AAP | ASSOCIATION OF AMERICAN PUBLISHERS

Electronic submission via *www.regulations.gov*

December 6, 2023

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

Reply Comments

The Association of American Publishers (AAP) appreciates the opportunity to submit these Reply Comments to the Copyright Office in the above referenced Notice of Inquiry regarding copyright law and artificial intelligence ("AI"). We filed previous comments on October 30, 2023, during the initial comment round.

Our submission today is in forceful opposition to the flawed and inaccurate assertions submitted by some tech companies and/or their investors in the first comment round, in which they position copyright and the protection of creative expression as an obstacle to innovation and progress. This is nonsense. Copyright is the engine of free expression, responsible for incentivizing invaluable, creative expression and information that is inherently transformational to both people and society and indisputably essential to democracy.

Copyright protection also fuels markets. The U.S. copyright industries collectively added more than $1.8 trillion in annual value to U.S. gross domestic product in 2021,[1] while offering a vital chain of financial return to support new authorship. Those that seek to weaken these protections disregard the premise that authorship cannot be taken for granted. They petition the government for cover from liability for their calculated disregard of authorship, also ignoring that rightsholders today already routinely license their works for all kinds of digital uses. Rather than working with copyright owners, these companies seek to appropriate literature and other invaluable intellectual property for their own commercial gain, and to bend the law to their will.

Government should have no role in bestowing legal or commercial advantages to AI companies at the expense of authors, publishers, and other creators.

Preface

Human authorship is more important than ever in the age of artificial intelligence.

It is not and should not be a copyright owner's burden to subsidize the development of generative AI ("Gen AI") technologies. The companies that benefit from the commercialization of this technology should be required not only to compensate rights holders for their past ingestion of copyrighted works to train Gen AI systems but also for their ongoing and future use of protected works to train new Gen AI systems or fine-tune their existing products.

[1] See Copyright Industries in the U.S. Economy – The 2022 Report at https://www.iipa.org/files/uploads/2022/12/IIPA-Report-2022_Interactive_12-12-2022-1.pdf.

Copyright owners vigorously contest the assertion by tech companies that copyright industries should cease objecting to the un-permissioned and uncompensated use of protected works to create training datasets for Gen AI systems because—as one tech commenter puts it—Gen AI developers have already invested significantly in this new technology on "settled expectations" of fair use.[2] In response, we note that their "expectations" have no basis in the law; this is an entirely new use with massive potential ramifications, and fair use is famously a balance of factors and equities—it depends on the particulars of each situation. Rights holder expectations have always been that AI companies should seek consent from rights holders for the use of their works and, further, that they should incorporate into their planning the simple, well-established principle of licensing what you do not own.

It is clear from several submissions in the initial comment round that Gen AI developers gambled on a flawed assumption that wholesale reproduction for Gen AI training is categorically fair use, and that policy makers would give them a free pass in the name of innovation, or even more incredible, hardship. It is deeply ironic that these billion-dollar companies bemoan the financial burden they would face if they were required to pay reasonable license fees to the copyright owners whose works are the very building blocks of Gen AI and whose livelihoods are threatened by the same systems. As astutely stated by one journalist, "[w]e know what they're up to: privatizing the profits they might collect from AI applications while sticking the creative public with the costs. Will we let them pull this same stunt again?"[3]

AAP must also respond to comments from "big tech" that characterize the "nascent" nature of the AI industry. These same arguments raised at the dawn of the Internet more than 30 years ago—i.e., that start-up technology companies should not be hampered by pre-existing public laws; that copyrights are an impediment to innovation; and that imposing restrictions on U.S. tech companies will impede their ability to dominate the global marketplace—are again being advanced as if new.

The Copyright Office is quite familiar—as is everyone—with these well-worn and disingenuous claims. Gen AI developers are not struggling "start-ups" that need a boost from the government. They count among their investors some of the largest and most profitable technology companies in the world and are valued, in some instances, between $80-90 billion dollars.[4] There is absolutely no public policy reason to create legal immunities for such companies, who face only the reasonable requirement that they seek the consent of, or licenses from, rights holders whose works they use for training their Gen AI systems.

It would be a grave error to repeat the past policy mistakes that allowed technology companies to achieve such an unhealthy, monopoly-like market dominance to the point that governments have struggled to curb their power, despite repeated attempts to moderate their

---

[2] See Andreesen Horowitz at 6.

[3] Michael Hiltzik, Column: *AI investors say they'll go broke if they have to pay for copyrighted works. Don't believe it*, L.A. TIMES, Nov. 16, 2023, *available at* https://www.latimes.com/business/story/2023-11-16/ai-investors-say-theyll-go-broke-if-they-have-to-pay-for-copyrighted-materials-dont-buy-it.

[4] See OpenAI in talks to sell shares at $86 billion valuation - Bloomberg News.

aggressive marketplace tactics.[5] To the contrary, there is every reason now to ensure that Gen AI companies develop their products and services fairly and transparently, and that includes respecting the authorship that is the building block of their businesses.

In related submissions, some tech commenters have tried to create a sense of crisis about a purported "global AI arms race," where the ascendancy and dominance of the U.S. AI industry hangs in the balance, and disregard for copyright law is the only way the U.S. can maintain its lead. For instance, one claims that "the overzealous enforcement of copyright when it comes to AI training—or the *ad hoc* limitation of the fair use doctrine that properly protects AI training—could cost the United States the battle for global AI dominance."[6] Another argues—seemingly oblivious to the irony of their assertion—that basic transparency requirements regarding training data would "disclose *their* trade-secret protected information," which would "accomplish little practical benefit, while simultaneously imposing burdensome reporting obligations that would force AI developers to disclose sensitive and valuable information to the world, including to our foreign rivals."[7]

This specter of the U.S. losing the battle for global AI dominance is speciously advanced as justification for ignoring the rights of authors and publishers. This could not be further from the truth. The issue of national security is certainly of deep concern to all American citizens, particularly where bad actors may use AI and Gen AI systems to sow misinformation or disinformation that undermine our democratic institutions and create other national security risks. This possibility points to the even greater need for authors and publishers that produce and disseminate vetted, fact-checked quality content.

Advancing the nation's technological and economic agenda is not a zero-sum game, and it should not become a race to the bottom. As noted by one member of the National AI Advisory Committee (NAIAC), "We want these technologies to be developed according to our norms and ethics."[8] The U.S. can continue to lead the responsible and ethical development of advanced Gen AI systems while respecting the rights of creators and producers, authors and publishers—without whom the creation and dissemination of the works essential to training Gen AI systems

---

[5] See, e.g., Subcommittee on Antitrust, Commercial, and Administrative Law of the Committee on the Judiciary of the House of Representatives, Investigation of Competition in Digital Markets, Majority Staff Report and Recommendations, Part I (July 2022); see also Department of Justice, Justice Department Sues Google for Monopolizing Digital Advertising Technologies (Jan. 24, 2023), https://www.justice.gov/opa/pr/justice-department-sues-google-monopolizing-digital-advertising-technologies; Klobuchar, Grassley, Colleagues to Introduce Bipartisan Legislation to Rein in Big Tech (Oct. 14, 2021), https://www.klobuchar.senate.gov/public/index.cfm/2021/10/klobuchar-grassley-colleagues-to-introduce-bipartisan-legislation-to-rein-in-big-tech; Federal Trade Commission, FTC Sues Facebook for Illegal Monopolization (Dec. 9, 2020), https://www.ftc.gov/news-events/news/press-releases/2020/12/ftc-sues-facebook-illegal-monopolization.

[6] Andreesen Horowitz at 8.

[7] Meta at 20.

[8] Michael Richards, AI's Role in Modernizing Intellectual Property and Bolstering National Security (Aug. 1, 2022) (quoting Yll Bajraktari), https://www.uschamber.com/technology/ais-role-in-modernizing-intellectual-property-and-bolstering-national-security.

would not be possible. For without their creations, their literary inventions—on what then will Gen AI systems train?

The U.S. is a global leader in copyright protection[9] and in copyright industry production. If the U.S. does not uphold its own laws to protect rights holders, other nations would be unlikely to do the same, harming its own creative industries and the livelihoods of all creators dependent upon a strong copyright framework. Claiming that AI "may be the most important technology our civilization has ever created"[10] is meritless if it is not responsibly and ethically developed, or its deployment and use not subject to the same common-sense regulations that apply to all other technologies.

In summary, U.S. innovation is not at risk. What is at risk are the important tenets of copyright law from the self-serving rhetoric and actions of Gen AI developers that would disregard the rights of authors and publishers who contribute to the country's rich cultural, economic, and intellectual life. Creating a digital royalty system to compensate copyright owners and other rights holders is neither a novel nor insoluble challenge. Today such systems already exist and are an established part of the Internet economy. In fact, such payment schemes are already fundamental to digital commerce for creative content—from music to literature to news. A Gen AI business that cannot contract with rightsholders for the building blocks of its technology is simply not innovative at all.

Respectfully, and as set forth in more detail in the pages that follow, AAP believes that the facts and legislative history lead to the following conclusions about copyright and artificial intelligence systems:

1. The large scale copying and use of copyrighted content to create the training datasets ingested by Gen AI models requires consent from rightsholders. Neither the Copyright Act nor case law supports any colorable application of fair use to excuse the wholesale unauthorized reproduction of copyrighted works to train Gen AI models.

2. The un-permissioned use of copyrighted works to train Gen AI models is inconsistent with the public policies animating copyright law and the objectives it is intended to promote.

3. Corrective legislation may be necessary if court decisions expand fair use beyond its appropriate boundaries to permit wholesale, un-permissioned use of copyrighted works. This would eviscerate the reproduction right and would be in sharp contrast to what Congress intended.

4. Transparency is an essential requirement. It is in the public interest to know what works of authorship have been ingested and an essential part of seeking proper consent to have such information clearly recorded. Such a requirement is not

[9] U.S. Chamber of Commerce, Global Innovation Policy Center, International IP Index, 2023 Eleventh Edition 7 (2023), https://www.uschamber.com/assets/documents/GIPC_IPIndex2023_FullReport_final.pdf.
[10] Andreesen Horowitz at 2.

burdensome and lends itself to further innovation in the field of digital rights enterprises. Virtually all stakeholders in the AI space are already subject to reporting requirements for privacy laws and do not dispute their ability to comply with the extensive existing reporting requirements for privacy laws.

1. Gen AI models copy and use the expressive content of copyrighted works for training, contrary to claims that any copying is only of unprotected or nonexpressive aspects of works.

With respect to the training of Gen AI and Large Language Models on text, the expressive content of works is being copied. Claims that ingestion is permitted by fair use, or even that any copying is non-infringing, inaccurately represent what is being copied.[11] The Gen AI training

[11] Andreessen Horowitz at 6 ("when an AI model is trained on copyrighted works, the purpose is not to store any of the potentially copyrightable content (that is, the protectable expression) of any work on which it is trained. Rather, training algorithms are designed to use training data to extract facts and statistical patterns across a broad body of examples of content—i.e., information that is not copyrightable."); Anthropic at 7 ("The copying is merely an intermediate step, extracting unprotectable elements about the entire corpus of works, in order to create new outputs. In this way, the use of the original copyrighted work is non-expressive; that is, it is not re-using the copyrighted expression to communicate it to users. To the extent copyrighted works are used in training data, it is for analysis (of statistical relationships between words and concepts) that is unrelated to any expressive purpose of the work."; Authors Alliance at 9 ("The use of copyrighted works as training data for generative AI models is in most cases a non-expressive use, as it is done as an intermediate step in producing non-infringing content, such as by extracting non-expressive information such as patterns, facts, and data in or about the work."); BSA at 8 ("In such a scenario, the AI developer would not be reproducing this text for its expressive purpose. Rather, the reproductions would be made solely for the purpose of extracting unprotected information about the English language–i.e., the correlations, patterns, and relationships among the 26 letters of the alphabet and the 1 million English language words, as they appear in thousands of stock phrases, figures of speech, similes, metaphors, grammar patterns, and common linguistic formulations and expressions. Neither these letters, words, and phrases, nor the mathematical patterns among them across thousands or millions of writings, are copyright protectable subject matter."); EFF at 2-3 ("it is crucial to note that a machine learning process is not limited to, or even primarily, identifying patterns that reflect copyrightable elements of an original work. Rather, much of its "learning" is focused on non-copyrightable properties of the subject being depicted (limes being green, trees having a vertical trunk with branches, cats having four legs, etc.) or scenes-a-faire that are free for all to use (formulaic premises and plot twists in language models, archetypal uses of visual language and color to convey mood)."); New Media Rights at 11 ("First, it must be understood that when OpenAI's natural language processing functions "read" text they do not analyze the meaning but instead the functionality of sentence structure and syntax. So, while the LLM is built by digesting creative material, that material is not processed for its copyrightable expression, but rather for its non-copyrightable aspects: the function of language itself."); OpenAI at 11 ("First, when undergoing pre-training, a model is not interested in the expressive aspects of individual copyrighted works. Instead, as described above, the pre-training process is a highly sophisticated computational process that teaches the model to analyze the structure and syntax of language and images in general terms, to discern the statistical relationships between words, shapes, colors, textures, and concepts."); Program on Information Justice and Intellectual Property (PIJIP), American University Washington College of Law at 5

process does not extract the ideas, facts, or concepts being conveyed by an author, it solely extracts the exact expressive choices made to convey those ideas—i.e., the words an author used, and the order in which they were placed.[12] Such copying is unlikely to be transformative or fair use. The Copyright Office may wish to note the statement of a former executive of the Gen AI firm, Stability AI, who wrote,

> I've resigned from my role leading the Audio team at Stability AI, because I don't agree with the company's opinion that training generative AI models on copyrighted works is 'fair use'... But setting aside the fair use argument for a moment—since 'fair use' wasn't designed with generative AI in mind—training generative AI models in this way is, to me, wrong. Companies worth billions of dollars are, without permission, training generative AI models on creators' works, which are then being used to create new content that in many cases can compete with the original works. I don't see how this can be acceptable in a society that has set up the economics of the creative arts such that creators rely on copyright. [13]

These comments align with the recent Supreme Court ruling in *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, "It will not impoverish our world to require AWF to pay Goldsmith a fraction of the proceeds from its reuse of her copyrighted work. Recall, payments like these are incentives for artists to create original works in the first place."[14]

We also note a point raised by Google in its initial comments in this process: Gen AI models are not information retrieval systems.[15] The occurrence of "hallucinations" demonstrates that Gen AI models do not "understand" meaning or ideas. Even the use of the term

---

("Most of what gets 'extracted' from the training data is factual, consisting of information about grammatical relationships between words. This statistical data is often uncopyrighted factual information."); Project LEND at 6 ("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.'"); Professors Samuelson, Sprigman, and Sag at 8 ("Understanding the process of training foundation models is relevant to the generative AI systems' fair use defenses because the scope of copyright protection does not extend to 'statistical information' such as 'word frequencies, syntactic patterns, and thematic markers'."); Van Lindberg at 24 ("Rather than copying any expression, however, the model training process records facts about the work.").

[12] See, e.g., Gervais, Daniel at 3 ("[I]n the case of GenAI, the use by the machine is not mere character recognition; it is semantic in nature. The machines process the expression of ideas in the works to create new expression."); News Media Alliance at 33 ("LLMs typically ingest valuable media content for their written expression. To the extent they are ingesting this content so these published words can be analyzed 'in relation to all the other words in a sentence,' or their sequences of words identified, that analysis and identification is intended to capture the very expression that copyright protects. Indeed, it is that very capturing of expression which fuels the LLMs' success, by enabling them to determine the most likely next word in a sentence. That is why LLMs that are trained to generate their own expressive works 'copy expression for expression's sake.'").

[13] *Why I just resigned from my job in generative AI*, Music Business Worldwide, *available at* https://www.musicbusinessworldwide.com/why-just-resigned-from-my-job-generative-ai/.

[14] *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith,* 143 S. Ct. 1258, 1286 (2023).

[15] Google at 4.

"hallucination" may lead to a misinformed view of how Gen AI models work, since it applies a distinction between the accurate and inaccurate representation of facts and reality by a model. In fact, a Gen AI model never generates output with reference to facts and reality; any similarity between its output and reality is merely a byproduct of the probabilities derived from the expression in its training materials.[16]

2. The Copyright Office should reject arguments to narrow the scope of exclusive rights.

Arguments made in favor of permitting the copying of copyrighted works to train Gen AI assert unfounded claims about the nature of the exclusive rights protected by copyright law and should be rejected. By reproducing copyrighted works to train Gen AI models, Gen AI developers are expropriating the value of those works, value which the copyright owner has created and to which he or she is entitled. The Copyright Act is clear that "the owner of copyright … has the exclusive rights … to reproduce the copyrighted work in copies,"[17] and "[a]nyone who violates any of the exclusive rights of the copyright owner … is an infringer of the copyright."[18] Contrary to the arguments of some respondents, the statutory framework draws no distinction between what respondents consider "legitimate" interests and otherwise; does not limit protection to some subset of uses that respondents believe copyright "is intended to protect;" nor supports the idea that some types of copying fall outside the scope of the exclusive rights, e.g., "non-expressive" copying.[19]

The plain meaning of the statute supports a broad reading, and the breadth of protection is a feature, not a bug. The House Report accompanying the 1976 Copyright Act explains, "The approach of the bill is to set forth the copyright owner's exclusive rights in broad terms in section 106, and then to provide various limitations, qualifications, or exemptions in the 12 sections that follow."[20] The drafters of the Act intended, by this broad language, to ensure that copyright

---

[16] Oliver Bown, "'Hallucinating' AIs Sound Creative, but Let's Not Celebrate Being Wrong, The MIT Press Reader (Oct. 13, 2023), https://thereader.mitpress.mit.edu/hallucinating-ais-sound-creative-but-lets-not-celebrate-being-wrong/.

[17] 17 U.S.C. § 106.

[18] 17 U.S.C. § 501(a).

[19] Andreessen Horowitz at 6 ("Gen AI model training is a productive, non-exploitative use of training material. That type of use does not exploit any protectable expression in any given work, and so it does not implicate any of the legitimate rightsholder interests that copyright law seeks to protect."); Anthropic at 7 ("the use of the original copyrighted work is non-expressive; that is, it is not re-using the copyrighted expression to communicate it to users."); BSA ("such functionality simply does not compete with any copyrighted works in any manner that copyright is intended to protect."); CCIA at 8 ("These non-expressive copies are not consumable by the public and do not function as market substitutes for copies of the ingested works."); Meta at 10 ("The process of training and developing AI models does not necessarily trigger the rights that copyright exists to protect."); TechNet at 4-5 ("the creation of intermediate copies in furtherance of the creation of a new and useful technological tool is not the kind of copying that violates copyright law.").

[20] H.R. 94-1476 at 61. See also House Comm. on the Judiciary, 87th Cong., Copyright Law Revision Part 6: Supplementary Report of the Register Of Copyrights on the General Revision of the U.S. Copyright Law: 1965 Revision Bill 14 (Comm. Print 1965) ("we believe that the author's rights should be stated in the

protection would encompass not only the breadth of technological uses known at the time of enactment, but also future technological uses.[21] They did so by, for example:

- Defining "copies" to include material objects in which a work is fixed "by any method now known *or later developed*, and from which the work can be perceived, reproduced, or otherwise communicated, either directly *or with the aid of a machine or device*,"[22] language intended by Congress—

> "to avoid the artificial and largely unjustifiable distinctions … under which statutory copyrightability in certain cases has been made to depend upon the form or medium in which the work is fixed. Under the bill it makes no difference what the form, manner, or medium of fixation may be-whether it is in words, numbers, notes, sounds, pictures, or any other graphic or symbolic indicia, whether embodied in a physical object in written, printed, photographic, sculptural, punched, magnetic, or any other stable form, and whether it is capable of perception directly or by means of any machine or device "now known or later developed."[23]

- Clarifying that "device," "machine," or "process," includes "one now known *or later developed*."[24]

- Creating, for the first time, the public display right.[25] In explaining this critical change, this Office presciently said:

> "Since the *Report* was issued in 1961 we have become increasingly aware of the enormous potential importance of showing, rather than distributing, copies as a means of disseminating an author's work. In addition to improved projection equipment, the use of closed-and open-circuit television for presenting images of graphic and textual material to large audiences of spectators could, in the near future, have drastic effects upon copyright owners' rights. Equally if not more significant for the future are the implications of information storage and retrieval devices; when linked together by communications satellites or other means, these could eventually provide libraries and individuals throughout the world with access

---

statute in broad terms, and that the specific limitations on them should not go any further than is shown to be necessary in the public interest.").

[21] House Comm. on the Judiciary, 87th Cong., Copyright Law Revision Part 6: Supplementary Report of the Register of Copyrights on the General Revision of the U.S. Copyright Law: 1965 Revision Bill 14 (Comm. Print 1965) ("A real danger to be guarded against is that of confining the scope of an author's rights on the basis of the present technology so that, as the years go by, his copyright loses much of its value because of unforeseen technical advances.").

[22] 17 U.S.C. § 101 (Emphasis added).

[23] H.R. 94-1476 at 52.

[24] Id (Emphasis added).

[25] 17 USC § 106(5).

> to a single copy of a work by transmission of electronic images. It is not inconceivable that, in certain areas at least, 'exhibition' may take over from 'reproduction' of 'copies' as the means of presenting authors' works to the public, and we are now convinced that a basic right of public exhibition should be expressly recognized in the statute.[26]"

- Expanding the public performance right.[27] These last two changes to the public display and performance rights were made in part because of Congress's observation that performances and displays would continue "to supplant markets for printed copies."[28]

Together, these features of the Copyright Act affirm a broad and unstrained reading of the exclusive rights. Contrary to arguments to read the reproduction right out of the Copyright Act, the text, structure, and legislative history demonstrate a goal of strengthening and supplementing that right. In analyzing issues related to the use of copyrighted works to train Gen AI, the Copyright Office should keep this view in mind and avoid redefining the exclusive rights.

3. The Copyright Office should reject analogies between human learning and Gen AI model training because they are reductive, unhelpful, and ignore the plain text of the Copyright Act.[29]

The analogy between human learning and Gen AI model training is illuminating in one respect: it demonstrates that the use of a copyrighted work to train Gen AI models is for the same intrinsic purpose as the author's purpose, which "seriously weakens a claimed fair use."[30]

---

[26] House Comm. on the Judiciary, 87th Cong., Copyright Law Revision Part 6: Supplementary Report of the Register Of Copyrights on the General Revision of the U.S. Copyright Law: 1965 Revision Bill 20 (Comm. Print 1965).

[27] 17 U.S.C. § 106(4).

[28] H.R. 94-1476 at 63.

[29] AIPLA at 12 ("We also note that the issues presented by AI infringement are analogous to human infringement. Conceptually, generative AIs use a similar process to that used by human authors."); Committee for Justice at 3 ("The similarities between machine learning in generative AI models and the way in which human creators learn from a lifetime of examples makes it possible to apply the law governing the latter to the former."); Creative Commons at 2 ("Generative AI can function in a similar way. Just as people learn from past works, generative AI is trained on previous works, analyzing past materials in order to extract underlying ideas and other information in order to build new works."); Public Knowledge at 6 ("[T]raining a GAI system is generally analogous to reading a book, looking at a photograph, admiring a painting, or listening to music."); Van Lindberg at 5-6 ("AI systems learn about language by hearing and reading language, just like humans. AI systems learn about art by observing art, just like humans. AI systems learn about music by hearing and reading music, just like humans. There is no other way for the system to 'learn.'").

[30] Weissmann v. Freeman, 868 F.2d 1313, 1324 (2d Cir. 1989); Soc'y of the Holy Transfiguration Monastery, Inc. v. Gregory, 689 F.3d 29, 60 (1st Cir. 2012) (citing Weissman); Divine Dharma Meditation Int'l, Inc. v. Inst. of Latent Energy Studies, No. 19-55264, 2021 U.S. App. LEXIS 25145, at *3 (9th Cir. Aug. 23, 2021)

But other than that, equating the copying done to train Gen AI models with human learning mispresents what is occurring and will lead to incorrect legal conclusions. A human does not copy a work, under the text of the Copyright Act, when he or she learns. By contrast, when a Gen AI developer uses a copyrighted work to train a model, copies are made under the plain text of the Act.

Such copying should not be treated as a mere technicality or an accident, and it is crucially important that the Office recognize the centrality of the exclusive rights to advancing the goals of copyright law. The drafters of the 1976 Act debated this very question of "inputting" of copyrighted works into computers and determined that it should be considered as implicating the copyright owner's exclusive rights, which would be infringement absent any exception.[31] Their conclusion was subsequently reaffirmed by the National Commission on New Technology Uses of Copyrighted Works.[32] This legislative history supports the intent of Congress that is best evidenced in the plain text of the Act. In this context, speculative analogizing about the differences between how humans learn and how AI "learns" is both misleading and harmful.

(same); Cambridge Univ. Press v. Patton, 769 F.3d 1232, 1263 (11th Cir. 2014) (verbatim copies of book excerpts for course reserves not transformative because they serve same intrinsic purpose as plaintiffs' works).

[31] House Comm. on the Judiciary, 87th Cong., Copyright Law Revision Part 6: Supplementary Report of the Register Of Copyrights on the General Revision of the U.S. Copyright Law: 1965 Revision Bill 18 ("An important question that has emerged since publication of the *Report* in 1961 involves computer uses of copyrighted materials. Mainly in an effort to stimulate a discussion of the issue, the preliminary draft of 1963 contained a provision granting an exclusive right 'to reproduce [the work] in any form in the programming or operation of an information storage and retrieval system.' We became convinced, however, that it would be a mistake for the statute, in trying to deal with such a new and evolving field as that of computer technology, to include an explicit provision that could later turn out to be too broad or too narrow. A much better approach, we feel, is to state the general concepts of copyright in language, such as that in section 106(a), which would be general in terms and broad enough to allow for adjustment to future changes in patterns of reproduction and other uses of authors' works. At the same time, we should emphasize here that, unless the doctrine of 'fair use' is applicable in a particular case, the bill contemplates that certain computer uses would come within the copyright owner's exclusive rights. It seems clear, for example, that the actual copying of entire works (or substantial portions of them) for 'input' or storage in a computer would constitute a 'reproduction' under clause (1), whatever form the 'copies' take: punch-cards, punched or magnetic tape, electronic storage units, etc. Similarly, at the 'output' end of the process, the "retrieval" or 'print-out' of an entire work (or a substantial part of it) in tangible copies would also come under copyright control.")

[32] Final Report of the National Commission on New Technology Uses of Copyrighted Works at 39 ("The protection afforded by section 106 of the new law seemingly would prohibit the unauthorized storage of a work within a computer memory, which would be merely one form of reproduction, one of the exclusive rights granted by copyright. Considering the act of storing a computerized data base in the memory of a computer as an exclusive right of the copyright proprietor appears consistent both with accepted copyright principles and with considerations of fair treatment for potentially affected parties. Making a copy of an entire work would normally, subject to some possible exception for fair use, be considered exclusively within the domain of the copyright proprietor. One would have to assume, however, that fair use would apply rarely to the reproduction in their entirety of such compendious works as data bases.")

Notwithstanding the plain text, there are clear factual differences between human learning and Gen AI training that render analogies unhelpful and reductive. Copyrighted works can be ingested into machines on a scale exponentially faster than a human can read. A machine can recall—verbatim—vast amounts of data instantaneously. And a model could output verbatim copies on a systemic and wholesale basis even though a Gen AI developer might seek to reduce infringing outputs by taking certain steps.

Finally, in no case does the Copyright Act permit unauthorized access to or acquisition of copyrighted works. Lawful access to authorized sources matters. A human is not permitted to illegally reproduce and download 183,000 copyrighted works (the number of infringing titles estimated in the "Books3" corpus)[33] in order to read or learn from them.

4. Gen AI models can easily reproduce verbatim works used for training, in whole or part. Developers must take affirmative steps to reduce such "memorization" or "overfitting" (with the caveat that such steps may never be 100% effective).

Some respondents attempt to dismiss the generation of outputs that are verbatim copies of copyrighted works used as inputs as an insignificant error but concede that developers must take affirmative steps to avoid this occurrence.[34] However the law must recognize things as they are, not as good-faith actors wish them to be. The Office should also take note that bad-faith or negligent actors may deliberately design Gen AI models to generate infringing outputs.

Memorization is an inherent characteristic of Gen AI models, and the law should recognize that Gen AI developers and operators have an obligation to design their models to avoid the generation of infringing outputs.

---

[33] Alex Reisner, These 183,000 Books are Fueling the Biggest Fight in Publishing and Tech, The Atlantic (Sep. 25, 2023), https://www.theatlantic.com/technology/archive/2023/09/books3-database-generative-ai-training-copyright-infringement/675363/.

[34] Anthropic at 6 ("We don't believe users should be able to create outputs using Claude that infringe copyrighted works. That is not an intended or permitted use of this technology, and we take steps to prevent it."); BSA at 11 ("In any well-designed AI model trained on a sufficiently large data set, computational analysis should never (or only in the rarest of circumstances) produce outputs that are 'substantially similar,' let alone 'virtually identical,' to any specific copyrighted work."); EFF at 4 ("memorization of specific documents is considered a bug, not a feature."); Google ("The possibility that AI models can occasionally, despite the best efforts of their developers, output content that replicates existing expression is a bug not a feature, and developers are taking a range of measures to limit that occurrence even further, including deduplication of training data."); New Media Rights at 12 ("While it is true that some LLMs can be designed to produce an output that resembles copyrighted inputs, that does not appear to be the case for OpenAI's current functionality."); OpenAI at 7 ("Because our models do not have access to training information after they have learned from it, they are unlikely to duplicate training data in their outputs. In fact, verbatim repetition or 'memorization' of training data is generally considered by AI developers to be a bug to be corrected, rather than a feature to be pursued. OpenAI has employed numerous measures to reduce the incidence of this happening, and we regularly update our practices to deploy more."); Van Lindberg at 35 ("While it might be possible to make guesses, to find particular 'memorized' inputs, or to construct toy AI models that have this property, it is not possible for any current system that is commercially reasonable or academically interesting.").

5. Transparency is essential to the responsible and ethical development of trustworthy AI systems.

Transparency promotes accountability and is necessary to build public trust in AI technologies and in the products and services into which AI technologies may be embedded. Specifically with respect to the (un-permissioned) use of copyrighted works to train Gen AI systems, accurate record-keeping and appropriate disclosure of copyrighted works incorporated into training datasets will provide rights holders with the information necessary to identify the party from whom compensation is owing for such use (i.e., the training dataset creator and/or the Gen AI developer). Where licensing is occurring, and as licensing models continue to evolve, accurate record keeping of how and which copyrighted works are used to train Gen AI systems may provide the mechanism through which revenue sharing models may be defined (for instance, based on how much a specific copyrighted work used as an input contributed to a particular output).

6. The U.S. should continue to encourage other countries to maintain robust copyright protections for creators and copyright owners regarding the use of their works in training Gen AI models.

Several submissions noted that international consistency regarding the copyright legal framework for training AI models on copyrighted works would be beneficial. However, there is disagreement over whether that consistency should aim to protect human creators and copyright owners or subsidize AI developers.[35]

The U.S. should continue to encourage other countries to maintain robust copyright protections, which would not be contrary to facilitating continued investment in and development of Gen AI models. In particular, AAP suggests that Japan and Singapore be encouraged to reassess the TDM exceptions in their laws in view of subsequent developments. Notably, the existing TDM exceptions in the copyright laws of several countries preceded the

[35] *Compare, e.g.,* A2IM-RIAA at 8-10; Authors Guild at 6-7; Copyright Alliance at 19-22; DPE, AFL-CIO at 4-5; Getty Images at 6-7; NMA at 13-16; NMPA at 4-6; NWU at 7-8; *with* Anthropic at 4-5; Creative Commons at 2-3; Stability AI at 8; Wikimedia at 4-5.

public emergence of Gen AI in late 2022;[36] said exceptions may not apply to Gen AI;[37] and they likely violate the Berne Convention if applied to Gen AI.[38]

The existing TDM exceptions, if applied to Gen AI, are also flawed policy. The livelihoods and professions of authors, publishers, and all those integral to the publishing endeavor, and by

[36] E.g. AAP at 5 ("We note that TDM exceptions were adopted before the rise of Gen AI, and the original aim was to facilitate computational analysis of large amounts of text as may be embodied in scholarly and scientific articles for research purposes."); UMG at 16-18 ("[T]here are some policies, including ones that were adopted years ago before the rise of generative AI [(i.e. in Singapore and Japan)], we believe the U.S. should avoid. Whatever their historical merit, generative AI poses threats that render them obsolete and damaging for the creative community, the music industry, and the general integrity of intellectual property law.").

[37] E.g., AAP at 21 ("[T]he [EU DSM] Directive preceded the rapid growth and significant attention to Gen AI systems over the last year. The Directive does not refer to AI, much less Gen AI, nor does it suggest a scientific research context would be permissible where text, sounds, or images would be used to generate new text, sounds, or images."); CCC at 2-5 ("[S]ome countries have provided exceptions or limitations for specific subsets of AI (namely text and data mining (TDM)), but it is unclear how these provisions apply to rapidly changing technologies such as those now seen in generative AI models.").

[38] E.g. Copyright Alliance at 19-22 (discussing the approaches in the European Union, Japan, Singapore, and the United Kingdom, and concluding that "none of these approaches should be considered in the United States, as this would not only require a change to the Copyright Act but could also potentially result in U.S. noncompliance under the Berne Convention for the Protection of Literary and Artistic Works[, since] such exceptions may not be compliant under the Berne three-step test"); Gervais at 2 ("The legislative changes in the European Union, Japan, Singapore, and Switzerland, to name just those, are all very different. For one thing, they have different limits. It is not clear that all of them are compatible with the TRIPS Agreement's dual three-step test that applies to exceptions and limitations to the right of reproduction (namely article 9(2) of the Berne Convention, incorporated into TRIPS) and article 13 of TRIPS. It thus does not seem warranted to follow any one of those approaches at this juncture."); Kernochan Center at 3-4 ("The opt-out [of Article 4 of the EU DSM Directive] poses problems of compatibility with Berne Convention norms, because it conditions the scope of protection on a declaratory obligation. Berne art 5(1) provides 'The enjoyment and the exercise of these rights [including reproduction and communication to the public] shall not be subject to any formality.' As detailed in Jane C. Ginsburg, Berne-Forbidden Formalities and Mass Digitization, 96 BOSTON U. L. REV. 745, 758-68 (2016), available at http://ssrn.com/abstract=2772176, a back-door limitation on the scope of protection through a formality-freighted opt-out from an exception offends Berne just as much as a front-door imposition of a formality as a condition of protection."); MPA at 12-14 ("MPA submits that these types of exemptions [i.e. Japan's and Singapore's] are bad policy, and they likely fail to comply with the Berne Convention's "three-step" test."); NMPA at 4-6 ("Categorical exemptions from copyright liability violate the three-step test of the Berne Convention, which limits permissible exceptions to the reproduction right only to 'certain special cases, provided that such reproduction does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author.' Sweeping carveouts to copyright protection, such as those [in the EU, UK, Singapore, and Japan] for TDM, go significantly beyond 'special cases.' Such carveouts may also stifle the developing marketplace for licensing copyrighted works for use in AI development and thereby run afoul of the prohibition against interference with the normal exploitation of the work by the owner. . . . Opt-out regimes may also violate Article 5(2) of the Berne Convention which prohibits conditioning copyright protection upon any formality."); PPA at 9-10 ("PPA considers any exception that broadly allows scraping of copyrighted works without the authorization of the copyright owner to violate all three steps of the three-step test that governs permissible exceptions to copyright in international instruments.")

extension the immense public benefit that results from their efforts,[39] should not be put at risk to give private commercial entities, including Gen AI developers and the billion-dollar companies that fund them, subsidies in the form of uncompensated and un-permissioned transfers of their intellectual property. Marginally improving the profit margins of Gen AI developers by sparing them the "inconvenience" of seeking authorization to use copyrighted works and providing reasonable compensation is not a reason to precipitate an existential crisis for the creative sector. Apart from being both fundamentally unfair and contrary to copyright law principles, the harm to creators and the U.S. economy would far exceed any benefit to AI developers.

AAP recommends the U.S. exercise its leadership role internationally to protect the U.S. creative sector and ensure a global level playing field by opposing ill-advised copyright exceptions. We recognize that it is not within the government's power to determine outcomes in other countries, but the effects of misguided policy elsewhere can be mitigated.

Some have argued that requiring Gen AI developers to seek permission from, and compensating rights holders for the use of their works, would make the U.S. less attractive to AI investors in contrast to countries where long standing principles of copyright may be disregarded. We note that these costs are not only relatively minor compared to the overall expenditures of Gen AI developers but also that they are but among a long list of costs that influence where a company chooses to locate its business. These costs include availability of a robust talent pool, quality of the technological infrastructure, proximity to universities with strong programs in relevant fields, access to private equity and venture capital, time-zone compatibility, proximity to customer markets, local languages, local culture/quality of life, tax liability, servers, network infrastructure, real estate, research and development, payment processing, energy, sales, marketing, customer service, legal, finance, human resources, corporate communications and policy, and other administrative and professional services, among others.

According to Tortoise AI, the U.S. remains in first position overall in accelerating AI development, as evaluated against the three pillars of investment, innovation, and implementation.[40] Tortoise's Global AI Index (GAII) evaluates a country's readiness across several factors including talent, infrastructure, operating environment (regulatory framework), research, development, government strategy, and commercial ventures. That the fourth iteration of Tortoise's index, published June 28, 2023, continues to rank the U.S. as No. 1 on its GAII indicates that the existing U.S. framework—with its robust intellectual property protection framework—continues to secure the United States' competitive advantage as a global powerhouse for AI development. With the U.S. already ranking high for many of the GAII indicators, suggestions

---

[39] *See* AAP at 27 ("Licensing fees are an important source of income for U.S. creators and rightsholders and support the continued investment in new human-created works. The importance of sustaining the U.S. publishing industry cannot be understated. AAP members publish high-quality literary works, including works that present novel ideas and new facts unearthed by authors; hold governments, businesses, and citizens accountable; contribute to a vibrant culture; educate, and inspire Americans of all ages; and report on scientific progress. Trustworthy Gen AI systems require high-quality new publications to remain state-of-the-art, and a flourishing publishing industry is best positioned to increase the value of Gen AI systems. A flourishing publishing industry will also help protect against some of the potential ills of Gen AI systems, including misinformation and bias.").

[40] See Tortoise Global AI Index at https://www.tortoisemedia.com/intelligence/global-ai/#rankings.

that a change in policy, such as by weakening copyright protections to permit the taking of copyrighted works without permission or compensation are gravely misplaced. It is precisely the U.S.' strong copyright and IP framework that is driving investment in content creation and curation, in research and development, and thus, fueling the U.S.' competitive advantage in AI development.

Ultimately, the cost to license copyrighted works to train Gen AI models is only one cost among many, and costs themselves are but one factor that influences the choice of business location. The U.S. excels in the key factors identified in the GAII scorecard, ensuring that it will remain an epicenter for AI development. Further, the U.S. could possibly enhance its competitive advantage by requiring AI developers to comply with robust U.S. rules protecting creators regardless of domicile.[41]

Conclusion

AAP welcomes this further opportunity to provide its views regarding the issues raised in the Copyright Office's Notice of Inquiry and Request for Comments. Our analysis indicates that current case law does not support the application of fair use to permit the wholesale reproduction of copyrighted works by Gen AI developers to train AI models. The un-permissioned use of copyrighted works to train Gen AI systems is also inconsistent with the policies animating copyright law and the objectives it is intended to promote. We appreciate the Copyright Office's forward-thinking research into the intersection of copyright law and AI and look forward to continuing to work with the Copyright Office on this important matter.

[41] E.g. A2IM-RIAA at 8-10 ("[T]o prevent AI developers from geo-laundering[,] . . . an AI developer should not be able to import an AI model into the U.S. that ingested copyrighted works without authorization by claiming that their AI development occurred in another jurisdiction where they claim the ingestion of copyrighted works is legal."); CCC at 2-5 ("[T]he United States can protect domestic consumers, creators, and businesses, including technology companies, by requiring compliance with domestic rules, regardless of the source or domicile of the AI developer or the AI entities developed.").