Pearson

Electronic submission via www.regulations.gov

October 30, 2023

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

Introduction

For many decades, Pearson has been a leader in educational publishing. Today, we are the world's leading learning company, with more than 160 million users of our products and services worldwide. Approximately 20,000 Pearson employees, together with our authors, are committed to creating enriching learning experiences designed for real-life impact. We are invested in the future of learning, from its ability to unlock opportunities for learners and societies globally to learning through innovation. Learning is who we are.

Pearson is a trusted brand that represents high quality product offerings and enriching, vibrant learning experiences. Our company's unparalleled collection of high quality, proprietary intellectual property assets strengthen our ability to deliver an unmatched experience for the learner across their lifetime of learning. With our unrivalled depth of content and data, we expect generative AI to create significant positive opportunities for Pearson. However, like all new technologies, the use of generative AI must be approached in a thoughtful and considered manner to avoid doing more harm than good. As learners and educators place enormous trust in Pearson, it is essential that our content and copyrights be protected. Pearson welcomes the opportunity to participate in the United States Copyright Office's notice of inquiry on generative AI. While there are other risks associated with the use of generative AI, including cybersecurity, data privacy, bias and unfairness, our responses below focus on and illustrate the importance of copyright in our protection of our valuable business assets.

Pearson's Business

Pearson's purpose is to provide life to a lifetime of learning. We do this by creating vibrant learning experiences for learners in many settings including primary and secondary schools, higher education, the workforce, professional assessments, and English language programs. We create reliable and up-to-date materials by collaborating with trusted authors and expert content creators such as teachers, faculty, and educational institutions. The high-quality learning experiences we provide include not only literary and artistic works, traditionally in written text, but also data in many forms, including digital books, digital learning environments, audio visual content, computer programs, databases, maps, and drawings. This effort takes considerable investment in both time and resources and results in highly valuable creative products that are protected by copyright.

The worldwide educational publishing market was estimated to be worth more than 6 trillion U.S. dollars in 2022 and is expected to grow to 8 trillion U.S. dollars by 2030.[1] It contributes hundreds of millions of dollars to the U.S. economy and supports millions of high value jobs. It is essential that this socially important and valuable industry enjoys the protections under the law to which it is entitled.

General Questions

1. As described above, generative AI systems have the ability to produce material that would be copyrightable if it were created by a human author. What are your views on the potential benefits and risks of this technology? How is the use of this technology currently affecting or likely to affect creators, copyright owners, technology developers, researchers, and the public?

We are cautiously optimistic about the potential of generative AI to improve education and learning, provided the appropriate steps are taken to mitigate the risks. We see potential benefits in personalized learning, increasing access and equity, improving outcomes, and in feedback and engagement. However, the risks of generative AI include cybersecurity, data privacy and the introduction of bias and unfairness, as well the delivery of faulty, unreliable data.

Unfortunately, the lack of transparency about generative AI training data and processes makes it difficult to know if a specific generative AI system complies with all relevant copyright law, other intellectual property laws, database rights, data protection laws, and international laws and treaties, such as the Berne Convention. It is essential that generative AI companies be transparent about the content and material used to train their systems, trace the provenance of such content by disclosing their processes, track the copyrighted works used, and retain the copyright management information of such content.

With regard to copyright interests in particular, most generative AI systems currently available have been trained using copyrighted content without the authorization of rights holders. These impermissible activities must be addressed not only because they use valuable content without the remuneration to copyright holders, but also because the unauthorized and unrestrained use of copyrighted materials disincentivizes innovation, discourages creation of new high-quality content, and results in inaccurate and less valuable machine-produced derivative content and works.

Fundamentally, the exclusive rights of copyright should not be ignored, and generative AI companies should not be allowed to avoid or circumvent the copyright or other intellectual property rights of others simply in the name of innovation. Copyright holders have in no way automatically relinquished any exclusive rights in their content or given up their right to consider market or licensing opportunities for their content in the generative AI context. It is critical that generative AI companies obtain the consent of the copyright owners for the use of their content to avoid infringement and the other negative outcomes discussed above. Pearson's content is much more than just the 1s and 0s of "data", it represents valuable creative assets

[1] *Global Education's $8 Trillion Reboot*, MORGAN Stanley (June 7, 2023), https://www.morganstanley.com/ideas/education-system-technology-reboot ("The global education market is slated to reach an estimated $8 trillion in value by 2030 from $6 trillion in 2022.").

developed by experts with substantial investment that generative AI developers seek to use to produce quality models and outputs.

2. Does the increasing use or distribution of AI-generated material raise any unique issues for your sector or industry as compared to other copyright stakeholders?

As the world's largest learning company, we see a variety of unique issues related to the increasing use or distribution of AI-generated material in the education and workforce space.

First, the use of AI systems for cheating and plagiarism is well-known. The availability of these tools for unethical purposes poses challenges not only for students, teachers, and the integrity of academic systems, but also undermines the usefulness and market for educational products. If AI companies negotiated appropriate licenses with educational publishers, we might be able to set limits on the amount of any one work that is ingested and the sort of output that the program could deliver, making AI systems less prone to serve as tools for cheating.[2]

Second, educational publishers like Pearson are particularly concerned with the reliability and validity of outputs generated by these generative AI systems. In the learning environment, high-quality, vetted content is critical to providing learners with positive outcomes. Likewise, accuracy and pedagogy are of paramount value. We understand that generative AI systems have used material they have sourced indiscriminately from the internet to train their models. The quality of this material, its accuracy, and its sources are unverified, and these systems return outputs that are inaccurate or misleading. To add to our concern, we have seen generative AI models produce inaccurate content and misinformation and attributed that content to Pearson products.

That generative AI models output incorrect responses and attribute them to authors and publishers who have spent significant resources in establishing their favorable goodwill and strong reputations in delivering quality educational material is alarming. It will remain critical in the learning industry that humans are involved in the creation, development, and deployment of generative AI systems to ensure responsible sourcing of training data and use of these technologies.

Third, we believe the scale of unauthorized copying of educational materials is likely greater than other forms of content. As discussed above, developers of generative AI systems use datasets comprised of data that has been indiscriminately sourced from the internet. They use automated web-crawling tools to ingest large volumes of data, including huge datasets of eBooks, often pirated, such as the "Books3" repository and others.[3] The unlawful presence of pirated material on the internet is a battle that educational publishers continue to fight, but the unchecked sourcing and scraping of educational content and material from around the internet to train generative AI models further propagates the harms of online piracy by dramatically expanding the reach of these pirated sources.

[2] To be clear, the opportunity to control the use or non-use of an entity's intellectual property, including whether or not to license its intellectual property, vests with the intellectual property owner.

[3] Alex Reisner, *What I Found in a Database Meta Uses to Train Generative AI*, THE ATLANTIC (Sept. 25, 2023), https://www.theatlantic.com/technology/archive/2023/09/books3-ai-training-meta-copyright-infringement-lawsuit/675411/.

Fourth, the use of generative AI risks undermining copyright's core value of promoting an incentive to create.[4] It is no stretch to believe that AI systems have been or will be used to create wholly AI-generated textbooks trained on and designed to replace those of responsible publishers like Pearson. Not only would such AI-generated content infringe rights in the underlying content on which such outputs produced and risk publication of unreliable content, but it would materially impact the incentive of true creators to create new works for the advancement of the arts and sciences.

4. Are there any statutory or regulatory approaches that have been adopted or are under consideration in other countries that relate to copyright and AI that should be considered or avoided in the United States? How important a factor is international consistency in this area across borders?

The United States should carefully consider other countries' approaches to copyright and AI and the implications those approaches have in the U.S. Of note are two international examples below.

European Union. The European Parliament's proposed EU AI Act would require providers of generative "AI Systems" to publish summaries of copyrighted data used for training the AI's model.[5] Specifically, providers of generative AI systems would need to "make publicly available a sufficiently detailed summary of the use of training data protected under copyright law."[6] Generative AI companies in the United States should face the same requirement.

Requiring transparency and adequate disclosures of training data from the developers of generative AI models is necessary for several reasons. First, it would help copyright holders protect their works. If generative AI models were required to provide a detailed summary of their training data, this would better facilitate a determination of whether copyrighted works have been used without authorization. Second, mandatory disclosure would likely help to produce AI models that are of superior quality. Currently, serious allegations have been made against generative AI models for sourcing third-party copyrighted materials

[4] *See, e.g.*, David De Cremer et al., *How Generative AI Could Disrupt Creative Work*, Harv. Bus. Rev. (Apr. 13, 2023), https://hbr.org/2023/04/how-generative-ai-could-disrupt-creative-work ("[G]enerative AI significantly changes the incentive structure for creators, and raises risks for businesses and society. If cheaply made generative AI undercuts authentic human content, there's a real risk that innovation will slow down over time as humans make less and less new art and content.").

[5] Eur. Parliament, *EU AI Act: first regulation on artificial intelligence* (June 14, 2023), https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. The European Union is currently considering the EU AI Act. The EU AI Act would cover any "AI system" that is developed by an EU-based provider, as well as any systems that are developed outside of the EU and placed on the EU market. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules of Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM (2021) 206 final (Apr. 21, 2021). However, the member states are still negotiating the definition of AI System. It remains to be seen whether there will be consensus or alignment on a singular definition or whether individual governments will be left to define AI Systems independently during transposition, potentially creating inconsistency in the Act's implementation.

[6] Eur. Parliament, *EU AI Act: first regulation on artificial intelligence* (June 14, 2023), https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

from pirated websites.[7] In addition to infringement, this type of illicit sourcing contributes to poor quality training data and reflects poorly on the rights holders of the copyrighted material that has been collected from unauthorized sources. Consistent, mandatory disclosure requirements would disrupt this unauthorized sourcing and help ensure that AI models train on high-quality, authentic, and authorized works. Consistent disclosure requirements and transparency surrounding the use of materials used in the training process would also improve the ability of users and government entities to scrutinize the training process. Importantly, such requirements likely would spur self-regulation by the AI companies to avoid running afoul of copyright law in the first place.

Under European Union copyright law, there are two exceptions for content scraping (sometimes referred to as text and data mining ("TDM")), including an exception for research and cultural heritage institutions and an exception that covers TDM for any purpose.[8] This second exception grants copyright holders the right to opt out of the exception and expressly reserve such use of works to themselves. While this approach reflects a compromise between the technology sector and rights holders, it does not, in our view, fully protect the rights holders. Rather, it places an impossible burden on rights holders to notify each and every AI company of its intent to opt out of content scraping. Essentially, it creates a new "formality" on rights holders and inverts the benefits of ownership. Copyright ownership does not require that the owner of that right affirmatively declare that a third party does not have permission to use the owner's copyrighted work. Instead, copyright ownership gives that rights holder the right to control use of their work - to decide whether or not to grant someone permission to use that work and whether to keep that work for their own use. No affirmative act by the copyright owner such as an "opt out" is currently required. A better approach would be analogous to Australia's copyright law, described below, which requires a generative AI company to affirmatively seek the permission of a copyright holder to use its content.

Australia. Under Australian copyright law, there is no exception for TDM for AI development. Accordingly, in Australia, to engage in TDM to train an AI model, an AI company must affirmatively seek and obtain the permission of a copyright holder. This approach recognizes the priority, import and breadth that should be placed on copyright, and it ensures that copyrighted materials are more fully protected, without putting the onus on the copyright holder to continue to justify and establish its rights each time an AI model is developed. This Australian law withstood lobbying efforts and claims that the approach was inflexible and would significantly damage and even paralyze innovation in Australia compared to other parts of the world. Notwithstanding this resistance, AI innovation does not appear to have been foreclosed or even been stifled in Australia. Thus, such concerns should be considered with extreme caution as generative AI models continue to forge ahead with development and deployment, even in jurisdictions that affirmatively require AI companies to obtain consent to ingest copyrighted materials.

[7] *See, e.g.*, Complaint, *Authors Guild v. OpenAI*, No. 1:23 cv 08292 (S.D.N.Y. Sep. 18, 2023); Complaint, *Chabon v. OpenAI*, No. 3:23 cv 04625 (N.D. Cal. Sept. 8, 2023); Complaint, *Silverman v. OpenAI*, No. 3:23 cv 03416 (N.D. Cal. July 7, 2023); Complaint, *Tremblay v. OpenAI*, No. 3:23 cv 03223 (N.D. Cal June 28, 2023).

[8] EU Directive 2019/790 (April 17, 2019).

8. Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use? Please discuss any case law you believe relevant to this question.

8.1. In light of the Supreme Court's recent decisions in Google v. Oracle America and Andy Warhol Foundation v. Goldsmith, how should the "purpose and character" of the use of copyrighted works to train an AI model be evaluated? What is the relevant use to be analyzed? Do different stages of training, such as pre-training and fine-tuning, raise different considerations under the first fair use factor?

8.2. How should the analysis apply to entities that collect and distribute copyrighted material for training but may not themselves engage in the training?

8.3. The use of copyrighted materials in a training dataset or to train generative AI models may be done for noncommercial or research purposes. How should the fair use analysis apply if AI models or datasets are later adapted for use of a commercial nature? Does it make a difference if funding for these noncommercial or research uses is provided by for-profit developers of AI systems?

8.4. What quantity of training materials do developers of generative AI models use for training? Does the volume of material used to train an AI model affect the fair use analysis? If so, how?

8.5. Under the fourth factor of the fair use analysis, how should the effect on the potential market for or value of a copyrighted work used to train an AI model be measured? Should the inquiry be whether the outputs of the AI system incorporating the model compete with a particular copyrighted work, the body of works of the same author, or the market for that general class of works?

While this inquiry asks about circumstances where the unauthorized use of copyrighted works to train AI models constitutes fair use, Pearson cannot point to any such examples with respect to its own works in the current AI landscape. Indeed, Pearson's position is that its copyrighted works cannot be used in AI training without our permission. Although fair use is a fact-specific inquiry, Pearson believes the case law fully supports this position, including the Supreme Court's recent decision in *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258, 1275 (2023). In *Warhol*, among other things, the Court warned that "an overbroad concept of transformative use, one that includes any further purpose, or any different character, would narrow the copyright owner's exclusive right to create derivative works."[9]

A lack of fair use in the AI training context is also supported by the fact that entire copyrighted works, if not entire pirate databases consisting of tens of thousands of entire copyrighted works, have been ingested into generative AI systems without permission. And still further support for a lack of fair use can be found in the commercial nature of most generative AI systems, the absence of a fundamentally different purpose or use from the educational purpose for which educational works are created, and the market harm from lost sales, lost licensing revenue, and the usurpation of the right to decide whether to exercise a market opportunity or not. While by no means an exhaustive list of concerns, these issues warrant careful attention by the Copyright Office to the assertions of AI developers and their supporters with respect to fair use,

[9] 143 S. Ct. 1258, 1275 (2023).

especially those that advocate for any type of blanket fair use exception in the AI training context, which is wholly unjustified.

11. What legal, technical or practical issues might there be with respect to obtaining appropriate licenses for training? Who, if anyone, should be responsible for securing them (for example when the curator of a training dataset, the developer who trains an AI model, and the company employing that model in an AI system are different entities and may have different commercial or noncommercial roles)?

As a publisher, Pearson works every day to negotiate and obtain appropriate rights to publish content. When it comes to generative AI development, this process should be no different. If a company developing a generative AI product wants to use Pearson's copyrighted works in its training data, just like any other licensing prospect, Pearson's rights include the right to consider whether Pearson has an interest in licensing such content and whether mutually agreeable terms can be reached. The companies developing generative AI are some of the largest and most profitable in the world. They should not be given free rein to ignore the exclusive rights that have been conferred on copyright owners to pursue and promote their creativity. Further, allowing generative AI developers to shift the licensing burden onto rights holders would allow developers to gain an unfair advantage at the expense of rights holders and other stakeholders and undermine the rights vested in those who create, which are founded in the Constitution and enumerated in the Copyright Act.

18. Under copyright law, are there circumstances when a human using a generative AI system should be considered the "author" of material produced by the system? If so, what factors are relevant to that determination? For example, is selecting what material an AI model is trained on and/or providing an iterative series of text commands or prompts sufficient to claim authorship of the resulting output?

The capabilities of generative AI are just beginning to be understood. Given that the AI landscape is still being charted, we believe it is premature to foreclose the prospect that a human using a generative AI system could be considered the "author" of material produced by that system in limited circumstances. The formulation of a non-exhaustive list of factors would be useful to assessing whether there is sufficient human contribution and creativity to a generative AI output for that output to be protected by copyright. Moreover, the focus should not just be on the output. A similar list to assess how input, which is the product of sufficient human creativity, can establish or support copyright protection in certain outputs from a generative AI system would also be useful.

Generally, we would not expect that the act of providing an iterative series of text commands or prompts is sufficient to claim authorship of the resulting output. However, just as the Copyright Office suggests that copyright can only protect material that is the product of human activity, we believe that further consideration should be given to whether a claim of authorship in output may exist where the input itself is a representation of the original intellectual conception of an author and, thus, subject to copyright protection. For example, when a generative AI system is used to translate or copy edit a copyrighted work, is that resulting output sufficient for copyright protection because the input is protected notwithstanding that the translation or copy editing is provided by a generative AI system? We believe that these examples tend to illustrate the potential need for flexibility and further consideration to be given in the context of determining what human activity is needed for copyright protection in a work that uses generative AI, so that copyright holders do not experience unintended consequences from too stringent an application of the human component.

22. Can AI-generated outputs implicate the exclusive rights of preexisting copyrighted works, such as the right of reproduction or the derivative work right? If so, in what circumstances?

AI-generated outputs can certainly implicate the exclusive rights of preexisting copyrighted works, such as the right of reproduction or the derivative work right. As a threshold matter, Pearson notes that this question does not ask whether copyright infringement can arise from inputs but rather focuses solely on outputs.

Copyright holders enjoy exclusive rights in copyrighted works, including the rights to reproduce, distribute, perform, and display copies of the copyrighted work and to prepare derivative works based upon the copyrighted work.[10] These rights do not change simply because the potentially infringing works were partially or wholly created by software. Instead, the *prime facie* question of infringement focuses on the facial content of the allegedly infringing work. If, for example, an AI model outputted material substantially similar to a Pearson textbook, there could be no doubt that such output would implicate (and violate) Pearson's exclusive rights in copyright. Similarly, if an AI model were asked to generate a sequel to a copyrighted work, and that sequel would be a derivative work had it been created by a human given the same instruction, then that output would implicate (and violate) the copyright holder's right to prepare derivative works.

In short, there is no reason that an otherwise infringing output would fail to be considered infringing simply because it is AI-generated, nor is there any reason to allow AI developers to escape accountability in this situation.

26. If a generative AI system is trained on copyrighted works containing copyright management information, how does 17 U.S.C. 1202(b) apply to the treatment of that information in outputs of the system?

Under 17 U.S.C. § 1202(c), "copyright management information" ("CMI") is information conveyed in connection with copies of a work (including in digital form) such as its title, its author, its owner, terms and conditions for its use, and other information. This definition includes, for example, an author's name appearing next to copyrighted images.

For good reason, the statute protects against the removal or alteration of CMI. It is a violation of § 1202(b) to, with reason to know that it will induce, facilitate, or conceal copyright infringement, remove, or alter the information from a work, distribute it knowing that it was altered without the copyright owner's authorization, or distribute copyrighted works knowing that the information has been removed or altered without the copyright owner's authorization.

[10] *See* 17 U.S.C. § 106.

As with other aspects of copyright law, the answer here is not materially transformed in the generative AI context. If a person were to read an article that included CMI and then impermissibly distribute copies of the article with that information removed, that conduct would violate § 1202(b). Likewise, if an AI model removed CMI in its ingestion of a copyrighted work of a third party or were trained on that work with CMI and then distributed a copy of the work with the CMI removed, such conduct would violate § 1202(b).

Thank you for considering our views.