cognizable copyright interest only extends as far as the protectable expression embodied in the copyrighted work. Therefore, a licensing market for the non-expressive elements of a copyrighted work is not within the cognizable copyright interests that courts recognize. Even if some users acquire licenses to make non-expressive uses of copyrighted works, every potential fair user shouldn't be forced to, especially when they are in a noncommercial, educational setting like an academic or research library.[68]

Ultimately, it is fair use for the scholar to train her Pulitzer AI Project on the 94 copyrighted works at issue to extract the unprotectable elements, including the underlying themes and facts that distinguish these books. The scholar transforms the works into unprotectable data, uses a reasonable amount in relation to her transformative purpose, and has not negatively impacted the market for the expressive works. Where a researcher copies entire works in a nonprofit educational setting for purposes of extracting and studying the unprotected ideas, that use should be recognized as fair.

The European Union has already passed the Copyright in the Digital Single Market Directive which requires members states to adopt a limitation or exclusion "for reproductions and extractions made by research organizations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access,"[69] and to "allow cultural heritage institutions to make copies of any works or other subject matter that are permanently in their collections, in any format or medium, for purposes of preservation of such works or other subject matter and to the extent necessary for such preservation."[70] These exceptions are not overridable by contract.[71] By holding that the reproduction and use of copyrighted works for the training of AI is fair use, the Copyright Office would be aligning itself with established precedent in the EU and foster research and scholarship the world over. The EU adopted this measure to decrease legal uncertainty around TDM and research, foster private-public partnerships, and achieve the goal of a "Digital Single Market: the free movement of goods, persons, services, and capital, where individuals and businesses can

---

[65] Benjamin Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Columbia Journal of Law & the Arts 45 (2017).

[66] Mark Lemley, *Fair Learning*, 99 Texas Law Review 4 (2020).

[67] Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 66 Journal of the Copyright Society of the USA, 291 (2019).

[68] *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913 (2d Cir. 1994) (any use may be licensed but not all markets are relevant.Limiting the fourth factor to traditional, reasonable, and likely-to-be exploited markets, the fourth factor avoids the problem of circularity.), *Google LLC v. Oracle America, Inc.*, 593 US _, 33 (2021) (the Supreme Court did not hold the failure to come to a licensing agreement with Oracle against Google).

[69] Directive 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, art. 3, 2019).

[70] Directive 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, art. 6, 2019).

[71] Directive 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, art. 7, 2019).

seamlessly access and exercise online activities under conditions of fair competition, and a high level of consumer protection."[72]

Recommendation 2: In response to Question 8.3, it should make no difference in the fair use analysis if the funding for these nonprofit educational or research uses is provided by for-profit developers of AI systems.

Receiving funding from a for-profit entity for nonprofit educational research should not change the fair use analysis. As mentioned above, when copying is done to further a non-expressive use that is unlikely to act as a market substitute, commerciality plays much less of a role than it does in other contexts. The Supreme Court in *Google v. Oracle* held that Google's copying to increase interoperability, which had nothing to do with the authors' actual expression, was a fair use.[73] The Supreme Court also pointed to considering "the public benefit the copying will likely produce."[74] The market harm that copyright holders may suffer from a loss of licensing for training AI purposes, which as mentioned above is not within the scope of the cognizable market recognized by copyright, does not outweigh the public benefit of using AI to aid research and development in a noncommercial educational environment.

Some researchers, like the hypothetical Pulitzer AI Project researcher, may want to build their own limited corpuses from their library's existing digital resources, while other researchers, like those who want to create more advanced AI models, may need access to much wider swaths of digital materials. Universities should be able to partner with for-profit enterprises to digitize their print collections and increase access for those researchers.[75] Libraries that lack the resources to do this digitization independently should not be penalized for working with for-profit partners. Nonprofit research has a long history of help from for-profit enterprises, which have resulted in extremely beneficial programs. Since the University of California partnered with the Google Library Project back in 2006, 4.6 million volumes from the UC collections have been digitized and deposited into the HathiTrust Digital Library, 4.4 million of which were digitized by Google.[76] Because of this huge collection, when HathiTrust opened its Emergency Temporary Access Service during the early days of the COVID-19 pandemic, UC students, faculty, and researchers were able to access UC library materials that would have been otherwise completely unavailable.[77] University

[72] Bottis, M., Papadopoulos, M., Zampakolas, C., & Ganatsiou, P. (2019). Text and Data Mining in Directive 2019/790/EU Enhancing WebHarvesting and Web-Archiving in Libraries and Archives. Open Journal of Philosophy, 9, 369-395.https://doi.org/10.4236/ojpp.2019.93024.

[73] *Google LLC v. Oracle America, Inc.*, 593 US _ (2021), *see also* Mark Lemley, *Fair Learning*, 99 Texas Law Review 4 (2020).

[74] *Google LLC v. Oracle America, Inc.*, 593 US_ (2021). (PAGE #).

[75] This comment is not addressing those cases where a university researcher uses totally commercially developed corpuses because it is outside the scope of Project LEND, which is focused on how best to increase the access to and use of the UC libraries' materials.

[76] UCnet, *UC Libraries Partner with Google to Digitize Books* (Aug. 9, 2006), https://ucnet.universityofcalifornia.edu/news/2020/05/from-the-archive-uc-libraries-partner-with-google-to-digitize-books.html#:~:text=The%20following%20press%20release%20was,books%20from%20the%20libraries'%20collect ions (last visited Sept. 26, 2023), California Digital Library, *Where to Find Our Books*, https://cdlib.org/services/pad/massdig/where-to-find-our-books/ (last visited Oct. 13, 2023).

[77] HathiTrust's Emergency Temporary Access Service provided digital access to UC faculty, students, and staff to the copyrighted materials that the member institution had a copy of in their physical collections. *See* UC Santa Cruz, University Library, *HathiTrust Provides Emergency Temporary Access to Copyrighted Books*, https://library.ucsc.edu/ETAS (last visited Sept. 26, 2023).

partnerships with for-profit companies have produced significant benefits for students and researchers and should be able to continue without hindrance or negative effects on the fair use analysis.

Recommendation 3: In response to Questions 9 and 10, the Copyright Office should not adopt an opt in or out regime or a voluntary licensing scheme.

By establishing an opt in/out regime or a voluntary licensing scheme, the Copyright Office would reduce the amount of available content upon which to build the "progress of the Useful Arts and sciences" and would be artificially creating a market that could harm the fair use analysis, over-enforce copyright rights, and intrude into an area in which no additional enforcement is needed.[78] An opt in/out regime or a voluntary licensing scheme could create a *de facto* new copyright right for the unprotectable elements of a copyrighted work and enable copyright holders to overreach the protections the Copyright Act has granted them and stop uses that are unequivocally fair and beneficial.[79] Any administrative system of an opt in or out regime or voluntary licensing scheme would be extremely hard to implement and hard to change as more is learned about AI and copyright.

Academic and research libraries already face an erosion of their fair use rights by the terms contained in their contracts with various publishers and database holders. Any opt in/out regime or voluntary licensing scheme could exacerbate this effect and have the added consequence of pricing out those who cannot afford the licensing fees, halting many uses, including research and scholarship. As the Association of Research Libraries has noted, there has been a "disturbing trend" toward licenses for digital content that include "terms that prohibit certain uses that would otherwise be lawful under the US Copyright Act and related regulations."[80] For example, the licenses may place general restrictions on fair use or more specific restrictions on certain uses like text and data mining, requiring the libraries to pay additional, excessively high fees if they want to engage in such activities, even though the Copyright Office itself has held that text and data mining is a fair use.[81] Implementing an opt in/out regime or a voluntary licensing scheme will only strengthen the licensors' ability to contract out certain uses and erode libraries' ability to exercise their lawful rights.

As discussed in more detail below, AI systems can imbed the bias of their training data into their models. An opt in/out regime or a voluntary licensing scheme might increase bias because the less comprehensive the database is, the more biased it is likely to be. In contrast, including all copyrighted materials in training datasets may create a more comprehensive dataset and likely decrease how biased the AI might be.

---

[78] U.S. Const. art 1, §. 8, cl. 8.

[79] Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 66 Journal of the Copyright Society of the USA, 291 (2019).

[80] Association of Research Libraries, Katherine Klosek, *Copyright and Contracts: Issues & Strategies*, 3 July 22, 2022.

[81] United States Copyright Office, Section 1201 Rulemaking: 8th Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention, Recommendation of the Register of Copyrights, 105-124 (Oct. 2021).

Recommendation 4: In response to Question 34, the Copyright Office should consider how AI reproduces biases found in its training data and acknowledge that training AI models on copyrighted works may reduce bias.

There is plentiful evidence that AI can replicate the biases of its training data.[82] As Charles Babbage said back in 1864, "On two occasions I have been asked–'Pray Mr. Babbage, if you put into the machine the wrong figures, … will the right answers come out?' … I am not able to rightly apprehend the kind of confusion of ideas that could provoke such a question," or more succinctly "garbage in, garbage out."[83] All copyrighted material encodes bias. One way to reduce bias is to train AI systems on the widest array of works possible. To bring this back to the Pulitzer AI Project, if the researcher can only use the works in the public domain, the ones from 1918 to 1927, she would get a very skewed picture of what the Prize means. Today the Pulitzer Prize for Fiction is awarded "for distinguished fiction by an American author, preferably dealing with American life."[84] However, between 1918 and 1927, the prize was awarded to "the American novel published during the year which shall best present the *wholesome* atmosphere of American Life and the highest standard of American *manners* and *manhood*."[85] The books awarded in this time period reflect a certain viewpoint on gender, race, the United States of America, and what it means to be an American. However, if the research can train the AI on every book that's won the Pulitzer Prize, she will get a much more accurate and nuanced picture.

As this example illustrates, the inclusion of copyrighted works in training data may reduce the bias of an AI.[86] The larger the dataset and the more varied its inputs, the less biased an AI might be. If the creators of these models cannot use copyrighted works or are required to pay licensing fees for the huge amount of content they require to train the model, they may turn to biased, low-friction data (older works in the public domain, for example), which will produce outputs that replicate the biases of that data.

Conclusion

A flourishing research and scholarly environment depends on the ability to build from pre-existing knowledge. AI has the potential to greatly aid and further developments in the arts, sciences, and more. The Copyright Office has the opportunity to recognize the important role AI could play in nonprofit educational institutions and support them by prioritizing these users and uses when

---

[82] Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 Wash. L. Rev. 579 (2018), Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 Proceedings of Mach. Learning Res. 1-15 (2018), Tolga Bolukbasi et al., *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* (2016).

[83] Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 Wash. L. Rev. 579, 585 fn. 25 (2018), quoting Charles Babbage, *Passages from the Life of a Philosopher* 67 (1864).

[84] The Pulitzer Prizes, Books Submission FAQ, Guidelines and Requirements, https://www.pulitzer.org/page/books-submission-faq-guidelines-and-requirements (last visited Sept. 26, 2023).

[85] Rebecca Chambers, Choosing America's Story: How the Pulitzer Prize for Fiction has Defined the Way Americans Think About America and Themselves (April 1, 2017) (unpublished B.A. thesis, Bucknell University) (citing W.J. Stuckey, The Pulitzer Prize Novels: A Critical backward Look 7, 2d ed. 1981), https://digitalcommons.bucknell.edu/cgi/viewcontent.cgi?article=1417&context=honors_theses (last visited Sept. 26, 2023) (emphasis added).

[86] Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, W. Law Rev. 579 (2018).

creating a framework for governing AI. The training of AI models on copyrighted works and the use of the outputs in scholarship should be a fair use. Unnecessary restrictions, including suppression of private-public partnerships, or the establishment of opt in or out regimes or voluntary licensing schemes will hinder UC researchers and students from contributing important work.

We would be pleased to provide additional information on the above matters or to elaborate on aspects that would be of assistance to the Copyright Office's inquiry. We can be contacted at bchambers@berkeley.edu.

Respectfully,

Rebecca Chambers, with and on behalf of Project LEND.

Project LEND Signatories[87]

Rebecca Chambers,
Berkeley Center for Law & Technology
UC Berkeley School of Law
Project LEND Legal Fellow

Molly Van Houweling
Harold C. Hohbach Distinguished Professor of Patent Law and Intellectual Property
Co-Director, Berkeley Center for Law & Technology
UC Berkeley School of Law
Project LEND Advisor

Rice Majors
Associate University Librarian, Scholarly Resources
UC Davis
Project LEND Principal Investigator

Erik Mitchell
Audrey Geisel University Librarian
UC San Diego
Project LEND Co-Principal Investigator

Rachael G. Samberg
Scholarly Communication Officer & Program Director
UC Berkeley Library
Project LEND Legal Working Group

Katie Fortney
Copyright Policy & Education Officer
California Digital Library, UC Office of the President

---

[87] Institutional affiliation for identification purposes only.

Project LEND Legal Working Group

Roger Smith
Associate University Librarian Scholarly Resources and Services
UC San Diego
Project LEND Collections Working Group

Günter Waibel
Associate Vice Provost & Executive Director
California Digital Library, UC Office of the President
Project LEND Co-Principal Investigator

Peter Brantley
Director Online Strategy, Library
UC Davis
Project LEND Technical Working Group

Appendix A

Many UC faculty members employ some sort of AI to assist with their research.[88]

Example 1: UC Berkeley Professor Bryan Wagner,[89] Dept. of English, UC Berkeley, trained WordSeer, a user interface created by Adita Muralidharan that allows users to pose questions to a particular set of texts. Professor Wagner created a set of digitized American slave narratives. He then posed two questions, "what does God do?" and "how was God described?" to better interpret American slaves' relationships with God. WordSeer returned thousands of results ranked by frequency. Overall, it turns out, slaves felt God was merciful and benevolent.[90] These connections could not have been drawn as quickly without the help of the WordSeer tool.

Example 2: UC Berkeley Professor Timothy Tangherlini, Dept. of Scandinavian, UC Berkeley, uses computational methods to study topic change and geographic distribution in large folklore corpora.[91] He has also more recently used AI to help understand how conspiracy theories spread as seen in a recent paper, *A Real-Time Platform for Contextualized Conspiracy Theory Analysis*.[92]

Example 3: UC Berkeley Professor Justin Hosbey, Assistant Professor, Dept. of City and Regional Planning, UC Berkeley, uses "research methods from the digital and spatial humanities to understand and visualize how the post-Katrina privatization of neighborhood schools in low-income and working-class Black communities has fractured, but not broken, space and placemaking in Black New Orleans."[93]

Example 4: UC San Diego Professor Shlomo Dubnov is working on Project REACH to train an AI on archival recordings of the legendary Belgian jazz musician Toots Thielemans to generate new jazz improvisations.[94]

Example 5: UC Berkeley professor Alison Gopnik compares how children and AI perform in the same tasks and environments to understand the abilities of existing AI systems and create new ones. Young children's learning may be an important model for AI.[95]

---

[88] Find UC Faculty, https://vcresearch.berkeley.edu/faculty-expertise (last visited Oct. 4, 2023).
[89] Bryan Wagner, Research UC Berkeley, https://vcresearch.berkeley.edu/faculty/bryan-wagner (last visited Oct. 4, 2023).
[90] Nicole Freeling, *Trimming Time in the Stacks*, (Dec. 20, 2011) https://research.universityofcalifornia.edu/stories/2011/12/wordseer.html (last visited Oct. 4, 2023).
[91] Timothy Tangherlini, https://scandinavian.berkeley.edu/people/timothy-tangherlini/ (last visited Oct. 4, 2023).
[92] Timothy Tangherlini, *A Real-Time Platform for Contextualized Conspiracy Theory Analysis*, 2021 International Conference on Data Mining Workshops (2021) https://www.academia.edu/86741615/A_real_time_platform_for_contextualized_conspiracy_theory_analysis.
[93] Justin Hosbey, Research UC Berkeley, https://vcresearch.berkeley.edu/faculty/justin-hosbey (last visited Oct. 4, 2023), *see also* J.T. Roane and Justin Hosbey, *Mapping Black Ecologies*, Current Research in Digital humanities (Aug. 23, 2019) https://crdh.rrchnm.org/essays/v02-05-mapping-black-ecologies/ (last visited Oct. 4, 2023).
[94] Katherine Connor, Josh Baxt, *Creating Computing: Teaching Computers to Think Like a Human*, https://today.ucsd.edu/story/jazz (last visited Oct. 17, 2023).
[95] CITRIS, CDSS, and BAIR Present: Alison Gopnik and Imitation and Innovation in AI: What Four-year-olds Can do and AI Can't (Yet), https://citris-uc.org/event/citris-cdss-and-bair-present-alison-gopnik-on-imitation-and-innovation-in-ai/ (Sept. 27, 2023).

Example 6: UCLA offers a digital humanities minor (undergraduate) and certificate (graduate). In the past, students have worked on several projects, including those using photography as data such as Architectural Reconstructions on Broadway, Landscapes of Injustice: The Despossion of Japanese Americans, the Early Caribbean Digital Archive, and W.E.B Du Bois, the 1900 Paris Exposition, which showed the barriers and progress of African Americans after the abolishment of slavery.[96]

Example 7: UC Irvine Professor Emeritus Benjamin Colby uses artificial intelligence to analyze life histories, diaries, and other cultural data to interrogate the relationship between culture and stress, cultural evolution, and adaptive potential.[97]

---

[96] Instagram post, UCLA DH, https://www.instagram.com/p/CqD7jr5PyI5/?short_redirect=1&img_index=3 (posted Mar. 21, 2023) (last visited Oct. 16, 2023).
[97] UC Irvine, Faculty Profile System, Benjamin Colby, https://faculty.uci.edu/profile/?facultyId=2607 (last visited Oct. 17, 2023).