AUTHORS ALLIANCE | authorsalliance.org 510.480.8302 | 2108 N ST #8898 Sacramento, CA 95816

March 14, 2025

RE: Request for Information on the Development of an Artificial Intelligence (AI) Action Plan (*Sent via email to: ostp-ai-rfi@nitrd.gov*)

Introduction

Authors Alliance appreciates the opportunity to provide feedback to the Office of Science and Technology Policy (OSTP), on the Development of an Artificial Intelligence Action Plan.

Authors Alliance is a nonprofit organization with the mission to advance the interests of authors who want to serve the public good by sharing their creations broadly. We create resources to help authors understand and enjoy their rights and promote policies that make knowledge and culture available and discoverable. In addition, we advocate on their behalf before Congress, the courts, and other government entities.[1] Many of our members are academic researchers who are engaged with AI and text data mining research. Authors Alliance has played a key role in supporting their work, for example, by advocating before the U.S. Copyright Office for exemptions from the DMCA to conduct that work.

This response is primarily focused on copyright law and its impact on AI training and innovation. Right now, there are 39 AI lawsuits currently pending before district courts across the United States. These cases, along with threats of litigation from large copyright holders, have cast a shadow of uncertainty on AI development. Meanwhile, other jurisdictions including the EU, Japan, and others have taken decisive steps to provide clarity on certain aspects of how copyright law applies to AI and text data mining.

US Copyright law has played a major role in both developing the incredible creative industries homed in the US, as well as driving leading scientific research and commercial innovation.[2] The key to this innovation policy has been a thoughtful balance between providing a degree of control over copyrighted works to copyright holders while allowing for flexibility when it comes to technological innovation. Fair use has been the most critical part of this balance, consistently allowing new innovations - from home

[1] Authors Alliance, "About Us," https://www.authorsalliance.org/about/.
[2] Fred von Lohmann, "Fair Use as Innovation Policy," *Berkeley Technology Law Journal*, Vol. 23, No. 2, 2008, https://ssrn.com/abstract=1273385

video recorders to web search engines to unprecedented digitization projects. It has enabled creators to protect their expression while permitting others to build on, comment, and even criticize those ideas, as well as develop new products by extracting unprotectable facts and ideas from them.

One of the critical building blocks for AI development is access to high-quality data sets for AI training and refinement in order to extract facts and ideas from them, and identify patterns among them. Many of those materials are protected by copyright, and fair use has been the primary legal means asserted to gain access to those facts and ideas.[3] The most important thing the Federal Government can do in the copyright realm to protect American innovation is to protect access to works as training data by supporting the application of fair use.

The remainder of our comment explains how the Federal Government can do this by:

(1) Highlighting the role of fair use in AI model training and the need for clarity in preventing time consuming and unnecessary litigation
(2) Surfacing the problem of contractual override of fair use and its impact on AI development
(3) Acknowledging the importance of public data resources to AI and the need to ensure continued access to high-quality training datasets
(4) Emphasizing how AI policies can support both innovation and individuals whose livelihoods may be impacted by AI
(5) Considering how the U.S. might best expedite innovation in the development of AI

Throughout this response, we will aim to provide recommendations that balance the interests of authors, the public good, and the needs of a thriving AI research environment. Thank you for your time and consideration.

Dave Hansen
Executive Director, Authors Alliance

[3] To be clear, not every use an AI company makes will be fair use. For example, implementation of AI models in tools that allow for reproducing verbatim, entire copies of creative works as an AI output may be a step too far, as we have argued elsewhere.

AI Innovation, Copyright, and Fair Use

A foundational legal issue in AI development is the status of AI training under copyright law. Under U.S. law, the right of fair use, codified in 17 U.S.C. § 107, provides flexibility that has long advanced technology, including allowing unlicensed full copies to be used in search engines and text and data mining. Courts have consistently upheld that transformative uses—those that add new meaning or purpose to copyrighted works—are highly likely to be considered fair use.[4]

AI *training*, which involves processing large bodies of copyrighted works to develop generalized machine learning models, is supported by a strong argument for fair use. Legislative action explicitly recognizing AI training as fair use would go a long way to prevent protracted and innovation-stifling litigation. To be absolutely clear, we believe that AI training will be found to be a fair use by U.S. courts. However, this will take time and may slow the development of AI in some quarters for a number of years.[5] If the goal of this administration is to speed the development of AI as much as possible, legislative interventions offer a means to achieve that goal.[6]

Current litigation against AI developers highlights the need for proactive legal protections. As of March 2025, over 35 lawsuits have been filed against AI companies.[7] Without clear statutory recognition of AI training as fair use, developers face unpredictable and costly legal challenges. While large, well-capitalized corporations are in a better position to absorb the costs of litigation, we are particularly concerned with the chilling effects of litigation on smaller, less well-funded startups and noncommercial researchers.

We strongly believe that innovations in AI development are likely to come from both large corporations and smaller research teams. Recently, the emergence of DeepSeek provided us with a vivid example of the disruptive potential of smaller-scale actors in the

[4] See Authors Guild v. Google, Inc., 804 F.3d 202, 214 (2d Cir. 2015) ("Transformative uses tend to favor a fair use finding because a transformative use is one that communicates something new and different from the original or expands its utility, thus serving copyright's overall objective of contributing to public knowledge.")
[5] For example, *Google v. Oracle* took over 10 years to reach the U.S. Supreme Court, where at last a 6-2 majority held that Google's use of the Java APIs was fair use, and *Authors Guild v. Google* also took over 10 years for it to be denied cert by the Supreme Court, thus sustaining Google's use as fair.
[6] Joshua Levine and Tim Hwang, "Copyright, AI, and Great Power Competition," January 2025, https://www.thefai.org/posts/copyright-ai-and-great-power-competition
[7] Chat GPT is Eating the World, "Master List of Copyright Lawsuits vs. AI Companies in the U.S.," https://chatgptiseatingtheworld.com/2025/01/07/updated-the-master-list-of-ai-copyright-lawsuits-current-total-38/

AI space.[8] Given that smaller actors may be unwilling to take on the legal risk represented by AI development, this administration should make clear—possibly through intervention in these suits– that it supports the application of fair use to AI training. It should also encourage Congress to amend the Copyright Act to explicitly include AI training as an illustrative example of fair use[9] and provide standalone exceptions or safe harbors specifically designed to permit AI training and development.

Section 1202(b) of the Copyright Act[10] has also Become a Stumbling Block in the Training of AI

In many of the copyright infringement lawsuits brought against AI developers, plaintiffs allege violations of 17 U.S.C. § 1202(b). Broadly speaking, Section 1202(b) prohibits the "removal or alteration of Copyright Management Information (CMI)."[11] CMI is poorly defined in the statute, which is just one of many problems created by 1202(b). Violations of 1202(b) come with sizable statutory damage awards – between $2,500 and $25,000 for each violation. Courts are unlikely to find AI developers in violation of 1202(b), but this issue has attracted plaintiffs and continues to make its way through the courts.

Section 1202(b) was codified into law at a time when we were still referring to the internet as the "information superhighway" and CMI was compared to a car's license plate.[12] It was a little used provision of the law for twenty years, and has only recently been reinvigorated in the context of AI litigation. It is a poor fit for the present moment.

---

[8] Alex Tapscott, "How DeepSeek is upending AI innovation and investment after sending tech leaders reeling*," New York Post,* February 1, 2025, https://nypost.com/2025/02/01/tech/how-deepseek-is-upending-ai-innovation-and-investment/ ("Despite concerns about DeepSeek security and that it possibly copied rival ChatGPT, the news sent US AI leaders reeling, causing them to lose more than $1 trillion in total market value — including nearly $600 billion from chip king Nvidia alone.")

[9] 17 U.S.C. § 107 ("Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research…)

[10] 17 U.S.C. § 1202(b).

[11] Maria Crusey, "Copyright Management Information, 1202(b), and AI," https://www.authorsalliance.org/2024/10/30/copyright-management-information-1202b-and-ai/

[12] Information Infrastructure Task Force, *Intellectual Property and the National Information Infrastructure: The Report of the Working Group on Intellectual Property Rights*, (1995), 235, https://www.eff.org/files/filenode/DMCA/ntia_dmca_white_paper.pdf ("Copyright management information will serve as a kind of license plate for a work on the information superhighway, from which a user may obtain important information about the work.")

We strongly believe that the outright repeal of 1202(b) would have little negative impact on the functioning of the Copyright Act. After all, there was very little 1202(b) litigation prior to 2020. At minimum, AI developers should be granted broad immunity from 1202(b) claims, not simply because the claims are frivolous, but because the removal of CMI can often be a necessary and appropriate step in training AI models. CMI, if left in AI datasets, will frequently create a form of noise for AI models that risks degrading their quality. Removing CMI should be an accepted and uncontroversial option for AI developers, rather than a senseless legal requirement that they must find ways to design around.

If 1202(b) remains a viable option for plaintiffs, we anticipate a wave of copyright troll lawsuits, given the possibility of high statutory damage awards. It could well lead to death by a thousand lawsuits and might stifle the development of AI for years to come.

Contractual Overrides of Fair Use and Their Impact on AI Development

While fair use serves as a critical legal doctrine in support of AI development, its effectiveness can be undermined by contractual agreements that restrict these rights - a phenomenon known as "contractual override."[13] This occurs when private parties impose terms, often through licensing agreements or terms of service, that limit or entirely prohibit uses otherwise permissible under fair use. Such contractual restrictions pose significant challenges to AI research and development.

Nature and Mechanism of Contractual Overrides

Typical sources of contractual override include:

1. **Licensing Agreements:** Publishers and content providers may include clauses in their licensing agreements that explicitly restrict activities like text and data mining (TDM) or the use of content for AI training. For instance, a license for access to a digital database might prohibit copying or analyzing the content, even for non-commercial research purposes.

2. **Terms of Service (ToS):** Online platforms often have terms of service agreements that users must accept to access content. These terms can include prohibitions against data scraping, analysis, or other activities essential for AI training, effectively limiting the application of fair use in these contexts.

Impact on AI Research and Development

[13] Dave Hansen, “How to Evade Fair Use in Two Easy Steps,” https://www.authorsalliance.org/2023/02/23/fair-use-week-2023-how-to-evade-fair-use-in-two-easy-steps/

Contractual overrides may undermine AI development in several ways:

- **Inhibition of Research:** Researchers and developers may find themselves unable to utilize vast amounts of digital content for AI training due to restrictive contractual terms, stifling innovation and the advancement of AI technologies.

- **Legal Uncertainty:** Even when a use might qualify as fair under copyright law, the presence of contractual restrictions creates legal ambiguity, which could discourage researchers from pursuing projects due to fear of litigation.

- **Disparities in Global Research:** Unlike the U.S., many countries have already enacted laws that prevent contracts from overriding statutory exceptions for activities like Text and Data Mining. For example, the European Union's Directive on Copyright in the Digital Single Market ensures that contractual terms cannot override exceptions for TDM by research organizations.[14] These disparities in international law place U.S. researchers at a disadvantage, as they must navigate both copyright law and restrictive contracts.

Contract law should not be permitted to override fair use. Policymakers should consider statutory limitations on contractual override, similar to approaches taken in the EU and other jurisdictions.[15] To mitigate the adverse effects of contractual overrides on AI development, we would recommend that this administration work with Congress to enact legislation that limits the ability of private contracts to override fair use rights broadly, and particularly for purposes related to AI research and development. This would place the U.S. in a similar position to other jurisdictions that protect statutory exceptions from contractual override.

Public Data Resources and High-Quality AI Training Sets

AI models rely on extensive datasets to improve accuracy and overall quality. However, concerns have emerged that AI developers may soon hit a "data wall," wherein the

[14] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, https://eur-lex.europa.eu/eli/dir/2019/790/oj ("Article 7: Any contractual provision contrary to the exceptions provided for in Articles 3, 5 and 6 shall be unenforceable.")
[15] Jonathan Band, "Protecting User Rights Against Contract Override," https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1099&context=research ("This compilation assembles the copyright override prevention clauses adopted in 48 countries over the past 30 years.")

availability of high-quality, freely accessible training data diminishes.[16] To counteract this, the U.S. government should invest in large-scale data annotation projects and leverage public archives for AI training, ensuring that U.S.-based AI systems maintain a competitive advantage.

Additionally, the United States possesses vast, high quality publicly funded collections that could be leveraged for AI training. Each day, the Library of Congress alone receives some 15,000 items and adds more than 10,000 items to its collections.[17] Its collections include audio recordings, maps, books, film, and photographs - a rich set of resources for training AI. And the scale of these collections is vast - its National Audio-Visual Conservation Center contains "millions of sound recordings and film, television and video items, representing more than a century of audiovisual production."[18] Expanding access to collections like these, while simultaneously transforming them into datasets specific to AI training, and ensuring that they are properly annotated could support AI systems that are more accurate, reliable, and far richer than any currently available.

The Authors Alliance has a keen interest in this work and is currently working toward making a public interest AI training corpus a reality.[19] We appreciate that librarians and archivists have a deep and hard-won understanding of managing large-scale analog and digital collections; it would be wise to tap into that deep expertise in the coming years. The United States should seriously consider leveraging that expertise to build a large-scale corpus for AI training.

Beyond these collections, the federal government also sponsors the creation of large, varied and high-quality research that should also be leveraged for these purposes. Currently, federal agencies have implemented public access plans to provide readers access to tax-payer funded research produced pursuant to federal grants. The federal government should also consider providing access to these research materials for AI training and development purposes.

---

[16] Kevin Roose, "The Data That Powers A.I. Is Disappearing Fast," *The New York Times*, July 19, 2024, https://www.nytimes.com/2024/07/19/technology/ai-data-restrictions.html

[17] Library of Congress, "Fascinating Facts," accessed March 10, 2025, https://www.loc.gov/about/fascinating-facts/

[18] Id.

[19] Authors Alliance, "The Public Interest Corpus: An Update and Opportunities for Co-Development," https://www.authorsalliance.org/2025/02/24/the-public-interest-corpus-an-update-and-opportunities-for-co-development/

The U.S. already has made some efforts in this direction: the National Artificial Intelligence Research Resource Pilot (NAIRR) being among the most prominent.[20] We recommend that efforts like NAIRR be extended and further supported with the above considerations in mind.

AI, Workforce Development, and Copyright's Role

The emergence of AI has raised concerns about workforce displacement, particularly in creative industries such as journalism, literature, and visual arts. While AI tools offer new opportunities for content creation, they need not come at the cost of human authorship. Instead of restricting AI training through excessive copyright barriers, policymakers should focus on investment and leveraging the skills of authors in contributing to AI training, equipping individuals with the skills and resources necessary to work alongside AI to facilitate its development.

Similar to historical shifts in industrial automation, AI should augment human labor and creativity, rather than replacing it outright. This will best be best accomplished if new creative labor and authorship informs the continued development of AI. This administration should fund creative work on a large scale, in the service of generating data that can fill in any current gaps surfaced by AI developers. Here, we imagine that there may be opportunities to grow and sustain nationwide oral history projects, documentary photography and mapping projects, regional digitization of ephemera, and other similar work.

Government-funded projects could be immediately made available for AI development. Combined with the digitization and annotation of collections held in memory institutions, these investments would pay massive dividends in helping create dynamic and highest-quality public data sets for AI development.

Maximizing U.S. Competitiveness in the Development of AI

To accelerate the development and deployment of artificial intelligence (AI) technologies, the federal government might draw inspiration from its rapid mobilization during the COVID-19 pandemic. During the pandemic, Operation Warp Speed (OWS) demonstrated the effectiveness of public-private collaborations in expediting vaccine development.

By combining government resources with private sector expertise, OWS facilitated the swift creation and distribution of COVID-19 vaccines. A similar approach in the AI sector

[20] National Artificial Intelligence Research Resource Pilot, accessed March 14, 2025, https://nairrpilot.org/

could involve the formation of alliances between federal agencies, libraries and archives, and technology companies to accelerate AI research, development, and implementation.

Again, the United States could generate public datasets in response to specific needs surfaced by the AI development community. It could provide these datasets to AI developers and researchers for the express purpose of AI development, even if copyright may preclude their use for other purposes.

Conclusion

Artificial intelligence could well bring the next great leap forward in human knowledge, creativity, and innovation—but only if we foster it properly with favorable legal and policy foundations. The United States is positioned to continue to lead this charge, leveraging our deep traditions of innovation, robust research institutions, and unparalleled public knowledge repositories. However, without decisive action, we risk allowing legal uncertainty, restrictive contracts, and underutilized or completely untapped resources to stifle progress.

Fair use has long been a bedrock of American innovation. Recognizing AI training as fair use would not only protect this legacy but also ensure that AI development remains accessible to researchers, startups, and independent creators.

The United States has led in past waves of technological transformation by embracing bold, pragmatic policy solutions. Now, we must do so again. By embracing fair use, ensuring access to high-quality public data, and very intentionally building a highly creative workforce ready to engage with AI, we can establish a framework that sustains and accelerates our current levels of innovation.

Submission Statement: This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the AI Action Plan and associated documents without attribution.