STM Advancing trusted research

October 30, 2023

Suzanne V. Wilson, General Counsel and Associate Register of Copyrights
Maria Strong, Associate Register of Copyrights and Director of Policy and International Affairs
U.S. Copyright Office
101 Independence Ave. SE
Washington, DC 20559

RE: STM Response to Request for Information on Artificial Intelligence and Copyright (88 FR 59942, Docket No. 2023–6)

Dear Ms. Wilson and Ms. Strong,

STM, the International Association of Scientific, Technical and Medical Publishers, welcomes the opportunity to provide input to the U.S. Copyright Office on the broad and complex nexus of artificial intelligence (AI) and copyright. We look forward to working with the Office and with other stakeholders to support the development of an ethical, trustworthy, and human-centric AI ecosystem. For the scholarly publishing community, primary among our priorities in thinking about AI are preventing misinformation, respecting intellectual property, and supporting accurate, transparent, and reliable research and discovery processes.

STM and its member publishers understand that AI has the potential to radically change the way we work, learn, do research, and deal with information. It introduces both risks and rewards. STM and its members are actively considering how AI can be developed and deployed in an ethical, accountable, and trustworthy manner. Publishers are driving innovation in this space as well as contributing to AI development in at least two vital ways: to support research, research communication, and research integrity as well as by providing high quality, validated inputs for machine learning that improve the integrity and quality of AI tools.

AI is already being used in scholarly publishing to improve the efficiency of research and discovery, to enhance research integrity, and to identify promising areas of new research. It is being deployed internally by publishers to improve the management of articles and the process of peer review, to detect potential research misconduct, and to improve research recommendations. It is being used externally to support researcher efficiency and to help identify promising avenues of study. The future potential applications of AI for scholarly publishing are exciting. AI applications have the capacity to go beyond testing hypotheses against vast amounts of data to also creating new hypotheses, developing new theories, and exploring new connections.

Scholarly publishers are engaged with AI as key providers of high-quality content, context, and data that are used as inputs for machine learning and as training sets and information for AI. To support the availability and accuracy of these inputs and, therefore, AI trained on this material, STM's members make their content available under licensing schemes that support transparency and quality.[1] The curation and tagging services performed by publishers are the bedrock ingredient for high quality, trustworthy, and ethical outputs.

Providing this corpus of information in usable digital formats is a core expertise of publishers. STM's member publishers validate, normalize, tag, and enrich content. They deliver material in robust, interoperable, and globally consistent formats, and they create domain-specific ontologies. Publishers ensure that information comprises a trustworthy high quality input source with tremendous potential for use by AI systems across a broad range of applications. Importantly, STM publishers increasingly publish copyrighted works and associated datasets with AI ingestion technologies in mind. In other words, copyrighted content including information that could be accurately read and "ingested" appropriately without depending on a discursive narrative are published in ways that facilitate machine learning, and can be repurposed as high quality input for AI. Regardless of the purpose of an AI tool, the accuracy of the scientific record maintained by science and academic publishers helps to ensure that machine learning has both depth and accuracy.

STM also acknowledges the risks inherent in AI. STM and its members are working to address the potential for AI to spread misinformation, bias, and error throughout the scholarly record when fake papers, bad data, manipulated images, and inappropriate interventions in peer review are used to train AI systems. As part of this effort STM published a White Paper in May 2021 outlining best practice principles for an ethical, trustworthy, and human-centric AI ecosystem. Those principles remain relevant today and can be found here. Also relevant to the present Notice of Inquiry is STM's submission to the 2020 U.S. Patent and Trademark Office Notice of Inquiry regarding the need for intellectual property protection in the context of AI innovation. Overall, STM believes that the following guidelines are critical for any AI policy framework that seeks to balance the opportunities and risks of these technologies. The policy must:

- Respect and support existing intellectual property protections, including copyright, to ensure the continued development and availability of high quality, vetted content produced by researchers and publishers. Policy should make a clear distinction between raw data, structured data, and copyrighted works, and require the use of licensing for copyrighted content which, in turn, supports accountability and transparent provenance for key information such as training materials used by AI tools.

- Leverage and build upon existing efforts, including those of STM and our members, to promote high quality, accurate communications about research advances and prevent the introduction of misinformation, bias, and error. This will require engagement with stakeholders and

[1] *See, e.g.*, STM's Submission to the U.S. Patent and Trademark Office, Request for Comments, 84 FR 58141, pp 58141-58142: Intellectual Property Protection for Artificial Intelligence and Innovation, Jan. 10, 2020, at 1, available at https://www.uspto.gov/sites/default/files/documents/International%20Association%20of%20Scientifi_RFC-84-FR-58141.pdf.

consideration of how the introduction of AI affects the scholarly communication ecosystem.

- Encourage and, where possible, require public transparency and provenance of training data and content to enable users to understand and trace back results to their sources; and for users and right holders to understand what sources were used in training an AI system. In the scholarly sector, this could include requirements that an AI system and its outputs be placed into a chain of evidence such that the results can be reproduced and verified.

- Clarify that accountability lies in all providers and developers of AI solutions, including generative AI models, and through the whole life cycle of the AI. Providers should keep all records, documentation, and associated metadata for the duration of the existence of the AI system to support AI accountability throughout the value chain.

- Work to support and fund efforts to protect the integrity of research and education, which may be particularly vulnerable to misinformation or misinterpretation of AI outputs. In particular, a government policy should support the development and implementation of mechanisms to identify outputs (whether quantitative, visual or textual) created by AI. In the scholarly sector, these tools could include the detection of fake (synthetic) or manipulated image and data, identification of paper mills, and prohibitions against the use of inaccurate article versions or illegally sourced content.

Consistent with these guidelines, STM offers specific responses to relevant questions raised in the NoI, which are detailed below. STM and its members stand ready to work with the U.S. Copyright Office and other stakeholders to further support policy development that fosters the potential benefits of AI while reducing the risk of institutionalizing misinformation, bias, and error in these systems.

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?

The potential benefits of AI are potentially extraordinary. STM and its members currently use AI to improve the efficiency of research communications and discovery, to enhance research integrity, and to identify promising areas of new research. Some examples include:

- Evaluating and improving the readability of manuscripts. This could help researchers whose first language is not English write clearer papers, faster – advancing the progress of science and improving equity in the global scholarly community;
- Assisting researchers in literature reviews by analyzing the existing record and identifying relevant material, previously undiscovered relationships, knowledge gaps and research opportunities;
- Assisting authors to finding the most appropriate publication outlet for their manuscript; and editors the most appropriate reviewers;

- Assisting reviewers with tools that can detect plagiarism, examine the quality of citations, references and other metadata *e.g.*, tools that use machine learning to scan papers and identify text within the publication that indicate its integrity and reproducibility;
- Assisting editors with initial screenings for relevance to the journal's scope;
- Assisting authors and publishers by summarizing what can be highly technical research findings in terms understandable to a non-specialist audience. This can make knowledge more accessible to the general public as well as decision makers working with the findings to develop or adjust policy and legislation.

In addition to the potential benefits, AI also comes with risks that are real and potentially harmful. Some of those risks are relevant to the exercise of intellectual property rights and the functioning of law enforcement. Inaccurate data, bias, manipulated images, and other threats to the integrity of the scientific record existed prior to the emergence of AI and are exacerbated by it. The use of AI trained on erroneous or biased input can lead to equally flawed research outputs and introduce new inaccuracies and fresh biases to the scholarly record. Generative AI systems trained on materials other than the Version of Record pose a threat to scientific integrity as they may perpetuate and spread flawed, biased, and inaccurate outputs. Such outputs can contaminate the research landscape with disinformation and undermine the credibility of authentic works of scholarship, thereby endangering public health, safety, and security. There is also the potential for bad actors to use AI to flood the system with questionable but difficult to detect fraudulent data or information.

Most researchers and publishers are actively exploring how best to maximize the benefits and mitigate the risks of AI – and expect that their understanding of how to achieve each will evolve and improve over time. Fundamental to ensuring that AI systems have the highest quality and most trustworthy material on which to train are the incentives that intellectual property rights in general, and copyright in particular, offer for their creation, curation, and enhancement. STM is concerned that certain developers of AI systems are ignoring, and may continue to ignore, calls to respect the basic rights inherent in copyright, which require permission and/or compensation for use and re-use of content. The interests of those who seek to use copyrighted materials to train AI systems must not be prioritized over the rights copyright holders – especially when practical licensing solutions are readily available in the market. Licensing and respect for copyright already drive major sectors of the U.S. economy and can serve to power the progress, accountability, transparency, and accuracy of AI systems as well.

As the Office considers copyright in the context of AI, STM believes the following are key considerations:

- The safety and security limitations of generative AI are not yet fully understood. If these challenges can be met, and the right balance of automation and human creativity can be safely achieved, generative AI could become integral to scholarly publishing.
- Human intelligence remains critical to determining the validity, relevance, and significance of research reports. There should be no AI "black box" in the chain of scholarly discovery. People are a crucial part of evaluating the meaning of patterns identified by AI systems.
- Deficiencies in the accuracy and quality of information provided as outputs by various large language models (LLMs) have the potential to threaten public safety and infringe rights. While it should not be the role of AI to verify whether one party is correct or incorrect, there should be a

traceable chain of information to help users assess the accuracy and credibility of outputs. To obtain the best results, AI should provide information based on state-of-the-art trustworthy research, which in turn should be based on licensed use of the Version of Record (VoR).

- To foster trust in the outputs of AI systems, STM recommends government measures that enable auditing the information produced by generative AI for accuracy and absence of bias. Appropriate funding should be allocated for these measures, coupled with close collaboration with stakeholders like publishers.

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?

STM and its members have a specific interest in the integrity of the scholarly record, and they invest significantly in its creation, maintenance, and enhancement. Indeed, STM's mission is to advance trusted research worldwide. AI has the capacity to amplify the activities of bad actors who seek to spread misinformation or otherwise populate the scholarly record with inaccuracy and bias. Researchers and publishers are especially concerned that the data and content used to train AI models is accurate and unbiased, as errors or biases in AI-generated content can have serious research and ethical implications and could quickly erode the public's trust in science. The investments that STM and its members make to ensure that high-quality content continues to be vetted, produced, maintained, and enhanced rest on the foundational incentives of copyright.

Another risk that flawed AI systems bring to the educational and academic publishing sector is a dip in pedagogical rigor and quality. Ensuring AI developers respect intellectual property laws will be critical to the ongoing sustainability of educational and academic publishing. In this context, the quality of AI inputs is extremely important. It depends on ensuring that copyrighted materials are licensed from right holders or those right holders' trusted authorized agents. Professionally edited and peer-reviewed scientific and scholarly publications are vital to developing trustworthy AI systems. These copyrighted works, to which publishers dedicate countless hours of time and unparalleled technical expertise – and for which they take legal responsibility – must be properly licensed.

3. Please identify any papers or studies that you believe are relevant to this Notice. These may address, for example, the economic effects of generative AI on the creative industries or how different licensing regimes do or could operate to remunerate copyright owners and/or creators for the use of their works in training AI models. The Office requests that commenters provide a hyperlink to the identified papers.

Committee on Publication Ethics (COPE), *Artificial intelligence (AI) in decision making*, https://doi.org/10.24318/9kvAgrnJ

AI Ethics in Scholarly Communication (STM), https://www.stm-assoc.org/2021_04_29_STM_AI_White_Paper_April2021.pdf

Global Publishing and Journalism Organizations Unite to Release Comprehensive Global Principles for Artificial Intelligence (AI), https://www.stm-assoc.org/global-publishing-and-journalism-organizations-unite-to-release-comprehensive-global-principles-for-artificial-intelligence-ai/

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 transparency requirements being considered in the EU AI Act are a vital consideration for U.S. policy makers. Publishers and researchers should be able to know whether and when their material is being ingested by a generative AI system so that they can best support the appropriate use of content and the accuracy of generative AI outputs. Regulatory frameworks should enable rightsholders to object, suggest a license be agreed, or pursue legal remedies, if needed, to quell downstream iterations of their content that infringes their rights. Europe's risk-based approach to AI is worth consideration, ensuring that the most stringent requirements are being applied to AI systems that pose the highest risk, and we recommend that requirements for respecting copyright be universal. The global nature of today's technology demands a coordinated policy response among governments. The United States should work with the EU and other governments to develop a shared vision for a risk-based regulatory approach for addressing AI challenges and advancing norms around responsible AI governance.

Publishers need to know whether and to what extent AI has been used to generate content they are considering publishing. This is the only way they can assess whether individual parts or even the entire work they are considering does or does not enjoy copyright protection, among other qualities. Beyond that aspect, transparency is also important to identify liability risks. AI produces unpredictable output, and the output may include copyright infringements of third party works. Any party's use of AI to produce content or to train an AI must be made transparent. This is the only way right holders can evaluate whether they need to put safeguards in place to mitigate legal and ethical risks.

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.

The area in which STM views new legislation as potentially important is with respect to transparency and accountability. Especially at this juncture in time, when fair use case law is undecided, it seems clear that knowing whether and to what extent copyrighted content comprises training data for an AI is the only logical starting point for determining a range of issues, including but not limited to copyright infringement, inappropriate use of personal data, and bias. While STM does not believe new copyright legislation is necessary at this time, it does believe that the establishment of transparency standards is vital for both the protection of individuals and organizations, and to enable the enforcement of existing rights. Transparency will be the bedrock for market-based licensing solutions and should apply between nations and regions.

Federal agencies could conduct comprehensive reviews of existing regulatory protections and enforcement authorities to guarantee the extent of their application to the use and development of AI. Transparency in the disclosure of datasets used in AI models will be crucial.

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

STM understands that AI models are trained on all manner of copyright-protected content, including content that is not lawfully accessed, not properly licensed, or whose licensing terms have not been respected. STM is concerned with the lack of transparency and oversight regarding the accumulation of unauthorized use of restricted content as an input for generative AI training purposes.

As far as properly licensed training materials, STM's member publishers have collected and curated such materials and successfully licensed them for several years for various purposes and technologies, including text and data mining (TDM). Where those licenses did not include AI uses specifically, in some cases, they are being treated as templates for licensing materials for generative AI. Generative AI models are trained on a variety of copyright-protected materials across content types, including text, code, images, audio, video. These materials are collected and curated in a number of ways. Some companies have their own teams of experts who collect and curate training data. Others use crowdsourcing or scraping software to collect data from users or from third party datasets. Once the training data has been collected or acquired, it is often processed and curated to ensure that it is high quality and relevant to the task that the generative AI model is being trained to perform. This may involve removing duplicate data, correcting errors, and labeling the data. Some examples include:

- Text generation models: These models are trained on large corpora of text, such as books, articles. The models learn to identify patterns in the text and to generate new text similar to the text that they were trained on.
- Image generation models: These models are trained on large datasets of images, such as photographs, illustrations, and diagrams. The models learn to identify patterns in the images and to generate new images similar to the images that they were trained on.
- Code generation models: These models are trained on large datasets of code, such as software applications, libraries, and frameworks. The models learn to identify patterns in the code and to generate new code similar to the code that they were trained on.

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?

STM's members are key providers of information, materials, and data on which AI tools could be trained. By validating, normalizing, tagging, and enriching content, in addition to delivering material in robust, interoperable, and globally consistent formats as well as 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. STM and its members are extremely concerned that existing generative AI models are harvesting copyright-protected material for training purposes while disregarding existing copyright protection (along with security and privacy restrictions). Without proper laws and regulations tracking intake and mandating transparency, it will not be possible to fully gauge whether a copyright violation has been made, thereby undermining the potential of a vibrant licensing market for high quality, enriched, and bias-free training materials.

We note that STM's member publishers are already licensing their databases for use in text and data mining and that these license structures could potentially be a useful template for licensing to generative AI models and systems, where they do not include such uses already.

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.

Fair use is a fact-specific analysis. STM does not propose any specific circumstances under which an unauthorized use of copyrighted works to train AI models would constitute fair use, other than to point to the existing case law underpinning Section 107. STM does not believe that U.S. fair uses cases lodged at the time of this submission in the AI space constitute fair uses of copyrighted works. We also point out that several generative AI systems have included pirated content in their training datasets, which does not constitute a fair use of the underlying copyrighted works.

STM recommends that any use of copyrighted works to train AI models be licensed, both to comply with copyright and rightsholders' desire for the use of their works, as well as to enable transparency and accountability. This will also promote the reliability and accuracy of these tools.

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?

In the *Google* case, Google's use of the Java APIs was found to be transformative. The Court concluded that Google copied only what was necessary to allow programmers to work in a different computing environment but with a familiar programming language. Given the Court made a hard distinction between interface declarations and the implementation code, STM does not view the ruling as having impact on software code generally, or on other copyrightable works.

The *Warhol* case underscores the proposition that transformativeness under the first fair use factor demands a higher threshold of transformation than some recent lower courts have interpreted, especially when the follow-on work occupies the same commercial sphere as the original work. Each fair use analysis is specific to the facts at hand, so while pre-training and fine-tuning may be less relevant in some cases, they will be very relevant in others.

STM would like to underscore, however, that transformativeness, which was heavily emphasized in the Second Circuit's decision in *Authors Guild v. Google*, 721 F.3d 132 (2nd Cir. 2015), should not be presumed to apply in an analogous way to LLM ingestion. The "Google Books" decision, like all fair use decisions, was limited to the fact pattern at hand involving the making available of "snippets" of books. There was also a strong focus on the security of Google Books' systems such that copyrighted content in its inventory was deemed safe from downstream piracy; this is absent in the existing generative AI cases lodged in the courts. In particular, generative AI may inadvertently surface copyrighted content in its outputs. Three years after the Google Books decision, the Second Circuit decided that a service provider's unauthorized copying and re-distribution of audiovisual content deprived the content producer of access to revenue that properly belongs to the copyright holder. *Fox News Network LLC v. TVEyes, Inc.*, 883 F.3d 169 (2d Cir. 2018). For STM and its member publishers, our high quality, curated, and trusted content is highly valuable and protected by copyright. Using this material as training data for AI must be licensed and compensated.

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?

Like any fair use analysis, the four-factor test must be applied in a fact-specific way and in line with existing case law and the principles discussed above.

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?

STM believes it does make a difference and that fair use jurisprudence is robust enough to ensure that such facts would be included in a fair use analysis. Public-private partnerships of a variety of natures abound and STM believes these arrangements are often mutually beneficial, as well as important for the public. Regardless, to the extent a fair use analysis based on a nonprofit motivation becomes a for-profit project, any fair use analysis of the use of copyrighted materials at any point in the project, before or after it changed profit character, should be authorized and/or remunerated.

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?

In STM's view, the existing fair use doctrine provides a lens through which an analysis of this question can be undertaken. However, even if the entire menu of existing online content is ingested by an AI, an individual author's exclusive rights should not be overridden merely because the comparable percentage of his work in an overall infringement scheme is small; that would have the perverse effect of excusing largescale infringement for the very reason of its volume.

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?

All of these inquiries should be made. Fair use jurisprudence demonstrates that the specific facts in a given case all have some degree of relevance and the inquiry should incorporate all of these lines of questioning. STM would like to underscore that its members' corpora of training data are incomparable in their accuracy and quality and that there is a tremendous actual and potential market for this material.

9. Should copyright owners have to affirmatively consent (opt in) to the use of their works for training materials, or should they be provided with the means to object (opt out)?

Copyright owners should have to affirmatively consent to the use of their works. STM is aware of the EU's 2019 Copyright Directive, Articles 3-4, under which right holders are required to affirmatively opt out of their works being mined for commercial purposes.[2] STM is proactively working with its members on a seamless technological opt-out mechanism that will permit right holders to benefit from Art. 4. Notwithstanding the European approach, STM recommends that right owners should be able to consent to any use of their works. The Berne Convention's prohibition on any formality requirement renders eligible works copyrightable as soon as they come into being. This clearly means that all rights – including reproduction in the course of machine ingestion and maintenance – remain in the right holder's control without any action on the right holder's part. While STM recommends an opt-in over an opt-out mechanism, to the extent any opt-out mechanism were considered in the United States, it must include the guardrails included in the European framework (both the 2019 Copyright Directive and the proposed rules for an AI Act), including but not limited to that a) the content or other subject matter be lawfully accessed prior to any exception being applicable; b) right holders should be allowed to apply measures when there is a risk that the security and integrity of their systems or databases could be jeopardized; c) copies made for the purpose of TDM should be stored in a secure environment; and d) the technology companies should be transparent with respect to the material their AI systems ingest.

An opt-in approach has another advantage – mitigation of privacy issues. Sensitive research and corollary data including personally identifiable information (PII) from human subjects is vulnerable to scraping. Guardrails are imperative to protecting vulnerable populations from having their personal information improperly reused by generative AI. The best assurance that copyrighted research and data are not at risk is to introduce and maintain an opt-in mechanism under which right holders retain control of their research, publications, and data.

9.1. Should consent of the copyright owner be required for all uses of copyrighted works to train AI models or only commercial uses?

Consent should be required for all uses of copyrighted works. While STM is working on seamless technological identifiers to implement Article 4 of the 2019 EU copyright directive, we highlight for purposes of this inquiry that the system envisioned by Articles 3 and 4 did not take into account (nor could it have) the rise and scale of generative AI.

9.2. If an "opt out" approach were adopted, how would that process work for a copyright owner who objected to the use of their works for training? Are there technical tools that might facilitate this process, such as a technical flag or metadata indicating that an automated service should not collect and store a work for AI training uses?

[2] DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. Art. 4, *Exception or limitation for text and data mining*. 1. Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining. 2. Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining. 3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online. 4. This Article shall not affect the application of Article 3 of this Directive.

STM does not support the introduction of an opt out approach for generative AI training. In addition to the fact that it ignores the basis of the law's exclusive rights, it introduces practical problems that do not currently have a solution, i.e., many generative AI systems have scraped and ingested entire corpora of scientific journals from pirate websites. These pirated articles exist without permission and are being scraped and ingested without permission, carrying with them the potential for transmitting malware, ransomware, and viruses. An opt out mechanism only exacerbates this problem. No technological solution exists at this time that can detect when an AI system scrapes by ignoring or removing "do not scrape" metadata. STM is working on an effective metadata protocol at the article level to complement the existing W3C protocol, which applies at platform level, to ensure that articles separated from the right holder's website carry the relevant "do not scrape" metadata wherever they may appear on the web. However, this effort is meant only to address Art. 3 and 4 of the DSM with respect to text and data mining, and it is not clear that technological scraping tools respect such markers. In sum, technological tools do exist or are being devised for the TDM context, but at present they place significant burdens on right holders to protect their content, and it is not clear that they have been respected or will be in the future.

To the extent any opt-out mechanism were considered, it must include the guardrails included in the European framework, including but not limited to a) the content or other subject matter be lawfully accessed prior to any exception being applicable; b) right holders are allowed to apply measures when there is a risk that the security and integrity of their systems or databases could be jeopardized; c) copies made for the purpose of TDM should be stored in a secure environment; and d) technology companies should be transparent with respect to the material their AI systems ingest.

Record-keeping and transparency regarding training materials is key to ensuring trust – both from right holders and end users – and to ensuring excellence. An opt-out system cannot be considered without ensuring a level of transparency that ensures licensing conditions can be ascertained and preexisting rights respected.

9.3. What legal, technical, or practical obstacles are there to establishing or using such a process? Given the volume of works used in training, is it feasible to get consent in advance from copyright owners?

In STM's experience, it has been difficult to implement an opt-out mechanism that is both effective and efficient. Where right holders have developed such processes (e.g., W3C), they are not yet widely known or used by technology services, which tend to create their own mechanisms in preference to the needs of right holders. In today's environment, licensing in advance from copyright owners is readily available, and STM publishers successfully license content for a variety of purposes, including for AI training. Given the size and scope of content used by AI systems for training, and the lack of transparency about what has been used, it is extremely burdensome, if possible at all, for right holders to ascertain whether their content has been included in any dataset. With no advance consent from copyright owners, businesses risk operating in a context of legal uncertainty, sometimes leading them to favor lower-quality inputs. In the STM context, these low-quality inputs may undermine the integrity of the scholarly record and even lead to dangerous misinformation such as inaccurate information about public health and safety.

9.4. If an objection is not honored, what remedies should be available? Are existing remedies for infringement appropriate or should there be a separate cause of action?

A copyright owner could sue an AI developer for infringement if the developer uses the owner's work to train an LLM without permission. If the copyright owner is successful, remedies may include damages and/or injunctive relief. However, these remedies may not be adequate to deter AI developers from using copyrighted works without permission. Additionally, as AI tools grow in sophistication and complexity, it is expected that it will be more difficult to ascertain violations. This underscores the necessity of transparency at the ingestion stage.

With the exponential growth of use of AI tools, infringement may not be clear until a user applies the AI tool and makes available the outputs. Right holders may then be in an untenable situation of having to track an unknowable number of users, developers, and service providers, as well as taking action against the developer or provider. This creates an uncertain legal environment for AI developers and providers as well.

9.5. In cases where the human creator does not own the copyright—for example, because they have assigned it or because the work was made for hire—should they have a right to object to an AI model being trained on their work? If so, how would such a system work?

In the scholarly publishing space, publishers, when they are the copyright holder of a given work, work with authors and researchers to represent their interests and are also in position to make objections and other legal claims on authors' behalf when appropriate.

10. If copyright owners' consent is required to train generative AI models, how can or should licenses be obtained?

There could be many avenues for this in the United States and around the world. Direct licensing, either through a right holder or through authorized agents, is STM's recommended solution for this and already drives a thriving market with respect to text and data mining. There is a growing demand for non-consumptive use in TDM, and several cross-publisher services have been developed with a variety of offerings and features. Such infrastructures could potentially be repurposed for providing training data for large language models.[3] Open licenses may not be sufficient to provide consent for training, especially where the training process or the AI tool itself may not be able to abide by the conditions of those licenses.

10.1. Is direct voluntary licensing feasible in some or all creative sectors?

[3] "Aggregator-provided services include TDM Studio (ProQuest), Digital Scholar Lab (Gale), Nexis Data Lab (LexisNexis), HathiTrust Research Center (Indiana University and University of Illinois with HathiTrust), Constellate (JSTOR). TDM services and tools support non-consumptive text analysis by providing secure virtual environments and access to right-cleared and public domain materials. Some also provide analysis and visualization tools. Crossref TDM API (https://www.crossref.org/documentation/retrieve-metadata/restapi/text-and-data-mining) allows researchers to harvest full-text documents from participating members, regardless of whether the content is open access or subscription." *See* Oya Y. Rieger and Roger C. Schonfeld, ITHAKA S+R, *Common Scholarly Communication Infrastructure Landscape Review*, April 24, 2023, at 18-19. https://sr.ithaka.org/wp-content/uploads/2023/04/SR-Report-Common-Scholarly-Communication-Infrastructure-Landscape-Review042423.pdf

Speaking for the academic/scientific/medical publishing sector, direct voluntary licensing is not only feasible but already pervasive in our sector for a variety of types of uses.

10.2. Is a voluntary collective licensing scheme a feasible or desirable approach? Are there existing collective management organizations that are well-suited to provide those licenses, and are there legal or other impediments that would prevent those organizations from performing this role? Should Congress consider statutory or other changes, such as an antitrust exception, to facilitate negotiation of collective licenses?

Voluntary collective licensing may have a role to play in the generative AI environment at some point in the future. It is already possible for right holders to engage in this activity without legislative change or government intervention. STM does not see a need for either at this time.

10.3. Should Congress consider establishing a compulsory licensing regime? If so, what should such a regime look like? What activities should the license cover, what works would be subject to the license, and would copyright owners have the ability to opt out? How should royalty rates and terms be set, allocated, reported and distributed?

No. STM and its members do not support compulsory licensing schemes.

10.4. Is an extended collective licensing scheme a feasible or desirable approach?

STM does not consider ECL a feasible or desirable approach at this time.

10.5. Should licensing regimes vary based on the type of work at issue?

Licensing regimes have the salutary quality of being nimble and tailorable based on the type of work at issue and any other specifications the parties so choose.

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)?

STM notes that although the roles in the question above (developers, curators, companies) are evolving, each participant involved in generative AI development should be legally responsible for any actions that interface with copyrighted works.

12. Is it possible or feasible to identify the degree to which a particular work contributes to a particular output from a generative AI system? Please explain.

The context of generative AI does not necessarily change the way in which infringement is ascertained. If access to copyrighted works is established along with substantial similarity, infringement likely exists. This fact underscores the importance of transparency and fostering an environment for a healthy licensing regime.

13. What would be the economic impacts of a licensing requirement on the development and adoption of generative AI systems?

The impact will vary depending on the AI system and the works, content, and/or data being sought for ingestion or training purposes. STM's member publishers have been licensing content and data for machine learning for several years and are proud to be contributing to important AI applications grounded in high quality training materials. However, STM underscores the importance of licensing the peer-reviewed and vetted Version of Record (VoR). Only the VoR of a given scientific, technical, or medical journal article is supported by the publisher both before and after its publication. This means that any changes needed to published material that are vital to maintaining the integrity of the scientific record are made. Training generative AI on outdated or incorrect research reports has the real-world effect of perpetuating scientific and medical errors, to the detriment of society at large.

14. Please describe any other factors you believe are relevant with respect to potential copyright liability for training AI models.

STM would like to underscore the importance of transparency to the trustworthiness and quality of the generative AI ecosystem. Without having a clear provenance in training data and content, it will not be possible to determine, e.g., whether the training materials were ever vetted for toxicity, bias, and discrimination, or whether rights have been respected. One is vital to public health and safety and the other to ensuring that a trusted scientific record is used to train AI systems. A trusted scientific record depends on a legal system that protects that record.

Transparency & Recordkeeping

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? Should creators of training datasets have a similar obligation?

Yes. One of the issues publishers face in the United States and elsewhere is how opaque many AI firms are in relation to the way their models are trained, and on what content. Unless and until there is more transparency on this – potentially through regulation or legislation – appropriate enforcement of intellectual property rights will not be possible.

Such a system could function in a variety of ways. For example, AI developers and creators of training datasets could be required to submit their records to a central database or repository. The metadata could then be used to identify the copyright holder and the AI developer or repository could, from that point, be required to provide the right holder direct notice of the use of their works. AI developers and creators of training datasets should be required to send copyright owners a copy of their records or to provide them with a link to a database where they could access those records. AI developers and providers could also disclose basic but relevant information on input data to the public, with right holders or others with legitimate interest having the right to receive more details upon request.

15.1. What level of specificity should be required?

A right holder should be informed whether his/her/its work is included in training data. If materials have been licensed, the source of the license should be included as well.

15.2. To whom should disclosures be made?

To right holders and to potential users of such AI models.

15.3. What obligations, if any, should be placed on developers of AI systems that incorporate models from third parties?

In the interests of transparency, accountability, and risk mitigation developers of AI systems should be required to disclose the identity of the third-party provider; the training corpora used to train that third party system; and any identifiable risks associated with the training corpora. Transparency obligations applying upstream to AI developers will also enable compliance downstream by other developers and distributors of AI systems that incorporate models developed by such third parties.

15.4. What would be the cost or other impact of such a recordkeeping system for developers of AI models or systems, creators, consumers, or other relevant parties?

STM does not have data relevant to this question. However, we believe that such costs would be appropriate in the context of respecting copyright and licensing requirements and would not exceed current expenses to train and run AI models.

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

Copyright owners – in the absence of a concluded license – should be notified directly that their works have been used to train an AI model and should be able to search for that information. STM believes that fostering, encouraging, and requiring visibility into information sources will enable users and auditors to better mitigate bias and error, to ascertain that a model was trained on information and content collected with the consent of those involved, and to ensure legal and regulatory compliance. Such use should also be licensed.

17. Outside of copyright law, are there existing U.S. laws that could require developers of AI models or systems to retain or disclose records about the materials they used for training? If your comment applies only to a particular subset of generative AI technologies, please make that clear.

Privacy laws are of particular importance and will be implicated in many of the circumstances raised in this Notice.

Copyrightability

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?

Human authors use a wide range and variety of tools to create works; using an AI system as a tool in the process of creation constitutes circumstances under which a human using a system could be considered an author, presuming a modicum of human creativity inheres in the resultant work as is described in the Copyright Office's Compendium. STM believes that accountability and transparency are central to an AI system. Only human authors can be responsible for their authentic work products and therefore accountable and liable for any problems. Any information or analysis put out by generative AI tools needs to be transparently described. Generative AI tools themselves cannot be considered capable of initiating an original piece of research without direction by human authors, as their output is based on the detection of already existing informational patterns and data correlations.

In the STM ecosystem, human authors remain fully responsible for the accuracy of their journal articles. Each author or co-author is responsible for ensuring that questions related to the accuracy or integrity of any part of a work are researched and resolved. Authorship requires the ability to approve the final version of a work and to agree to its submission to a journal. Authors are also responsible for ensuring that the work is original, that the stated authors qualify for authorship, and that the work does not infringe any third-party rights. Use of AI tools must be disclosed and described.

19. Are any revisions to the Copyright Act necessary to clarify the human authorship requirement or to provide additional standards to determine when content including AI-generated material is subject to copyright protection?

STM believes that legislation, legislative history, and the history of the relevant international legal instruments (e.g., the Berne Convention) provide conclusive and compelling evidence that only humans can be authors for purposes of copyright law.

20. Is legal protection for AI-generated material desirable as a policy matter? Is legal protection for AI-generated material necessary to encourage development of generative AI technologies and systems? Does existing copyright protection for computer code that operates a generative AI system provide sufficient incentives?

No; no; and no comment.

20.1. If you believe protection is desirable, should it be a form of copyright or a separate sui generis right? If the latter, in what respects should protection for AI-generated material differ from copyright?

Protection is not desirable.

21. Does the Copyright Clause in the U.S. Constitution permit copyright protection for AI-generated material? Would such protection "promote the progress of science and useful arts"? If so, how?

No. Given the current lack of transparency regarding materials on which many AI systems have been trained, providing copyright protection for their outputs risks institutionalizing rights infringement, perpetuating error and bias, and creating significant legal uncertainty with regard to the actual ownership of such output.

Infringement

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?

Yes, in almost all circumstances. If a copyrighted work has been reproduced in the course of AI ingestion, the right of reproduction has been implicated. If a derivative of an original work has been made without permission or compensation to the right holder, the derivative work right has been implicated.

23. Is the substantial similarity test adequate to address claims of infringement based on outputs from a generative AI system, or is some other standard appropriate or necessary?

Without adequate insight into how an AI system ingests training data – and what that training data consists of – it is difficult to apply the substantial similarity test with certainty. The substantial similarity test may be adequate, but it will need to be conducted in the context of what is known and unknown in the generative AI environment.

24. How can copyright owners prove the element of copying (such as by demonstrating access to a copyrighted work) if the developer of the AI model does not maintain or make available records of what training material it used? Are existing civil discovery rules sufficient to address this situation?

This is a key issue, and STM appreciates the Office considering it. In the absence of such records, it may be quite difficult. While existing discovery rules in civil procedure may be adequate to prove an element of copying, there is no case law yet to confirm this. This only underscores the importance of transparency for generative AI systems and their developers and operators.

25. If AI-generated material is found to infringe a copyrighted work, who should be directly or secondarily liable—the developer of a generative AI model, the developer of the system incorporating that model, end users of the system, or other parties?

The entity that used the system to generate the infringing output is likely the direct infringer in such circumstances, but the developer of the generative AI model and the developer of the system incorporating that model into another product or service may also be directly or secondarily liable.

25.1. Do "open-source" AI models raise unique considerations with respect to infringement based on their outputs?

STM does not have a view on this question at this time.

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?

As with any infringement case, Section 1202(b) should apply in the context of generative AI. STM is monitoring legal cases as they unfold and notes this particular question is implicated in the case lodged by Getty Images against Stability AI in the district court of Delaware.

27. Please describe any other issues that you believe policymakers should consider with respect to potential copyright liability based on AI-generated output.

Labeling or Identification

28. Should the law require AI-generated material to be labeled or otherwise publicly identified as being generated by AI? If so, in what context should the requirement apply and how should it work?

Yes. STM is not currently aware of any single widely accepted standard or applicable law but, for example, in the context of manuscript submissions to STM member journals, authors who have used AI or AI-assisted tools in the course of their writing process are typically asked to insert a section at the end of their manuscript entitled e.g., "Declaration of Generative AI and AI-assisted technologies in the writing process." In that statement, authors specify the tool(s) used and the reason for using it.

28.1. Who should be responsible for identifying a work as AI-generated?

The human(s) who used the AI to generate the output and potentially the AI developer, especially with respect to transparency and integrity.

28.2. Are there technical or practical barriers to labeling or identification requirements?

To STM's knowledge, there are no unsurmountable barriers.

28.3. If a notification or labeling requirement is adopted, what should be the consequences of the failure to label a particular work or the removal of a label?

STM does not have a recommendation on the consequences at this time.

29. What tools exist or are in development to identify AI-generated material, including by standard-setting bodies? How accurate are these tools? What are their limitations?

STM is aware that efforts are being made to develop such tools, but this process is nascent, and we are not aware of any that are widely used or widely accepted. There do not yet seem to be any existing tools with the accuracy or ability to monitor the vast range of material being generated or the underlying corpora of copyrighted material that fostered its generation. Without a regime that requires transparency and accountability from AI developers, it may be challenging, if not impossible, to identify AI-generated material, especially textual.

Additional Questions About Issues Related to Copyright

30. What legal rights, if any, currently apply to AI-generated material that features the name or likeness, including vocal likeness, of a particular person?

STM has no comment on this question.

31. Should Congress establish a new federal right, similar to state law rights of publicity, that would apply to AI-generated material? If so, should it preempt state laws or set a ceiling or floor for state law protections? What should be the contours of such a right?

STM has no comment on this question.

32. Are there or should there be protections against an AI system generating outputs that imitate the artistic style of a human creator (such as an AI system producing visual works “in the style of ” a specific artist)? Who should be eligible for such protection? What form should it take?

STM has no comment on this question.

33. With respect to sound recordings, how does section 114(b) of the Copyright Act relate to state law, such as state right of publicity laws? Does this issue require legislative attention in the context of generative AI?

STM has no comment on this question.

34. Please identify any issues not mentioned above that the Copyright Office should consider in conducting this study.

STM has no comment on this at this time.

Respectfully submitted,

Dr. Caroline Sutton
CEO
STM