aiim
Association for Intelligent
Information Management

October 29, 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. S.E.
Washington, D.C. 20559-6000

Re: Artificial Intelligence and Copyright

Dear Ms. Wilson and Ms. Strong,

The Association for Intelligent Information Management (AIIM) is pleased to offer comments in response to the US Copyright Office's Artificial Intelligence (AI) proposed study.

AIIM is the world's leading association dedicated to the information management industry. Information management practitioners focus on the collection, processing, storage, security, retention, and accessibility of information in an organization. Information managers are also responsible for the accuracy and transparency of information. Our member organizations include AI technology leaders as well as users of AI, and specifically Narrow AI solutions, for the past two decades.

Since 1944, AIIM has transformed the way use their information to solve business problems, create new value and strengthen organizational performance. With over 66,000 community members representing various fields, including IT, records management, and knowledge management, AIIM fosters a vibrant community of information management leaders. Through independent research, comprehensive training, and professional certification programs, AIIM empowers these professionals to enhance their skills and better serve their organizations. Our vision is to create a world in which every organization benefits from information management to reach their desired business outcomes.

AIIM offers the following responses and comments to certain questions and issues raised by the Office in the notice. We appreciate the Office's efforts to study the policy issues raised by generative AI systems in copyright law and the opportunity to provide our member organizations' insights and expertise. Our goal is to uphold safety, security, and public trust as fundamental principles in the future of AI.

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?*

The information management industry has been using artificial intelligence for over two decades. The industry has successfully employed narrow AI solutions like robotic process automation (RPA) to streamline complex processes and to efficiently process massive volumes of information. The integration of AI into an organization's information management system can greatly enhance the process of information retrieval. AI can also help with maintaining the quality of data, including correcting errors in text data and other inconsistencies.

However, with this power comes risk. It is the job of information management practitioners and leaders to provide data that is accurate, relevant, current, complete, consistent, and compliant with applicable regulations and standards.

Information must be credible to be trusted and, ultimately, used by organizations and people. As information management professionals, AIIM's members are accountable to the customers they serve, including their colleagues, customers, and the public at-large.

A critical part of accountability is a record or audit trail of what was created by the technology and evidence that the organization has acted in a legal manner. As a result, information management professionals bear responsibility for documenting the AI process.

For this reason, AIIM believes that source citations are an important way to establish information provenance and to mitigate the risks of distributing and employing inaccurate, unverified, or unlawful information. Information management professionals require documentation of AI algorithms, data sources, and decision-making processes to help their stakeholders use AI-generated information. This aligns with the entire premise of copyright itself, which is intended to incentivize creativity to serve the purpose of enriching the general public by providing access to creative work.

Copyright law at its core is an economic regulation to incentivize creators to share their work to benefit the public, as established in the Constitution. To that end, it's important that the government work to support the applications of AI in a way that protects the content creators and the content consumers. Copyright assignments must be clear—reducing liability of the organization and making end-users more comfortable to use AI-generated content.

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?*

The classification of AI-generated materials is critical in information management. Information management practitioners strive for reliable access to verifiable and accurate information. The systems they create to manage information would need to differentiate AI generated content. As a result, we believe that the government's role in considering copyright policy should focus on classification to better enable information management professionals to create systems that can rate the verifiability and accuracy of the information.

If the origins of AI-generated content are maintained by developers, the training data would be more transparent. It would then make it easier to be verified later in the iterative path—and protect end-user businesses. Such audit trails would make it easier for information managers to assess what information is more valuable, or on the flip side, what information may be fraught with inaccuracy or bias. In the future, they may be way to embed the copyright of the original content within the subset of the newly generated content.

Information management professionals also may be responsible for ensuring that the internal teams that are developing the AI tools, or the vendors from which we purchase them, provide the means for auditing where the original information is coming from. In this context, these professionals are also risk managers as well as front-line implementers of any guidelines or regulations put forth to support AI-generated content. And to fulfill their responsibilities, it is critical to have access to information about copyrights.

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.*

Generative AI requires increased scrutiny and more stringent rules over its use to ensure the higher risks associated with the technology are fully addressed. A more complex and comprehensive policy approach is needed, including limits on how entities can use such systems.

AIIM recommends requiring generative AI engines to provide source information to users, preferably as metadata. This source information allows users to validate the copyright and ownership of materials and improves the credibility of outputs.

Additionally, metadata for source information could facilitate the feasibility of legislation related to content dissemination of synthetic content. Legislation could require that source information as metadata be used by content distributors, like social media platforms, to identify synthetic content, such as deep fakes. This type of legislation would protect copyright owners and consumers.

This level of legislation regarding generative AI is required to protect consumers and organizations. Legislation would also potentially aid innovation and adoption of generative AI, by reducing risk and liability for organizations investing in generative AI.

In contrast, AIIM believes government's role regarding Narrow AI applications should be to incentivize adoption of accountable, responsible AI. A flexible, universal framework applied at the federal level would provide entities with the information and surety they need to adopt the technology and fully incorporate it into their operations.

Training

6.4. Are some or all training materials retained by developers of AI models after training is complete, and for what purpose(s)? Please describe any relevant storage and retention practices.

While AIIM will not comment on the use or acquisition of copyright-protected training materials, AIIM can provide insight into record retention and disposition.

ISO 15489 defines records as "information created, received, and maintained as evidence and information by an organization or person, in pursuance of legal obligations or in the transaction of business." Organizations capture records to ensure they can maintain that evidentiary weight.

There is a cost to retaining records. Thus, there needs to be a clear case for considering a document, file, or dataset a record. Most organizations usually retain records due to a legal or stuatory requirements; financial obligations; historical archives; or an operational need.

In the case of AI developers, the value of a Large Language Model (LLM) is in in the size of the dataset and the currency and accuracy of the data. The more outdated a model, the less valuable its output. Thus, retaining training materials does not provide fiscal, operational, or historical benefits to an AI developer. The reason for retention would very likely need to be a legal or stuatory requirement.

While AIIM avidly supports records retention in this instance, it's important to understand the impact on and costs to business when considering regulations.

There are costs for developing and maintaining a record management program. For an AI developer to retain training materials at-scale, they would need to employ the principles of records management and would likely need to adopt enterprise systems, like Electronic Records Management systems to not only retain records, but also be able to comply with regulatory requirements and possible legal requests. Employees would need training and skills to be able to successfully develop and manage record retention policies and develop automated workflows to identify, extract, and store relevant training materials in a record repository. There are also costs to storing retained data and this would increase the storage needs of developers.

A central part of records management is making sure that records are kept for as long as they are needed, but for no longer. Disposition is defined by ISO 15489 as the "range of processes associated with implementing records retention, destruction or transfer decisions which are documented in disposition authorities or other instruments."

Organizations should dispose of records in accordance with their records policy and program. Doing so will minimize the risks associated with keeping information too long. The combination of the policy, retention schedule, and documentation of disposition will allow the organization to document that it kept what it was supposed to, as long as it was supposed to, and that it disposed of the information as part of the normal course of business, all of which will help to reduce risk.

Disposing of information will also help to manage the costs associated with storing that information, including the costs of managing that information over time, which helps the organization be more efficient. There are costs to reviewing records, transferring records to archives, and/or destroying records.

It is also important to note that organizations depend on information to be compliant. The advancement of AI is critical to national security and economic growth. For organizations to adopt AI at an enterprise scale, it is important that information is deemed to be compliant with any applicable copyright laws and that the source of the information is clear. Regulations and standards reduce organizational liability by providing clear guidelines around retention and disposition to developers and customers of 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?

Metadata and metadata automation could be used to facilitate an efficient opt-out approach that occurs before data is mined for training models. AIIM recommends an approach that currently web search engines follow where the robots.txt metadata file indicates whether it's acceptable for a website's content to be indexed or not. Many AI models based on publicly available datasets could draw upon such an approach.

After data has been used for training a model, if a copyright owner wanted to opt-out their request could be handled similarly to data subject requests under the General Data Protection Regulation where organizations (data controllers) establish legitimate interest to collect and use data. Data subjects may then submit requests to access or delete their data, but those requests are constrained by the legitimate interest of the organization. The copyright owner would need to know that their work is being used first and then submit an "opt-out request". The request would then need to be validated by the AI developer to ensure the copyright claim is authentic and justified.

This approach would become too onerous to track and maintain records at the individual level but may be possible at a business level. Commercial data providers can establish automated data clearing houses to facilitate automated exchanges.

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?

Interoperability of various record management or content management systems is a significant obstacle. Metadata from one system would need to be mapped to the metadata of the receiving system for opt outs to be recognized. Industry standards can help establish this, but standards are not mandatory without regulation and legislation.

It is not feasible to obtain consent for existing public data, but going forward terms of service should be updated and provide users with an option whether to use their data for training models.

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

Licenses are legal records. As such, it's important that they follow a chain of record, which chronologically documents the custody, control, transfer, analysis, and disposition of physical or electronic evidence. The most defensible way to obtain licenses would be from the copyright owner or their agent. Licensed materials could then be denoted as licensed using metadata in an AI model developer's records management system.

In addition, content creators may have the option of developing their own datasets to train AI, building upon lawfully sourced open-source generative AI content. This would have an audit trail baked into the content. Such creators may also license the use of such a tool to customers, collaborators, or partners.

It should be noted that generative AI engines and models have largely been built with publicly accessibly information. Security protocols and permissions prevent these models from using private data. The government and nonprofit sector could play a role in educating copyright holders about protecting their information and content online.

While we seek to provide information on managing licenses as legal records, AIIM's recommendation would be to mandate the inclusion of source information as metadata as part of legislation as opposed to requiring licensure to use content. Restrictions on content stymie innovation and human creativity. Generative AI has the potential to build upon existing human knowledge at a rapid scale by connecting and combining data from a vast number of sources. Licensure could slow the potential of generative AI whereas source citations would support and grow use of generative AI by building build trust in AI outputs amongst users while providing credit to the original copyright owners.

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?*

Source citations are an important way to establish the credibility and accuracy of AI output and would help bolster adoption of generative AI by reducing consumer risk and liability. While it's possible that market demand will drive AI developers to collect, retain, and disclose training records, a requirement should be considered.

> *15.2. To whom should disclosures be made?*

Disclosures should be made to the consumer or user of the generative AI application. This has been effectively accomplished through automated source citations.

> *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?*

The most sustainable solution would be an automated solution that extracts and records data from a large language model (LLM), recording vital metadata, such as the date the information was collected; authorship; publisher; and URL address. This record would be maintained for a specified period of time in a record management system. Publicly accessible disclosures could be a part of a “Digital Commons”, which would create greater transparency and trust around how copyrighted information was used.

Members of AIIM provide intelligent information management solutions, which include technologies such as Enterprise Content Management Systems, Digital Asset Management Systems, Document Management Systems, and other Records management solutions.

Records management solutions are used to capture and manage types of information more formally. By capturing and declaring a particular piece of information as a record, the organization is committing to manage it throughout the lifecycle in a way that preserves its evidentiary weight should it be needed for a legal case, a regulatory inquiry, etc. This means records have to be stored in such a way that they cannot be edited, altered, or deleted for their entire lifecycle.

It's important that the solution be automated due the expected vast volume of data that would need to be collected. AIIM provides an outline for executing and automating a metadata plan at https://info.aiim.org/aiim-blog/how-to-build-a-metadata-plan.

There would be a cost to AI developers to adopt and implement such systems, but record management system are widely available to organizations of all sizes.

Metadata in a Digital Asset Management System or other records management system can also reflect any license or copyright restrictions. For example, a digital photo might be licensed from the owner for a one-year campaign, and the rights to use it online expire after that period. An information management or DAM system can help monitor these deadlines and secure content appropriately.

Providing such source information on the original content used to generate AI-created content also adheres to the US Copyright Office’s century old value proposition of creating and making work in all forms accessible—with the intention of enriching the general public.

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?*

A language model is a probabilistic model that predicts the probability of the next word based on the sequence of previous words the model has seen and learned. In other words, language models are “probability engines”.[1] Since mathematical equations and facts in general cannot be copyrighted, AIIM would argue it is not possible to copyright probability engines either.

In the U.S. Copyright Office’s letter regarding the *Cancellation Decision regarding Zarya of the Dawn*[2]*, t*he Office explained that where a human author lacks sufficient creative control over the AI-generated components of a work, the human is not the “author” of those components for copyright purposes."

Creative control over AI-output is best determined by establishing the ownership of the source material. There is increasing availability of large language models (LLMs) built using an organization’s data. In this case, the source material for the LLM is likely authored by the organization. Thus, the organization or individual can claim authorship of the output.

Alternatively, if a user of generative AI can demonstrate a human has significantly altered AI output and, thus, produced a new work, this could be a case where authorship could be applied. For example, an employee of an organization uses a publicly available generative AI tool, like ChatGPT, to inspire a white paper. The employee uses the AI output to create an outline for the white paper and heavily edits the document, adding new content and removing generic or inaccurate language from AI output.

Infringement

> 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?*

AIIM wrote in response to the National Telecommunications and Information Administration’s (“NTIA”) request for comment on AI accountability policy, that “both the developers of AI systems and the entity using those systems should be responsible for the outcomes they produce. Importantly, this should apply to both the organization and their employees. The purchaser of or entity using the tool must be responsible for the manner in which they use the technology, and

[1] Association for Intelligent Information Management, *The AIIM Official CIP 2023 Study Guide* (October 2023).

[2] U.S. Copyright Office, *Cancellation Decision re: Zarya of the Dawn (VAu001480196)* at 1 (Feb. 21, 2023), *https://www.copyright.gov/docs/zarya-of-the-dawn.pdf* (letter from the Office to applicant canceling the original certificate and issuing a new one covering only the expressive material created by the applicant).

when purchasing that product or service, they should recognize and be held accountable for that risk."

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?*

AIIM believes that AI-generated material must be labeled to help end-users and particularly the public identify AI-generated content. Such labeling would help establish the level of credibility and accuracy of the AI-generated content and reduce end-user risk and liability. While we believe that enough consumer demand may eventually exist to drive AI developers to develop a content audit trail by collecting, retaining, and disclosing training records, a federal requirement should be considered. However, any governmental policy to require labeling (such as a watermark or source citations in metadata) should also include funding and support to better inform the public when material was generated using AI. Finally, any policy should be developed at a national level to avoid state-by-state regulations, which would be impossible to implement.

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

In AIIM's response to the President's Council of Advisors on Science and Technology (PCAST) Working Group Request for Input on Generative AI, AIIM noted that there are three tools AI developers can employ to facilitate the recognition of accurate or inaccurate information:

- Source citations and application of metadata in AI output and other media to allow consumers to trace and verify sources.
- Watermarking of AI-generated output so consumers can use software to detect AI output.
- Confidence Ratings where applications provide a score on the potential accuracy of the output.

In each of these three cases, the remediation needs to be embedded in the solution during development and before release to the public. The government and associations like AIIM can play a role in promoting this as a best practice for the industry and educating consumers about how to identify solution providers who offer these best practices as features.

###

AIIM's response was developed in collaboration with Rachna Choudry, Kashyap Kompella, Alan Pelz-Sharpe, and Harvey Spencer. Thank you for this opportunity to share our thoughts about successful and sustainable deployment of generative AI and copyright considerations. We look forward to working with you and in collaboration with our member organizations and other stakeholders to support this growing industry while balancing innovation with risk mitigation.

Sincerely,

Tori M. L

Tori Miller Liu, MBA, FASAE, CAE
President & CEO
Association for Intelligent Information Management (AIIM)
8403 Colesville Rd, #1100, Silver Spring, MD 20910 USA
https://www.aiim.org/