WILEY

October 30, 2023

Rhea Efthimiadis
Assistant to the General Counsel
U.S. Copyright Office
101 Independence Ave SE
Washington, DC 20540

RE: Request for Information on Artificial Intelligence and Copyright [Agency/Docket Number - Notice: 2023-6; Document Number: 2023-18624]

Dear Ms. Efthimiadis:

Thank you for the opportunity to share our views on artificial intelligence and copyright. We appreciate this important feedback mechanism and look forward to working with the U.S. Copyright Office and other stakeholders to deliver meaningful input and help develop policy outcomes that will harness the benefits of AI while ensuring robust copyright protections.

Founded in 1807, Wiley is one of the world's largest publishers and a global leader in research and learning. For more than 215 years, we've consistently operated within academic integrity principles, which has solidified our trusted status among academic institutions. This deep and diverse involvement with the academic community positions us as a dependable partner that can help navigate the complexities of integrating AI into research and learning. We are proud to support researchers and learners by providing them with the highest quality content and services, helping to improve outcomes and livelihoods and support national productivity and innovation.

AI, which is inclusive of generative large language models, has the potential to dramatically transform the scholarly research and learning sectors. Today, it can unlock significant productivity improvements by detecting plagiarism, extracting critical information, classifying and recommending content, conducting technical and language checks, and bringing together related information from disparate sources. However, it can just as easily produce fake papers, falsify the peer review process, and erode the protection of intellectual property (IP) rights.

Wiley sees AI as having an immense potential to strengthen our ability to deliver trusted, high-quality knowledge and knowledge solutions. Through AI, we see potential opportunities to enhance how we help the world's top scholars, researchers, and experts share their knowledge more effectively, as well as the consumers of this knowledge to absorb and apply it more quickly.

It is our firm belief that in order to be truly effective, any policy framework that aims to address AI should:

- Respect existing intellectual property protections and require the use of licensing for copyrighted materials in the training of AI language models;

- Require transparency and accountability from AI tools to ensure their legal compliance, support accuracy, and reduce the potential for erroneous output;

- Include expanded national funding for AI research and development (R&D) that recognizes the cost of peer-review, editing, publication, distribution, and long-term stewardship of research; and

- Protect the integrity of research by incorporating defensive mechanisms that can identify fraudulent research accelerated by AI and produced under the guise of legitimacy by "paper-mills."

Our over 200-year history of weathering technological disruptions testifies to our unwavering commitment to quality and integrity. Together, we are creating the tools and infrastructure to advance research and learning in the 21st century, and ensuring this system is imbued with the values of rigor and integrity; academic freedom; openness; partnership; diversity, equity, and inclusion (DE&I); and respect for innovation, commercialization, and IP rights. In the future, AI technology may be able to independently conduct primary research, generate innovative insights, and disseminate its own findings. It may also enable personalized learning at a mass scale. In that world, Wiley's role as a publisher—providing editorial oversight and a seal of approval—is more crucial than ever before.

1. What are your views on the potential benefits and risks of this technology? How is the use of this technology currently affecting or likely to affect creators, copyright owners, technology developers, researchers, and the public?

Human Rights

Wiley shares the concerns that have been expressed globally by individuals, industries and governments around individual human rights and privacy impacts by generative AI tools. Use of these tools will ultimately depend on their providers adequately addressing these inherent risks. The safety and security limitations of generative AI are not yet fully understood, and legal frameworks need to catch up with the technology. If these challenges can be met, and the right balance of efficiency with human creativity can be achieved, generative AI could become as integral to scholarly publishing as the internet itself.

Humans in the Loop

Wiley believes AI, especially in the context of research and learning, must include humans as part of the equation. While AI has the potential to introduce efficiency gains, the core products, IP, and artifacts of our work will still require close human oversight and management. Human intelligence is critical to discerning the validity of conclusions, so there should be no AI 'black box' in the chain of scholarly discovery. People are a crucial part of assigning meaning to patterns identified by automated systems. We recognize the importance of humans to determine bias in AI and the importance of human assessments and decision making, too. This oversight ensures the authenticity and quality of our services, enhancing the value we provide to our customers. As we continue to navigate and incorporate these technological advancements, we will remain steadfast in our commitment to uphold academic integrity and trust.

Application and Use Cases

Since the introduction of large language models and their rapid dissemination, we have been evaluating the potential applications of generative AI tools in scholarly publishing. We have tested a variety of commercial and publicly accessible tools, such as ChatGPT and Google Bard, for potential use cases related to authoring, submission and reviews, editing and production, publishing, and discovery and dissemination. Based on this work, we've found that AI can potentially support:

- **Authoring** by helping evaluate writing quality and giving reasonable feedback; rewriting manuscripts with better sentence structure and vocabulary to improve readability; suggesting research topics; and producing initial drafts of plain language summaries for discoverability;

- **Submission and review** by helping an editorial assistant run initial screenings. However, there are some significant limitations;

- **Editing** by helping copy edit;

- **Publishing** by helping enrich, classify, and translate content; and

- **Discovery and dissemination** by enhancing the quality of search results while also introducing new ways of discovering information.

Content Integrity

Deficiencies in the accuracy and quality of ostensibly factual information used by various generative large language models have the potential to threaten and infringe upon people's rights and safety. It should not be the role of an AI to say 'who is right' in an academic debate, but to provide information based on state-of-the-art research, based on licensed use of the Version of Record (VoR), and reflecting any corrections or retractions. The VoR provides transparent, linked, and up-to-date access to all associated research artifacts, which will ultimately validate the integrity of the information. The VoR also provides transparent access to all of the applicable

publication ethics practices and standards. Those failing to ensure the authenticity of information with respect to both the inputs and outputs of AI models should be held to account. We are committed to providing access to the final, trusted VoR as the best way to accelerate open science, where applicable, as it pertains to AI.

Wiley continues to demonstrate our commitment to integrity of information through our ongoing efforts to prevent the publishing of harmful and factually incorrect research in institutionally recognized journals. We are ready and willing to work collaboratively to make this a reality using the systems we have built to support scientific collaboration, integrity, and the research enterprise.

"Paper-mills," which produce fraudulent and fabricated research under the guise of legitimacy, use AI to increase output of content that increasingly looks authentic. The quantity of fake papers from "paper-mills" is increasing, and AI's ability to rapidly produce content lacking peer-review could exacerbate the spread of misinformation. Wiley, along with other publishers, is developing machine-learning and AI tools to support existing manuscript screening checks for text similarity and identification of discrepancies or similarities in images or statistical data that could indicate potential ethical or reproducibility problems. For instance, in collaboration with the International Association of Scientific, Technical, and Medical Publishers (STM), Wiley is helping to develop the Integrity Hub, which combines shared data and experiences and technological innovation to detect research integrity issues. The hub includes a tool working to combat the efforts of "paper-mills" that aim to take advantage of the publishing industry.

While there will always be a need for human oversight of automated systems, these tools and services are increasing in sophistication and are beginning to be applied at scale. Routine checks and tools enable publishers to identify patterns of potential systematic manipulation of the publication process, allowing for actions to correct and amend the scientific record, upholding trust and credibility of information. However, the stakes for ensuring content integrity have never been higher due to the rate at which AI can produce information. To support content integrity efforts, the U.S. Government should develop and deploy measures needed for the purpose of auditing the information produced by generative AI for accuracy and impartiality. Appropriate funding should be allocated for these measures, coupled with close collaboration with private institutions, including publishers and creators.

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 global nature of today's technology ecosystem demands a coordinated policy response. The establishment of standards for AI, nationally and internationally, is vital for both the protection of individuals and for the protection of organizations using the technology. The cross-border nature of the digital economy makes it necessary for international AI regulatory frameworks and technical standards to apply between nations and regions. The United States should work with key domestic

and international partners to develop a shared vision for a risk-based regulatory approach for addressing AI challenges and advancing norms around responsible AI governance.

The U.S. Government should also designate or establish an organization tasked with the centralized oversight of issues related to AI. As part of this effort, the organization should be tasked with requiring developers of generative AI models to respect third-party IP rights, requiring auditing mechanisms to verify the content input into these models are indeed authorized for use and ensuring application of appropriate penalties for non-compliance with IP laws. For example, the Artificial Intelligence Act proposed by the European Union would require developers of generative AI to publish summaries of copyright materials used for the purpose of training AI language models. These actions, among others, would help to level the playing field with the technology sector and protect intellectual property rights while simultaneously ensuring the accuracy and reliability of information critical to U.S. innovation. In the meantime, we will continue our efforts to establish guidelines, monitor potential violations, and provide licensing opportunities, including for use in connection with AI, in alignment with our commitment to protect our users' privacy and our authors' IP.

5. Is new legislation warranted to address copyright or related issues with generative AI? If so, what should it entail? Specific proposals and legislative text are not necessary, but the Office welcomes any proposals or text for review.

Standards and Best Practices

Until governmental action is taken, questions regarding compliance and best practices remain unanswered. However, any efforts to govern AI must avoid blanket regulation that applies to all AI generally, but rather establishes individual standards and regulations for each type of AI. These include, but are not limited to: Artificial Super Intelligence, Artificial General Intelligence, and Artificial Narrow Intelligence.

To ensure legal protections are keeping pace with evolving technologies, Wiley joins others in calling on federal agencies to conduct comprehensive reviews of existing regulatory protections and enforcement authorities to guarantee the extent of their application to the use and development of AI. The rights of organizations must be protected, particularly with respect to copyright or IP infringement by AI models. Of note, these comprehensive reviews should include, but are not limited to, the U.S. Patent and Trademark Office, U.S. Copyright Office, the Department of Education, the National Institutes of Health; as well as the Copyright Act of 1976, and the Family Educational Rights and Privacy Act (FERPA).

We support recent efforts to establish shared resources for advanced AI research, such as the National AI Research Resource (NAIRR), which offers a path for leveraging the unique capabilities of the public, private, and academic sectors. By providing AI researchers with open access to computational resources and high-quality data, the NAIRR has the potential to democratize and enhance America's AI capabilities. The long-term success of the NAIRR can be enhanced by encouraging and enabling research partnerships between universities and the public and private sectors.

We also support the National Institute of Standards and Technology's (NIST) efforts to implement an AI Risk Management Framework that will enhance organizational AI governance. We also support efforts to develop standardized testing frameworks and benchmarks to evaluate the performance of AI systems.

Legislative and Executive Action

We encourage efforts to improve public transparency and provenance to enable users to understand and trace AI outputs to their sources, and for users and rightsholders to understand how AI models were trained. Given that AI technologies are being integrated into applications that will impact the lives and well-being of individuals, transparent disclosure of the datasets in which AI models are trained is crucial. As noted, AI technologies should be audited and required to maintain transparent records as to (a) the types of content or materials used to create the training datasets ingested by AI foundation models; (b) the extent to which the content originates from a trusted and accurate high-quality source; (c) whether the material is copyright protected, and if so, whether the use of the content is licensed and from whom; and (d) the assessment and audit processes conducted to test the readiness of an AI system for deployment.

If any executive or legislative action is taken, it should include language that allows for the application of existing and future laws to unforeseen developments in the technology within a reasonable manner of time. In doing so, there will be assurances that ongoing efforts will remain applicable and have a long-lasting impact on the protection of consumers and clarity on the obligations of private entities engaging in the development or use of AI.

6. What kinds of copyright-protected training materials are used to train AI models, and how are those materials collected and curated?

Wiley remains concerned with the lack of transparency and oversight regarding the potential for illegal accumulation and unauthorized use of restricted content as an input for model training purposes (and for that information to be provided to others), as this will create significant implications for IP protection. Moreover, the terms and conditions for using generative AI models lack clear and open transparency as they are determined solely by the service provider, without any consideration for the existing IP rights of others. For example, the information entered into and received from these models is subject to licensing terms with no opportunity to negotiate proper protections. The terms of use may require you to represent that you own or have legal rights to use the input with the service, which is lacking review and enforcement by the provider. The terms of use also do not warrant that the output is original and does not protect from third-party IP infringement claims, breach of confidentiality claims, or claims of violation of privacy or other applicable laws.

6.2. To what extent are copyrighted works licensed from copyright owners for use as training materials? To your knowledge, what licensing models are currently being offered and used?

As a leading publisher, we are key providers of information, materials, and data on which AI tools could be trained. By validating, normalizing, tagging and enriching content, delivering material in robust, interoperable and globally consistent formats, and creating domain-specific ontologies,

publishers ensure that information is a trustworthy high-quality input source with tremendous potential for use by AI systems across a broad range of applications. Wiley supports authorized licensing of content to AI developers on reasonable terms, providing access to valuable curated material on which to train trustworthy AI systems that yield verifiable and reliable outputs. However, we believe that generative AI models may be harvesting copyright protected material for training purposes while disregarding the existing intellectual property (including copyright), security, privacy and other restrictions on use of that data and information.

15. In order to allow copyright owners to determine whether their works have been used, should developers of AI models be required to collect, retain, and disclose records regarding the materials used to train their models? What level of specificity should be required? To whom should disclosures be made? What obligations, if any, should be placed on developers of AI systems that incorporate models from third parties?

Wiley acknowledges recent efforts of some AI developers to integrate policies and protective measures aimed at mitigating copyright infringement, such as the release of publicly accessible source code from AI developers to enable copyright holders to deter the extraction of proprietary data and information by large language models from their websites. While these good-faith efforts seek to reduce the impact of copyright infringement, they once more shift the burden onto copyright holders, requiring the rightsholder to conduct due diligence and be proactive in the pursuit of financial restitution through legal means. It is imperative that copyright protection be the affirmative duty of the provider to obtain proper rights for use of copyrighted materials before using them in large language models.

Although generative artificial intelligence represents a new technological frontier, it is important to underscore that the framework of copyright protection is well-established. In this context, AI developers should be held accountable for copyright infringements, in alignment with the longstanding legal precedents widely applied for decades.

AI developers should be required to rigorously assess the extent to which copyrighted materials are integrated into their datasets. In the event copyrighted materials are identified, AI developers should be required to obtain explicit rightsholder consent or expunge the information before incorporating the dataset as an input for the training of large language models. In the case of third-party organizations, their acquisition of datasets should strictly adhere to publicly available information that explicitly excludes copyrighted materials. In support of transparency, integrity, and protection of copyright, AI developers should also be required to publicly disclose, in an easily accessible format that collectively retains new and previous information, the origin of materials by which large language models have been trained and confirmation that the information generated by outputs is accurate and up to date. These disclosures should also include the source of the materials, the identification of applicable rightsholders, note of any existing licenses permitting use, and the date on which the information was collected. For circumstances in which copyright infringement has occurred, AI developers should be required to notify the rightsholder immediately and disclose the extent to which the copyrighted information was disseminated to unauthorized parties.

16. What obligations, if any, should there be to notify copyright owners that their works have been used to train an AI model?

Wiley recognizes the importance of using high quality, peer reviewed, vetted material to create the training datasets for AI models. Accountability and transparency policies which provide assurances that high-quality materials are used in training AI systems, without infringing on IP, builds confidence and trust in the technology. Fostering, encouraging, or requiring visibility into information sources would enable users and auditors to better tackle bias, to ascertain that a model was trained on information collected with the consent of those involved, and to ensure legal and regulatory compliance. Ultimately, the responsibility rests with the providers of AI systems, who should be required to support AI accountability and appropriate compensation to rightsholders throughout the value chain.

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?

Authorship

While generative AI has the potential to facilitate collaboration and improve the quality, reproducibility, and transparency of research work, its current propensity for inaccurate or biased results also introduces potential risks to research integrity. For this reason, accountability and transparency are central; only authors are accountable for their work, and any information or analysis generated by generative AI tools needs to be transparently described at the point of submission.

With respect to authorship, Wiley's policy is that generative AI tools cannot be considered capable of initiating an original piece of research without direction by human authors. Authors remain fully responsible for the accuracy of the information in their work, and use of AI tools must be disclosed and described, transparently and in detail, in the Methods or Acknowledgements section. Tools that are used to improve spelling, grammar, and general editing are not included in the scope of these guidelines. Our full policy is available on our website here: Best Practice Guidelines on Research Integrity and Publishing Ethics.

Concluding Thoughts

Wiley is committed to a future in which research on the development and use of AI is inclusive. Ultimately, our mission is to serve researchers, learners, and professionals. We must ensure that under no circumstances will people be put at risk, or that the quality of the works we publish and the valuable services that journals and societies provide to communities be compromised. To do so, the protection of intellectual property as it applies to AI language models is paramount.

As with any new technology, we must be both visionary and vigilant. We look forward to working with the U.S. Copyright Office and the wider scientific community on these issues and are committed to working collaboratively to develop forward-looking partnerships that strengthen research and innovation. In support of this endeavor, we have also provided a response to the White

House Office of Science and Technology Policy's Request for Information on National Priorities for Artificial Intelligence. The stakes for content integrity have never been higher, and we must leverage the entrepreneurial spirit of the research community and private sector to enable our country's continued leadership in the scientific enterprise while protecting intellectual property.

Sincerely,

Deirdre Silver
Executive Vice President, General Counsel
Wiley