Before the United States Copyright Office

Artificial Intelligence and Copyright **Notice of Inquiry & Request for Comments**	**Docket No. 2023—6**

Copyright Clearance Center, Inc. (CCC) appreciates the opportunity to submit the following comments in response to the United States Copyright Office's (Copyright Office's) August 30, 2023 Notice of Inquiry and Request for Comments (NOI) on artificial intelligence (AI) and copyright.

About CCC

With more than forty years of expertise in copyright and information management, CCC designs and delivers innovative information solutions that power decision-making by helping people integrate and navigate data sources and content assets. We collaborate with our customers to accelerate discovery and progress by shortening the distance between data, information, and insight. Our offerings are always market-based.

CCC is the primary collective management organization (CMO) for text in the United States. We offer global corporate licensing on a fully voluntary, non-exclusive basis, plus academic licensing services primarily within the United States. Our mission is to advance copyright, accelerate knowledge, and power innovation. CCC manages copyright licensing and content delivery on behalf of — among others — (1) publishers who manage copyrights themselves and (2) non-publishers (such as pharmaceutical companies) who manage both their own copyrights and third-party materials. In addition, CCC services include copyright education, library staffing, library and publisher software development, API development, persistent identifiers, and data/metadata services.

CCC supports AI in many ways. These include: our licensing and software solutions; collaborations with rightsholders, users and trade associations; ongoing development of persistent identifiers (PIDs); and support of FAIR data principles. We understand how important it is to help the AI journeys of users by enabling development of ethical, reliable, and trustworthy AI systems to further business, pedagogical, educational, and research goals. We understand how important it is to support rightsholders by ensuring adequate remuneration for their investments in the content that drives AI systems.

CCC is a founding member of the International Federation of Reproduction Rights Organizations (IFRRO).

Responses to Questions

CCC is pleased to provide the following responses to the NOI's 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?

Developments in AI systems have incredible potential to support our society and economy in ways both familiar and yet unknown. To fulfil this potential, AI development must be paired with an appreciation of and respect for creators and copyright. Copyright is an engine of innovation, a key part of economic activity, and incentivizes the creation of foundational materials upon which AI is often built. Support for copyright is crucial to our culture, science, jobs and the advancement of AI itself.

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?

AI generated materials may both advance text publishing and hinder it. In sectors such as science, news, and book publishing, poor quality AI materials can generate bad science, promote misinformation, and lead to harmful results. This is not to say that such is the inevitable result of all AI; merely that it is a meaningful risk with respect to certain AI applications. AI can advance text publishing by providing tools for writing, checking, validating, and improving text-based works. It is also useful for primary research that may result in the creation of new content.

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.

Given the proliferation of relevant materials, we limit ourselves to recommending the following:

Daniel Gervais' brief memorandum on copyright and AI, available here; and Nancy Wolff and Elizabeth Safran's *Diving Deeper into Artificial Intelligence: Understanding the Risks in Incorporating AI Technology into the Workplace*, available here.

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 international landscape for copyright protections, exceptions, and limitations as they apply to AI is in flux and inconsistent. As a foundational matter, all countries with copyright systems have laws that apply to AI because copyright's exclusive rights,

exceptions, and limitations all apply to AI. For example, AI development and training requires the making of copies and, as such, any country with copyright laws has laws which govern AI use of copyrighted materials. Whether the copying is an actionable infringement or falls under a general or specific copyright exception varies by country.

As explained in more detail below, however, some countries have provided exceptions or limitations for specific subsets of AI (namely text and data mining (TDM)), but it is unclear how these provisions apply to rapidly changing technologies such as those now seen in generative AI models. These narrower TDM provisions, moreover, are far from uniform; for example, the commercial or noncommercial scope of existing provisions differ, as does whether the works TDM uses must be lawfully acquired. While governed by international treaties copyright law is national. As can be seen in the Singapore copyright exception, discussed below, this creates an opportunity for countries with few content industries to create broad exceptions designed to lure tech investment at the cost of harm to creators in other countries, such as the United States. To avoid a race to the bottom where the laxest jurisdiction enjoys an unfair advantage perhaps attracting problematic AI development, the United States can protect domestic consumers, creators, and businesses, including technology companies, by requiring compliance with domestic rules, regardless of the source or domicile of the AI developer or the AI entities developed.

Without policies supporting these goals, cross-border interactions become increasingly unclear and problematic.

These inconsistencies pose obvious challenges, but the key is to ensure that the United States has policies that support our flexible and balanced longstanding copyright system. Therefore, it is crucial that the United States:

- Adopt policies that encourage lawful acquisition of copyrighted materials and appropriate licensing,
- Require transparency in what copyrighted materials AI technologies use, and
- Avoid overbroad policies that do not recognize the wide range of use cases and licensing models and that undermine the longstanding strength of U.S. fair use consideration of all four Section 107 factors.

These considerations are based on a review of existing international provisions described below.

UK. Section 29A of the UK's Copyright, Designs and Patents Act 1988 contains a copyright exception that allows lawfully acquired content to be mined "for the sole purpose of research for a non-commercial purpose." The UK considered a broader commercial exception last year but that change in law is increasingly unlikely. A recent Parliamentary Report "warn[ed] that the Government's original plan to exempt text and data mining by AI from copyright protection risks reducing arts and cultural production to mere 'inputs' in AI development and shows a clear lack of understanding of the needs of the UK's creative industries." As such, commercial use of materials for AI remains infringing.

China. There is no copyright exception for TDM/AI in China. By contrast, the latest iteration of Chinese policy provides liability rules for use of copyrighted content; namely

that service providers are responsible for the legality of the source of data used for training generative AI, including ensuring that the data sets (1) do not infringe intellectual property rights, (2) comply with cybersecurity and personal information protection laws, and (3) are accurate, truthful, objective, and diverse.[1] The rules also require that service providers use data and underlying models with legitimate, labelled sources. The service provider must be able to provide details of training data used on request and must label AI generated content.

EU. The EU has adopted specific copyright rules for TDM. Article 3 of the Digital Single Market Copyright Directive (DSM) provides a mandatory exception in the field of scientific research benefiting "research organizations," as defined. In addition, there is an exception in Article 4 for commercial TDM, which does not apply if a rightsholder expressly reserves TDM rights. These provisions, however, were developed before the more widespread use of and familiarity with publicly available generative AI technologies, and thus the contours of Articles 3 and 4's relation to rapidly changing technologies are a topic of ongoing discussion.

In addition to the DSM Directive, the EU is, as of this writing, still in the process of reviewing its AI Act. While the AI Act is not primarily intended to regulate copyright and other intellectual property concerns, its broad reach currently involves a potential transparency requirement that would mandate generative AI platforms to provide "sufficiently detailed summaries" of copyrighted materials used.

Switzerland, Japan, and Singapore. These three countries all have copyright exceptions that extend to commercial TDM. In Japan, a broad copyright exception in legislation is tempered by an equally broad caveat that the rights and interests of copyright holders may not be unreasonably prejudiced by any TDM or data analysis activity. Switzerland has also adopted a TDM exception for any scientific research,[2] whether commercial or non-commercial. Switzerland and its courts consider international copyright treaties as self-executing and so the three-step test included in the Berne Convention, the TRIPS Agreement, the WIPO Copyright Treaty (WCT), WIPO Performances and Phonograms Treaty (WPPT), and the Marrakesh Treaty are all directly applicable. Thus, while Japanese and Swiss law both provide for commercial exceptions, they are limited in scope, both because (1) rightsholders now frequently license works for AI uses, and (2) generative AI systems have been shown to compete with underlying works. In terms of Japanese law, these uses "unreasonably prejudice the copyright holders." In terms of Swiss law, many commercial TDM uses could violate the three-step test.

By contrast, Singapore's copyright law includes a broad TDM exception, allowing TDM (as defined) for any purpose, including commercial purposes, without the ability to modify the arrangement by contract. The only arguable restriction is that the pre-existing copyright

[1] The "Interim Measures" are available in Chinese here: http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm. A useful Summary is provided by Bird & Bird here: https://www.twobirds.com/en/insights/2023/china/what-you-need-to-know-about-china%E2%80%99s-new-generative-ai-measures; *see also* https://www.theguardian.com/commentisfree/2023/apr/22/can-china-keep-generative-ai-under-its-control-well-it-contained-the-internet.

[2] Article 24d of the Swiss Copyright Act.

works will have to be sourced legally, but the law allows infringing materials to be used if the infringement was not "knowing."

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 is not necessary, but the Office welcomes any proposals or text for review.

With respect to copyright infringement, there are numerous cases pending in the United States, and the need for specific legislation and what it might provide will be driven in part by the end results of those cases. With respect to AI-generated materials, for reasons set forth herein, labelling should be required, but voluntary labelling is unlikely to be adequate. Also, some form of disclosure requirement with respect to materials used for AI training is advisable, particularly when materials are used without consent or license.

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

We are not aware of any major type of copyrighted material that is NOT used to train AI models, lawfully and unlawfully. In the text sector, materials are collected and curated lawfully by publishers, other rightsholders, intermediaries, and aggregators (such as CCC). Materials are, unfortunately, also unlawfully aggregated by pirate sites, so-called "shadow libraries" (i.e., pirate sites with better branding), and, if the allegations set forth in lawsuits are accurate, by developers of some well-known open-source generative AI models.

6.1. How or where do developers of AI models acquire the materials or datasets that their models are trained on? To what extent is training material first collected by third-party entities (such as academic researchers or private companies)?

In the text sector, developers of AI models — when acting lawfully — acquire materials and data sets from publishers, other rightsholders, websites that allow crawling, intermediaries, and aggregators (such as CCC). Significant amounts of content are available through licenses, including open licenses such as CC BY and CC BY-NC. Significant amounts of content are also available through the public domain. When acting unlawfully, AI developers receive materials from pirate sites, through downloading in violation of express terms and flags, and from so-called "shadow libraries," among other things.

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?

Copyrighted materials are licensed for AI use directly by rightsholders and collectively through rights aggregators such as CCC. CCC's collective licenses are non-exclusive, global, and fully voluntary. Our current AI-related offerings are focused on the corporate, research, academic and education markets.

Additionally, in science publishing, under "open access" business models, copyright owners employ open licensing which sometimes allows licensed reuse for AI under the terms of

such licenses. According to this report, open models accounted for 31% of articles, reviews and conference papers in 2021.

6.3. To what extent is noncopyrighted material (such as public domain works) used for AI training? Alternatively, to what extent is training material created or commissioned by developers of AI models?

Noncopyrighted works, including data, US government works, and public domain materials, are used in training AI.

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.

Humans communicate in natural language by placing words in sequences; the rules about what the sequencing and specific form of a word are dictated by the specific language (e.g., English). An essential part of the architecture for all software systems (and therefore AI systems) that process text is how to represent that text so that the functions of the system can be performed most efficiently.

Almost all large language models are based on the "transformer architecture," which invokes the "attention mechanism." The latter is a mechanism that allows the AI technology to view entire sentences, and even paragraphs, at once rather than as a mere sequence of characters. This allows the software to capture the various contexts within which a word can occur.

Therefore, a key step in the processing of a textual input in language models is the splitting of the user input into special "words" that the AI system can understand. Those special words are called "tokens." The component that is responsible for that is called a "tokenizer." There are many types of tokenizers. For example, OpenAI and Azure OpenAI use a subword tokenization method called "Byte-Pair Encoding (BPE)" for its Generative Pretrained Transformer (GPT)-based models. BPE is a method that merges the most frequently occurring pairs of characters or bytes into a single token, until a certain number of tokens or a vocabulary size is reached. The larger the vocabulary size, the more diverse and expressive the texts that the model can generate.

Once the AI system has mapped the input text into tokens, it encodes the tokens into numbers and converts the sequences (even up to multiple paragraphs) that it processed as vectors of numbers that we call "word embeddings." These are vector-space representations of the tokens that preserve their original natural language representation that was given as text. It is important to understand the role of word embeddings when it comes to copyright because the embeddings are the representations (or encodings) of entire sentences, paragraphs, and even documents, in a high-dimensional vector space. It is through the embeddings that the AI system captures and stores the meaning and the relationships of the words from the natural language.

Embeddings are used in practically every task that a generative AI system performs (e.g., text generation, text summarization, text classification, text translation, image generation, code generation, and so on).

Word embeddings are usually stored in vector databases but a detailed description of all the approaches to storage is beyond the scope of this response since there is a wide variety of vendors, processes, and practices that are in use.

7. To the extent that it informs your views, please briefly describe your personal knowledge of the process by which AI models are trained. The Office is particularly interested in:

7.1. How are training materials used and/or reproduced when training an AI model? Please include your understanding of the nature and duration of any reproduction of works that occur during the training process, as well as your views on the extent to which these activities implicate the exclusive rights of copyright owners.

Our response to question 6.4 above captures the nature of the reproduction of copyrighted works. As to the duration, in all known systems the word embeddings remain stored in the database permanently since their elimination would remove that part of the captured input and lead to inconsistencies in the calculations (internal to the model) that were performed during training. In fact, with careful fine-tuning and prompting, one should be able to reproduce verbatim all text used during the training of a large language model.

7.2. How are inferences gained from the training process stored or represented within an AI model?

The word embeddings discussed in detail in our answer to question 6.4 above are one way in which a direct representation of the input is represented and stored in an AI system. However, there is also an indirect mechanism of storage – namely the training of the parameters that make up the model itself. Although the exact description of what happens is beyond the scope of this response, all AI systems are fundamentally mathematical models with a very large number of free parameters. During the training phase, the values of these parameters are specified. *Ceteris paribus*, if the very same system processes two entirely different training sets, then the values of the model in these two cases will be very different. So, the specifics of the training set are implicitly captured (and stored in perpetuity) within the values of the AI model.

7.3. Is it possible for an AI model to "unlearn" inferences it gained from training on a particular piece of training material? If so, is it economically feasible? In addition to retraining a model, are there other ways to "unlearn" inferences from training?

For purposes of this response, we will limit the scope to an AI model that can be classified as a generative pretrained transformer. There is no general answer that is valid for every conceivable AI system since their number and their nature cover a very large spectrum.

For GPT models, the system does not "unlearn," but there are ways to address infringement, from licensing to mitigation. For example, identifying the word embeddings

that correspond to a particular training input should be feasible with relatively good accuracy and efficiency and those can be used in filters as a post processing task to constrain the output of the model. The most economically feasible approach is licensing of copyrighted material used in training and the appropriate remuneration of the rightsholders, if needed.

7.4. Absent access to the underlying dataset, is it possible to identify whether an AI model was trained on a particular piece of training material?

Yes, it should be possible, but the level of effort depends on the system and the kind of access to the system that is allowed. Some level of system interrogation is necessary to perform that task. This is one reason why we favor disclosure requirements in the absence of a license. Especially for smaller creators, placing the burden of identifying use on the rightsholders is inequitable, burdensome, and unnecessary given that the AI developers have the information as to what they have used.

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.

U.S. law has no specific rules governing the use of copyrighted materials to train AI. Rather, such uses fall under the general copyright regime. Under U.S. law, copying copyrighted content to train AI can state a cause of action for infringement.[3] Thus, such activities require a license to be non-infringing unless they fall under the fair use exception.

The application of fair use to an infringement is fact dependent. Copying for purposes of training an AI will usually entail copying the complete work. Whether the copying is for commercial or non-commercial research purposes will be considered. The courts will also look very closely at market harm under the fourth factor. As stated by the Supreme Court in *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 590 (1994) "[the fourth factor] requires courts to consider not only the extent of market harm caused by the particular actions of the alleged infringer, but also 'whether unrestricted and widespread conduct of the sort engaged in by the defendant … would result in a substantially adverse impact on the potential market' for the original." And, as reinforced by the recent Supreme Court decision in *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. ___ (2023), the impact of the infringing use on licensing is one of the key factors in determining market harm.

Relevant instructional cases include the cases mentioned above as well as *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169 (2d Cir. 2018), where the Second Circuit Court of Appeals rejected a fair use defense in a case of allegedly transformative compiling of recorded broadcasts into text searchable databases that allowed search and viewing of short excerpts. By contrast, the Second Circuit had previously considered the text mining of scanned books for non-commercial social science research in *Authors Guild v. Google, Inc.*

[3] *Thomson Reuters Enters. Ctr. GmbH v. ROSS Intelligence Inc.*, 529 F.Supp.3d 303 (D. Del. 2021) (downloading and copying of Westlaw database for the purpose of training AI).

721 F.3d 132 (2d Cir. 2015), and held that copies made and used for a specific purpose involving snippets would likely fall under fair use.

There are currently multiple pending cases in the U.S. relating to use of copyrighted content for the development of AI systems. Congress has expressed interest in the issue by including language in the SAFE Innovation Framework that the Framework will "support our creators by addressing copyright concerns, protect intellectual property, and address liability."

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?

Fair use analysis will certainly treat non-commercial uses more leniently than commercial uses under factor 1. As explored in this blog, under *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, the application of the "transformative" doctrine under the first factor is likely to be narrower than previously believed by many. As fair use is fact dependent, different stages of training may have different analyses.

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?

Unless licensed or subject to a valid fair use defense, entities that collect and distribute copyrighted materials may be direct infringers as a result of making, tokenizing, storing, and distributing copies, as well as contributory infringers for providing infringing content to others for their further infringement.

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?

As noted above, the fair use analysis involves evaluation of four non-exclusive factors set forth in Section 107 and is entirely fact-specific. Noncommercial use is more likely to be fair use but is not automatically so. Commercial use is less likely to be fair use but is not automatically unexcused infringement. As is discussed in *Basic Books, Inc. v Kinko's Graphics Corp.*, 752 F. Supp. 1522 (S.D.N.Y. 1991), and *Princeton Univ. Press v. Michigan Document Servs., Inc.*, 99 F.3d 1381 (6th Cir. 1996), the same activity can be fair use when done by one person and infringing when done by another (particularly commercially). The specific lesson from these cases is that the fair use analysis evaluates the USE and not the USER. Ultimately, when noncommercial entities make copyright-sensitive uses of copyright-protected materials for the benefit of commercial entities and/or with the goal, intention and effect of commercial activity (*see Cambridge University Press v. Albert* , 906 F.3d (11th Cir. 2018) (use by a university intended primarily to reduce the university's costs posed "a severe threat of market harm," is presumptively tantamount to a commercial

use and is in any event not a fair use)), those uses are less likely to be deemed a fair use. Additionally, in sectors such as academic and scientific publishing whose primary markets are non-commercial research, fair use analysis will look closely at market harm even with non-commercial use.

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?

The quantity of materials used varies based on the developer, its goals, its technologies, its market, the availability of licensed content, and even its views on copyright law. Fair use's third factor, the amount and substantiality of use, however, focuses on the amount and substantiality of material copied from any one copyrighted work. It does not consider the percentage that the copied work comprises of all infringement by the defendant. As such, the overall volume of materials used should, at best, be irrelevant in a fair use analysis. With that stated, most infringing AI systems use entire works, so factor three would favor the rightsholder.

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 the alternatives set forth in the question are potentially relevant to factor four in a fair use analysis. Any single use of a copyrighted work may affect any of the identified markets and the relevance to any particular analysis may vary from use to use. The effect of the use on the licensing market for the work is also highly relevant to fair use analysis. In brief, the fair use analysis may not be "short-circuited" by way of any *a priori* conclusions that only one of these options is applicable.

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 is, and should remain, an opt in regime. Placing the burden of asserting rights on the copyright holders is inequitable, burdensome, and largely impractical. Only those making copies know what they are copying in the first instance and thus the copyright owners are not in a position to opt out.

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

All uses should require consent, subject to any relevant defenses.

9.2. If an "opt out" approach was 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?

There is good reason that copyright is an "opt in" regime. Some AI developers have gathered content by routinely ignoring flags, copyright notices and metadata. Thus, while there are protocols and flags that can be used and are used by rightsholders and honored by ethical AI developers, they are no substitute for placing the responsibility for compliance on the user. Moreover, requiring flags and metadata assumes that the content resides on a server or website under the control of the rightsholder. This is not always true. For example, in the recent case of *Am. Soc'y for Testing & Materials v. Public.Resource.Org, Inc.*, 82 F.4th 1262 (D.C. Cir. 2023), the Court of Appeals for the District of Columbia Circuit ruled that the non-commercial posting of technical standards incorporated into reference by law is fair use. It would be problematic to assume that the entity posting the standards over the objection of copyright owners would take steps to reserve the copyright owner's AI rights.

Finally, for smaller creators, any obligation to adopt technical protection measures or flags is unfair and unduly burdensome.

Technical flags and metadata are useful for AI developers who act ethically and have another great value; where ignored by AI developers they can provide evidence of willfulness.

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?

It is feasible to acquire advance consent of copyright owners. It is not feasible to place the burden on rightsholders to police their rights without knowing who is using their works without authorization and how the works are being used.

The burden of implementing technical measures, flags, and metadata may, depending on the sector, be involved, complicated and costly to copyright owners. In the recent past, international sector-wide initiatives such as ACAP have absorbed significant time and resources on the part of rightsholders and users seeking to act ethically, only to be rejected by the tech industry. Current efforts of note include the W3C Text and Data Mining Rights Reservation Protocol.

As noted above, as a practical matter, a copyright holder may have no control over websites where its content is held. This is especially true where content is posted in violation of copyright or under a copyright exception.

There is certainly enough copyrightable material available under license to build reliable, workable, and trustworthy AI. Just because a developer wants to use "everything" does not mean it needs to do so, is entitled to do so, or has the right to do so. Nor should governments and courts twist or modify the law to accommodate them.

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?

Remedies are available under copyright and contract law where jurisdiction is possible in the United States and assets exist to enforce judgments. Injunctive relief against infringing technologies should also be available. Failure to honor objections can be evidence of willfulness entitling to rightsholders to increased damages.

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?

Permission of the owner of the relevant right under copyright should be adequate.

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

Copyrighted materials can be licensed for AI use directly from rightsholders and collectively through collective management organizations such as CCC.

In science publishing and in other fields where open licenses are used, copyright owners frequently license reuse for AI under the terms of such licenses.

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

Direct voluntary licensing is feasible and commonly utilized in text and images. Among many others, the prominent copyright holders Associated Press, Getty Images and vAIsual all offer licenses. We offer no opinion as to other sectors.

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?

CCC already offers market-based, global non-exclusive voluntary licenses to support AI in the commercial research, schools, and education technology sectors. These licenses were built with rightsholders and users based on agreed understandings of needs and market conditions. We are well suited to provide and have proven experience of providing these and other AI licenses for the text sectors at a minimum.

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?

As an intermediary serving both rightsholders and users, we believe this is a question best answered by those communities. We can support whatever license type is decided.

With that said, compulsory licenses are typically provided in the United States only where there has been a market failure. For example, in certain music, cable, and satellite use cases, Congress has stepped in to provide specific statutory licenses, but only after extensive study, discussion, and review. At this point in the development of AI technologies, the market is providing options for licensing works, including those offered by CCC and described above.

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

As an intermediary serving both the rightsholders and users, we believe this is a question best answered by those communities. CCC can support whatever license type is decided. That said, we believe it worthwhile to explain how an extended collective license (ECL) works.

ECL presupposes the existence of a voluntary license between relevant parties representing a substantial majority of both licensors and licensees for a specifically-described, licensable use that, for explicit reasons, is unable to encompass all necessary works of potential-but-absent licensors. At that time, the parties participating in the voluntary license system petition the government to extend the existing licensing system to the potential-but-absent licensors on condition that the potential-but-absent licensors and their works are treated on an equal basis with the participating licensors and their works. Upon the government's grant of such a petition, an ECL is created along with some mechanism to administer it.

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

By necessity, licensing regimes vary based on the type of work at issue — specifically, it is extremely unlikely that there is any AI technology (any more than that there is any human being) that will make the same use of (or need the same license for), for example, a written text, a musical score, a musical performance, a photograph, painting or sculpture, a building blueprint, a motion picture, and a computer program.

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)?

Licensing is licensing, whether producing a book, making a film, distributing music online, or using for AI. Everyone making, storing, tokenizing and/or distributing copies needs a license unless covered by an exception. License scope and price are typically negotiable. The price paid by the curator is likely to be lower if it does not secure further distribution

rights for the developer or the company deploying the AI technology. These are all part of normal value proposition discussions.

At CCC, some licenses we issue are personal to the licensee, while others allow the licensee to work with contractors, and yet others allow the rights to flow upstream and downstream more fully.

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.

It is possible to identify the degree to which a particular work contributes to a particular output from a generative AI system. The exact method and the degree of difficulty for doing so will depend largely on (1) what the exact AI architecture is and how the AI system was trained, (2) what evaluation tools and system access are available, and (3) the definition of "contribution" and more specifically the "degree of contribution." For example, one approach for a system like ChatGPT could be based on the novelty of tokens introduced in the word embeddings by a specific work — that would require access to the vector database with the word embeddings of the AI system. Another approach could be based on the number of weights adjusted during training when the work is processed — that would require access to data during training (and even a modification of training to allow for this test). Hence, the technical feasibility, the economic viability, and level of objectivity in performing such a task would span a wide range of options. The creator of an AI system can decide whether a particular work is of value to the objectives that they want to achieve and should ensure that any copyrighted work used is properly licensed and the rightsholders are appropriately remunerated.

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

There is a licensing requirement today. Some comply and others do not. Those who act in compliance with the law are required to compete with infringers. Those who license works for the development and adoption of AI systems benefit economically from those licenses and lose economically when developers infringe.

When copyrighted works are used for the training of an AI system, the output (or results) from the AI system may compete directly with the works contained in the training set (e.g., by substituting for that work), or it/they may compete by replacing the functionality without replacing the work (e.g., by providing the information contained in the work but not excerpts themselves). Likewise, the use of content for training without rightsholder consent preempts and usurps the current and future licensing markets for use of copyrighted content in training and similar applications. The consequences of not allowing copyright rightsholders to freely set and/or reject licensing terms for use of their materials by AI systems presents a huge economic and social risk and undermines copyright law, the incentives it creates, and innovation overall.

Licensing enables the lawful use of copyrighted materials and is widely used to accommodate new technologies. Individual licenses and collective licensing solutions are available in a wide variety of markets, including markets for text, image, and audio works.

Collective licensing provides a harmonized collection of rights from many rightsholders – and is particularly beneficial when there is a need to use large numbers of materials from numerous rightsholders, as is often the case when using works for AI. The economic consequences of requiring licenses will be to bolster creators, the U.S. economy, and our culture.

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

Intentionally left blank.

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.

15.1. What level of specificity should be required?

At a minimum, it should be specific enough for a rightsholder to know that their materials were used when used without consent. Where licensed, licensing terms could set and vary any disclosure requirement.

15.2. To whom should disclosures be made?

Disclosures should be to the rightsholder to the extent material is used without explicit permission of the rightsholder. Where a license exists, disclosure and reporting should be set in the license terms.

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

As with any user of intellectual property, developers using systems that incorporate third party models should ensure that adequate licenses/chain of titles attach to the models or else they risk liability.

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?

We are unaware of any industries where maintaining source records and chain of title information is an unreasonable burden. In general, it is a best practice.

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

To the extent materials were not licensed, the obligation should be absolute, regardless of whether the use is under a copyright exception.

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?

NYC's Automated Employment Decision Tool law requires employers who use AI in hiring to disclose this fact to candidates and to retain records for annual audits to ensure no bias results from AI use.

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?

Intentionally left blank.

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?

Intentionally left blank.

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?

Intentionally left blank.

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?

Intentionally left blank.

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?

Intentionally left blank.

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

AI-generated outputs implicate the exclusive rights in preexisting copyrighted works, such as the rights of reproduction or the rights to create derivative works when the preexisting work are used in training and, depending on the system and output, subsequently.

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?

Existing copyright law should be applied.

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?

If the developer of the AI model does not maintain or make available records, copyright owners will often struggle to prove copying. Civil discovery rules are not useful for acquiring records that do not exist. This is why transparency and disclosure are critical.

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?

Existing copyright law should be applied.

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

Many "open-source" models seem to have been built on copyrighted content without rightsholder consent.

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?

Nothing about generative AI makes it exempt from the rules of 17 U.S.C. 1202(b). The requirements of 17 U.S.C. 1202(b) can be modified in a license.

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

Intentionally left blank.

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?

Wholly AI-generated material should be labeled or otherwise publicly identified as being generated by AI.

Research suggests that as more content is machine generated, it will be critical for artificial intelligence (AI) developers to distinguish it from human-created content to ensure the continued development of quality AI. This is because AI will need to avoid training on AI-developed content. Thus, the requirement of disclosure when content is AI generated and proper standards for such disclosure will both protect the public through transparent labeling and warnings and will aid in future AI development.

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

Any person or entity publicly distributing a work should be responsible for identifying it as AI-generated.

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

There are no major technical or practical barriers that prevent all forms of labeling or identification, which can be accomplished in a wide range of manners.

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?

The consequences of the failure to label a particular work or the removal of a label should include fines and injunctive relief, and potentially additional penalties when the failure/removal is intended to defraud or carries the risk of harm.

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?

Intentionally left blank.

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?

Intentionally left blank.

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?

Intentionally left blank.

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?

Intentionally left blank.

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 requires legislative attention in the context of generative AI?

Intentionally left blank.

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

Intentionally left blank.

Respectfully submitted on October 30, 2023 by:

Catherine Zaller Rowland

Catherine Zaller Rowland
VP, General Counsel

Roy S. Kaufman

Roy S. Kaufman
Managing Director, Business Development