Before the
United States Copyright Office
Washington, DC

In re

Artificial Intelligence and Copyright

Docket No. 2023-6, COLC-2023-0006

COMMENTS OF
THE COMPUTER & COMMUNICATIONS INDUSTRY ASSOCIATION (CCIA)

In response to the notice of inquiry and request for comments published by the U.S. Copyright Office ("the Office") in the Federal Register at 88 Fed. Reg. 59942 (Aug. 30, 2023), and extended at 88 Fed. Reg. 65205 (Sept. 21, 2023), the Computer & Communications Industry Association ("CCIA")[1] submits the following comments. CCIA appreciates the opportunity to provide input on these important issues and also participated in the Office's listening sessions on copyright and AI this spring.

CCIA's members are leaders in AI innovation. Not only are they developing and deploying a range of new AI-powered products for personal and enterprise users, they have developed popular open-source machine learning frameworks that others are now using in both academia and industry. CCIA members therefore have a significant interest in ensuring that the development and use of AI technology is promoted, rather than suppressed, by the U.S. copyright system.

CCIA believes that existing U.S. copyright law is capable of addressing issues related to artificial intelligence and serves to promote creative activity in AI technology. While unique

[1] CCIA is an international, not-for-profit trade association representing a broad cross section of communications and technology firms. For more than 50 years, CCIA has promoted open markets, open systems, and open networks. CCIA members employ more than 1.6 million workers, invest more than $100 billion in research and development, and contribute trillions of dollars in productivity to the global economy. A list of CCIA members is available at https://www.ccianet.org/members.

issues might arise in the future that may require additional legislation or regulation, the technology-neutral nature of the Copyright Act is sufficient to address present issues regarding AI and copyright.

Before responding to individual questions, CCIA wishes to highlight its recent white paper, “Understanding AI: A Guide to Sensible Governance.”[2] The paper is intended to serve as a guide for policymakers to craft rules that maximize the benefits of AI while reducing the potential risks. With smart regulation and governance, the United States can continue to lead the world in AI innovation. AI is not a single technology, but rather a family of related, but distinct, technologies, each of which may be applied in significantly different contexts. Responsible AI deployment can be best achieved through flexible, considered regulation that avoids unintended consequences.

I. 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?

Generative AI tools from technology developers are already benefiting the public and other stakeholders by democratizing accessibility, including enabling translation, speech recognition, computational photography, and AI toolkits others can use to create new works. AI can also enable creators’ new creative processes. For example, the author of Zarya of the Dawn used AI imagery to illustrate her text, and various small role-playing game publishers are using AI to generate imagery to illustrate their books. Additionally, researchers of all types — students, academics, and public and private sector employees — can increase efficiency and

[2] CCIA, *Understanding AI: A Guide to Sensible Governance* (June 2023), https://ccianet.org/library/understanding-ai-guide-to-sensible-governance/.

improve the accuracy and timeliness of their work. Small businesses are using AI tools to improve their efficiency, operations, and competitiveness. These tools support small businesses through sales analysis, marketing and design, customer support, and back office management.

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?

Many of CCIA's members are already innovating and investing in the AI space. Further, many generative AI systems—including third-party generative AI systems—are trained and/or run on compute services provided by CCIA's members. To the extent a revision or reinterpretation of copyright law disincentivizes the creation or use of new AI technologies, it would harm CCIA members who provide compute resources utilized by AI model developers and AI application developers as there would be reduced demand for their services.

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.

There has been a lot of recent scholarship from technical and legal experts on artificial intelligence. As discussed in the introduction, CCIA recently released a paper on AI.[3] There are also a number of new technical papers which describe current research into training techniques, potential techniques for limiting the output of copyrighted works, and fair use,[4] as well as many new and forthcoming pieces oriented towards the legal status and interaction of copyright and AI from law professors.[5]

[3] CCIA, *Understanding AI: A Guide to Sensible Governance* (June 2023), https://ccianet.org/wp-content/uploads/2023/06/CCIA_Understanding-AI.pdf.

[4] *See, e.g.*, Peter Henderson, et al., *Foundation Models and Fair Use*, arXiv (Mar. 2023), https://arxiv.org/pdf/2303.15715.pdf; Nicholas Carlini, et al., *Extracting Training Data from Diffusion Models*, arXiv (Jan. 2023), https://arxiv.org/pdf/2301.13188.pdf; Nikhil Vyas, et al., *On Provable Copyright Protection for Generative Models*, arXiv (July 2023), https://arxiv.org/pdf/2302.10870.pdf.

[5] *See, e.g.*, Pamela Samuelson, *Generative AI meets copyright*, Science (July 2023), https://www.science.org/doi/10.1126/science.adi0656; Amanda Levendowski, *How Copyright Law Can Fix*

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 flexible and balanced copyright law regime in the U.S. has been key to American success in innovation in emerging technologies like AI. The U.S. leads the way in AI development in large part due to the fair use right.

Other countries may be approaching these issues based on their unique legal frameworks and domestic industry. Japan and Singapore have enacted specific AI exceptions that do not require compensation, while the Israeli Ministry of Justice issued an opinion that its fair use provision, modeled on the U.S. fair use doctrine, permits the training of AI systems without compensation.[6] The EU's recent Directive on Copyright in the Digital Single Market established two exceptions for text and data mining (TDM). TDM for scientific research is permitted without compensation, while TDM for all other uses is permitted subject to an express opt-out by the copyright owner.

While an AI-specific exception for training without compensation could be useful in providing certainty in the United States, the flexible fair use approach provides a valuable floor that permits training for AI systems and must be maintained.

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.

Artificial Intelligence's Implicit Bias Problem, 93 Wash. L. Rev. 579 (2018), https://digitalcommons.law.uw.edu/cgi/viewcontent.cgi?article=5042&context=wlrl; Daryl Lim, *AI, Equity, and the IP Gap*, 75 SMU L. Rev. 815 (2022) ("The result is an IP system that perpetuates inequity when elite groups own an increasingly large share of IP rights"), https://scholar.smu.edu/cgi/viewcontent.cgi?article=4939; Matt Sag, *Copyright Safety for Generative AI*, 61 Houston L. Rev. (forthcoming 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4438593; Mark Lemley & Bryan Casey, *Fair Learning*, 99 Tex. L. Rev. 743 (2021), https://texaslawreview.org/fair-learning/.

[6] Jonathan Band, *Israel Ministry of Justice Issues Opinion Supporting the Use of Copyrighted Works for Machine Learning*, Disruptive Competition Project (Jan. 19, 2023), https://www.project-disco.org/intellectual-property/011823-israel-ministry-of-justice-issues-opinion-supporting-the-use-of-copyrighted-works-for-machine-learning/.

The existing U.S. legal framework is sufficient to address intellectual property issues related to AI. While no legislative or regulatory amendments are needed at this time, an AI-specific exception for training without compensation could provide additional certainty to AI system developers.

However, any report issued by the Office as a result of this inquiry should clearly state that fair use permits the ingestion of copyrighted material in the course of an AI process.

II. Training

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.

Various types of AI models will rely on different training processes. Training materials may not be used at all for some forms of AI, while the generative AI foundation models that have received the most attention will use large training datasets. Because of this technological variation, a single answer for all AI will be necessarily incomplete.

However, while copies of training materials may be made initially, as explained *infra* in response to Question 8, numerous appellate courts have correctly found this to be fair use.

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

Inferences are not directly gained or represented within an AI model. Instead, an AI model consists of a collection of linkages between billions of nodes in a directed graph. These linkages, and in particular the strength of the linkages between nodes, are what determine the behavior of the model. However, no specific node or linkage maps to a specific inference, and any given piece of training data will likely have a small impact on a large number of nodes and linkages.

This lack of any capability to link the final model to specific inputs or concepts is well-understood in the AI community, with AI researchers stating that "as of early 2023, there is no technique that would allow us to lay out in any satisfactory way what kinds of knowledge, reasoning, or goals a model is using when it produces some output."[7]

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?

To CCIA's knowledge, this is not presently feasible without fully retraining a model; the specific impact on inferences derived from a particular piece of training material is not retained, to the extent it even exists in the first place. Given the large expense of retraining a model, including significant energy consumption, there is no economically feasible way to 'unlearn' inferences from a particular piece of training data.

However, this area of technology continues to rapidly develop, and unlearning might become economically feasible in the future. At the same time, there is no guarantee that this circumstance will come to pass. Accordingly, CCIA recommends that the Office conduct its analysis and make its recommendations based on the assumption that unlearning is not feasible, but leave open the possibility of revisiting the question if unlearning approaches become available. Of course, the technical feasibility of unlearning is a different issue from the desirability of doing so from a policy perspective.

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?

[7] Samuel Bowman, *Eight Things to Know about Large Language Models* (2023), https://cims.nyu.edu/~sbowman/eightthings.pdf.

Absent access to the underlying dataset, so-called “extraction” attacks can, under particular circumstances, provide non-conclusive evidence of whether an AI model was trained on a particular piece of training material. However, because these types of attacks can also be used to extract private information, there is ongoing work aimed at preventing them. Further, differences in training or model specifics can cause these attacks to fail or succeed at higher rates, meaning that they do not provide a general mechanism for determining if an AI model was trained on a particular piece of data.

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.

The existing statutory framework and related case law concerning the fair use right, 17 U.S.C. § 107, clearly permit the ingestion of large amounts of copyrightable material for the purpose of an AI algorithm or process learning its function. Numerous appellate courts have correctly found the mass copying of raw material to build databases, including commercial databases, for automated computational analysis to be fair use under 17 U.S.C. § 107. *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015); *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014); *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 640 (4th Cir. 2009); *Perfect 10 v. Amazon.com, Inc.*, 508 F.3d 1146, 1165 (9th Cir. 2007); *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818 (9th Cir. 2003). Training AI is a form of this computational analysis. Judge Leval’s opinion in *Google* provides the clearest analysis of why the creation of datasets for computational analysis, and their subsequent uses in AI training, are fair uses.

To help prevent this issue from being relitigated in every case involving an AI training database, the Office’s report should draw a bright line stating that uses of copyrighted materials as data in the creation and deployment of AI machine learning systems are fair uses. Such clear guidance not only would conserve judicial resources, it would prevent erroneous decisions. This bright line would benefit innovators, courts, and the public.

AI algorithms and other processes often require the ingestion of large amounts of data. Assembling that data may entail converting it into a more usable format, e.g., translating image files into mathematical image representations. In addition, backup copies of the materials may be necessary to protect against loss of data in the event of system failure. Temporary reproductions of portions of the material in a computer's random access memory are a normal part of any computer program, including the process of training an AI algorithm. These copies are not viewable or consumable by the outside world. These non-expressive copies are not consumable by the public and do not function as market substitutes for copies of the ingested works.[8]

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?

There are two relevant uses to consider: (1) ingestion and training, and (2) output.

With respect to ingestion and training, and as noted above, a strong weight of existing case law is in favor of finding that ingestion and training uses are highly transformative fair uses. Expressive works are ingested for the purpose of understanding what expression is and how it relates to other expression, not for the purpose of commercializing that expression. This sort of highly transformative use is most analogous to the sort of text and data mining at issue in the *Google Books* litigation. However, while the *Google Books* litigation dealt with taking text and data and making it searchable, in the case of generative AI, the AI takes text and data and makes

[8] Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 Nw. U. L. Rev. 1607 (2009) https://lawecommons.luc.edu/cgi/viewcontent.cgi?article=1068&context=facpubs.

it into a tool to create new works entirely. The creation of new works is at the core of copyright and thus the kind of transformation at issue in generative AI is a higher level of transformation than that present in Google Books.

With respect to output, the relevant use is determined by the user of a generative model. Much like a tape recorder can be used to infringe copyright or to record a new work, the output of an AI model can be used in a variety of ways. The model developer and operator have created a system with substantial non-infringing uses. The user of the generative model is the entity that directs and controls what the model will output; as such, they bear responsibility for any infringement.[9] Existing copyright law addresses the fair use question with respect to output; where an output would be a fair use, it would be so regardless of whether a human or AI created the work.

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?

The underlying purpose of the use is the same, so the entities should not be treated differently. These entities provide additional broad benefits, such as reducing barriers to entry for smaller firms by reducing the cost of acquiring a dataset and providing a standardized point of comparison for comparative testing of models after development. Further, treating collection of materials differently might raise issues in other contexts, such as archives like the Internet Archive which also collect and distribute materials.

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?

[9] To be sure, the fair use analysis may be different if a model is fine-tuned in a manner that is more likely to produce infringing outputs.

Because the use is so highly transformative, and has no impact on the market for any of the works being ingested, little weight should be given to the commercial nature of the firm engaged in the AI activity. Neither commercial nor noncommercial uses of AI models or datasets derived from copyrighted materials should be considered copyright infringements.

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?

Training datasets started out large and have continued to grow over time. The original paper introducing the transformer structure utilized an English-German translation dataset of about 4.5 million sentence-pairs, totaling approximately 2 GB in size. Recent datasets are significantly larger. Modern large language models (LLMs) like LLAMA-2 and PaLM-2 use trillions of input tokens as their dataset, representing multiple terabytes of data. And generative AIs designed to generate images can rely on even larger datasets. For example, LAION-5B, a text-image dataset, contains nearly 6 billion image-text pairs, with even a reduced-resolution version of the dataset totaling approximately 50TB in size and higher-resolution versions reaching hundreds of terabytes.

The large volume of training materials underscores that the contribution of each work, and the impact on the market for each work, is *de minimis*. It also illustrates the problems an opt-in regime would create.

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?

The fourth fair use factor analysis should continue to focus on the effect of the market for that particular copyrighted work allegedly infringed alone.

However, the output of the AI system is not relevant to the question of whether ingestion is a fair use. Generative AI systems will typically have substantial non-infringing uses. If substantial non-infringing uses exist, then the model itself should continue to receive fair use protection, much like a VCR manufacturer has fair use protection for time-shifting uses by customers even though the VCR could also be used to copy a copyrighted work.

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)?

Because ingestion is a fair use, no affirmative consent is required by law. However, copyright owners who wish to may have effective means of opting out of allowing their works as training materials. For example, some may be able to put their content behind technological protection measures such as paywalls, which are legally protected by 17 U.S.C. § 1201. Copyright holders making content available on the web are able to use the widely-used robots.txt exclusion protocol to prevent the work posted to their websites from being crawled by specific AI bots.

Technical experts and standard setting organizations such as the World Wide Web Consortium (W3C) or the Internet Engineering Task Force (IETF) could work to develop an exclusion protocol with more granularity that would permit search engine bots but exclude other bots, or would permit a bot to ingest data from a site for some uses but not others. Some major AI developers are already beginning work on such a standard. Several companies have recently announced extensions that will allow website publishers to allow search bots but exclude AI training bots,[10] though a more universal approach that does not rely on identifying and excluding specific bots would be helpful.

[10] Danielle Romain, *An update on web publisher controls* (Sept. 28, 2023), https://blog.google/technology/ai/an-update-on-web-publisher-controls/; Microsoft Bing Blog, *Announcing new options for webmasters to control usage*

Further, given the enormous scale required for LLM creation, an opt-in regime is effectively going to block AI development. It will also likely have negative impacts on equity. While obtaining permission from, *e.g.*, songwriters may be viable through existing collective licensing groups, training data created in less common languages or from various subcultures is far less likely to be organized and the appropriate entity to contact for permission may even be impossible to determine.

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

As noted above, major AI companies are developing mechanisms to allow copyright holders effective choice in whether to allow their content for training, and standards bodies are likely to follow suit. Nonetheless, ingestion is a fair use regardless of the commercial nature of the enterprise, so the consent of the copyright owner should never be legally required. This is not an “unfair” result; authors cannot prevent human authors from using their work for inspiration or training purposes so long as the final product does not infringe upon the original work.

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?

As noted above, an enhanced robots.txt would be an ideal way to achieve this for Web data.

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?

of their content in Bing Chat (Sept. 22, 2023), https://blogs.bing.com/webmaster/september-2023/Announcing-new-options-for-webmasters-to-control-usage-of-their-content-in-Bing-Chat.

It is infeasible to get consent in advance given the huge volume of works involved. However, it is feasible—and significant efforts are already underway—to establish technical mechanisms for opting out.

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?

This would likely be a consideration separate from copyright law and the scope of this inquiry. However, while CCIA's members would not intentionally violate such an objection, an objection like robots.txt should not impact the fair use analysis.

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?

No one should have the "right" to object to an AI model being trained on their work. As noted above, technical means to allow a copyright owner to indicate an objection to training are being developed. An employee should not have the ability to force the employer to deploy these technical means.

Works for hire are owned by the employer, rather than the employee or commissioned creator. The human creator who created a work for hire was already compensated for their work and has no copyright interest. If a right to object to the use of a work for hire existed, it would belong to the employer. However, given the volume of copyrighted works owned by large employers, allowing employers to take this type of action would exclude large swaths of data that would aid in technological progress and the quality of AI systems and create significant barriers to entry for small entities wishing to develop new AI technologies. Theoretically, this could lead to a scenario in which AI models are trained solely with public domain works, leading to an incredibly limited scope of potential output and innovation and potentially to perpetuation of biases.[11]

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

Copyright owners’ consent should not be required to train generative AI models, nor would there be an efficient way to obtain it. At best, large AI developers with vast resources can engage in voluntary licensing over large datasets from prolific copyright owners. This would likely unfairly allocate leverage to large corporations, while sweeping over smaller creators with fewer works to license. Licensing creates barriers that would result in unrepresentative and less diverse datasets.

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

Direct voluntary licensing for AI systems would be infeasible in most — if not all — creative sectors, at least in combination with an opt-in system of licensing. Especially in the digital age, when large volumes of work are produced and published online each day, it is dubious that any licensing process will be able to keep up with non-AI innovation, calling into question the technology’s utility. Furthermore, it is unlikely that developers will expend the resources to enter into licensing agreements with less prominent creators, resulting in an undiversified dataset composed primarily of work from the largest (and likely, the most litigious) copyright holders.

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?

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

[11] *See* Levendowski, *supra* note 5.

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?

Congress should not consider establishing a compulsory licensing regime. There is no principled basis for establishing such a regime. Just as a reader does not need to pay for learning from a book, an AI system should not have to pay for learning from content posted on a website.

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

No and no.

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)?

Much of the material on which generative AIs are trained may lack any identified or identifiable author from whom to obtain a license. Even where an author might be identified, contacting them might be difficult or impossible. And because individual licenses would need to be obtained from each and every author in the billions of works being used as training data, the scale of transaction cost required to develop an AI model would be economically infeasible for even the largest entities.

In contrast, an approach that combines opt-outs via Web exclusion with metadata indicating opt-outs limits the transaction costs significantly while still allowing those who wish to either restrict use of their work for training or wish to receive compensation for it to do so.

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.

As a general matter, because of the enormous number of works ingested, it is not possible to identify the degree to which a particular work contributes to a particular output from a generative AI system. However, if a particular work appears in many different online locations and thus is ingested repeatedly, the AI may appear to have "memorized" the work when in fact the computational analysis in the model has been inadvertently distorted. AI developers already employ techniques like deduplication to avoid this problem and are working on additional mitigation measures to prevent this undesired phenomenon.

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

Licensing requirements would be economically inefficient and difficult to enforce. The advancement of AI systems is consistent with the goals of intellectual property protection under the Constitution — to promote progress, creativity, and innovation. AI system developers are incentivized to advance their technologies by the widespread adoption and interest in these technologies. If they are limited to a certain set of licensed materials, they will have fewer capabilities and compel fewer users. Furthermore, with the sheer volume of content produced each day, it would be nearly impossible for AI systems to remain current, which is an important advantage to using open-source AI tools.

This could also result in anti-competitive behavior from entities with more resources to license more materials than their competitors. Even if such licensing is non-exclusive, it will create a network effect, compelling more users to gravitate towards the AI system with access to the most training materials, and consequently the most capabilities. This would both discourage new entrants and potentially create a new monopoly on creative output, which could be harmful to innovation and progress as a whole. Mandating licensing agreements for generative AI would lead to inferior technologies, fewer competitors in the marketplace and hindered innovation generally.

III. 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?

Neither private AI developers nor creators of training datasets should be required to collect, retain and disclose records regarding the materials to train their models outside of the litigation context. First, there is no analogous obligation for human individuals and organizations. Movie directors do not have to disclose mood boards used to inspire their sets, for instance. Second, this would be both a resource-intensive and logistically intangible goal which would hinder the progress of science, counter to copyright law's goals. Third, even if executed properly, the utility of such record disclosures is dubious. It would be difficult to trace every usage or application of a work in an AI system whose output won't even necessarily reflect that usage.

There may be other reasons for developers and dataset creators of AI systems used in a government or non-profit capacity to collect, retain and disclose records regarding the materials to train their models, but not for copyright purposes.

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

There should be no obligation to notify copyright owners that their works have been used to train an AI model. This would be akin to an aspiring artist notifying all artists whose work they've studied or used to train that they have done so. This would serve no purpose but only deter creativity, especially when there is no clear way to tell how the work the AI system ingested was used and by whom. There is no way to ensure that the works were even used in a capacity that could enable infringement. Notifying copyright owners preemptively would create more problems than it would solve.

IV. Generative AI Outputs

a. 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?

Under current Copyright Office guidelines, humans who use AI to create a work "may claim copyright protection for their own contributions to that work," excluding any AI-generated content that is more than *de minimis*. This would extend certain protections to an end user, given that the human exercised sufficient creative control over the work's expression, and "actually formed" the traditional elements of authorship.

Any report authored as a result of this request should state that a work produced by an AI algorithm or process, absent sufficient contribution of a natural person to the resulting work, should not qualify as a work of authorship protectable under U.S. copyright law. Sufficient contribution could occur either via a human author significantly changing the AI's output into a final work, or by a human author exerting sufficient control over the output of a generative AI; however, so-called "prompt engineering" should per se be insufficient for a human to obtain copyright in the output.

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?

No, copyright law has been clear on this for more than a century. As properly interpreted by the Copyright Office, a work produced by an AI algorithm or process, without the involvement of a natural person contributing to the resulting work does not qualify as a work of authorship protectable under U.S. copyright law. This interpretation follows in a long line of cases and guidance finding that only a natural person can create a work of authorship protectable by copyright.

The Office currently refuses to register a work that was not created by a human being. The most recent version of the Compendium of U.S. Copyright Office Practices cites several cases from the 1880s in explaining that "copyright law only protects 'the fruits of intellectual labor' that 'are founded in the creative powers of the mind,'" and "because copyright law is limited to 'original intellectual conceptions of the author,' the Office will refuse to register a claim if it determines that a human being did not create the work." Compendium of U.S. Copyright Office Practices, Third Edition, at § 306 (citations omitted). The Office adds that it "will not register works produced by a machine or mere mechanical process that operates randomly or automatically without any creative input or intervention from a human author." *Id.* at § 313.2.

There is no need for this provision to change. Artists who incorporate technology into their artistic process can still obtain a copyright on their works, so long as the human artist has contributed a sufficient amount of original material to the combined work. Work created by AI systems should be held to the same standards as any other work.

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. All that should be protected is the human contribution. Computers don't need incentives; only people do. And existing incentives—both legal, such as copyrights and patents, and non-legal, such as first-mover advantages and a desire to supply a commercial need—will suffice to ensure the development of generative AI technologies.

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?

At the outset, it is at least uncertain if the meaning of "author" at the time of the Founding would have included a machine. Congress may entirely lack the power to grant copyright protection to the output of an AI.

Even if Congress has such a power, withholding copyright protection from a work resulting from an AI process for which there was no expressive contribution by a natural person is justifiable on policy grounds. The AI algorithm, and the computer that runs it, does not require the economic incentive provided by copyright in order to create works. Indeed, AI is capable of quickly producing an enormous array of works. Recognizing copyright in such output could quickly create a minefield of legal issues, leading to litigation and uncertainty.

To be sure, the human creator of the software that runs the AI algorithm or process would receive a copyright in the expressive aspects of the AI software (and perhaps a patent for inventions in the AI software). No additional copyright incentive is necessary to encourage the creation of AI software.

Because there is no policy justification for awarding copyright to the output of an AI process, it would not "promote the progress of science" to do so.

b. 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, when they produce material that is substantially similar in protected expression. This could be the result of a user requesting output that is substantially similar to protected expression and/or of the AI failing to operate as intended, *e.g.*, the "memorization" problem discussed above. AI developers are employing mitigation measures to prevent this outcome and continue to develop additional measures.

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?

The substantial similarity test is adequate to address claims of infringement based on outputs from a generative AI system, as these cases should not be treated differently from cases involving infringement by a human. Copyright law has adapted to new technologies throughout history, and AI is no different.

If the output of an AI system resembles existing copyrighted material, then the ordinary analysis of whether copyright infringement has occurred would apply. In short, the question would be whether the AI system had access to the allegedly infringed work, and whether the AI system's output is substantially similar in protected expression to the allegedly infringed work. The first question can be answered by examining whether the work in question was part of the training data used by the AI system. If it was not used in training, the AI system did not have access to it. The second question is answered as it would be in any other copyright case.

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?

Generally, any liability should lie on the end-user who requests and publishes a copyright-infringing work. Much like many other areas of technology, including photography, AI systems are strong examples of a "staple article or commodity of commerce suitable for substantial noninfringing use." *Sony Corp. of America v. Universal City Studios, Inc.*, 464 U.S. 417, 440-42 (1984). Misuse of AI systems to infringe copyright, much like misuse of a VCR or computer to impermissibly replicate copyrighted content, is attributable to the user, not the manufacturer of the system being abused.

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

Open-source AI models comprise many of the more popular generative AI models, including LLMs, translation tools, and chatbots, and any regulation or legislation concerning AI from this point forward should take open-source AI into account. Open-source AI should be treated the same as other forms of AI and as humans when it comes to alleged infringement—and in particular, with respect to outputs, open-source model developers should not be liable but rather users who generate and publish an allegedly infringing work.

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?

17 U.S.C. § 1202(b) is irrelevant to the outputs of the system. Rights under § 1202(b) are limited to only removals and alterations that will “induce, enable, facilitate, or conceal an infringement of any right under this title.” Even if copyright management information were to be removed in the training process, the AI provider does not do it in a way that will knowingly aid in infringement of any right under Title 17. The training works retain their copyright management information, and the output of the generative AI system is not a work or copy of a work with copyright management information. Accordingly, § 1202(b) generally cannot apply to the output of a generative AI.

c. 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?

Practices are quickly evolving in the industry, and the government should encourage that development, before rushing to legislate. Industry practices may provide helpful data on what labeling—if any—is helpful without being unworkable for consumers or companies.

One historical analog that should be considered is from the realm of open-source licenses, where the original Berkeley Software Distribution (BSD) license contained a requirement which obligated those incorporating BSD-licensed software to include a disclosure in any advertising of their product that their software included a contribution from the copyright holder. When incorporating only a single BSD-licensed piece of software, this was manageable, but as more and more projects began to use the BSD license and projects began to incorporate different BSD-licensed software, the disclosure requirement quickly became unmanageable, with some software requiring upwards of 75 disclosures in every piece of advertising material. Ultimately, this clause was removed from future BSD licenses. Similarly, if an AI disclosure requirement were generally in place, such a disclosure would likely quickly become unmanageable when multiple AI tools are used in combination to create a given piece of material. This would likely lead to an overwhelmingly large disclosure which most users would simply skip through or ignore, much like GDPR cookie management click-throughs.

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?

A number of tools are in existence or being developed, especially in the context of detecting academic malfeasance. However, they are inaccurate in practice and often lead to far greater problems. For example, one article examined AI detection techniques applied to human writing from native English speakers and native Chinese speakers writing in English. It found that the AI detectors had an average false positive rate of 61.3%, with even the best classifier still having a false positive rate of 19.8%.[12] Another preprint paper tested 14 different AI content detectors, finding that even the best of them had significant errors, and that no tool had an

[12] Weixin Liang, et al., *GPT detectors are biased against non-native English writers*, ScienceDirect (July 2023), https://www.sciencedirect.com/science/article/pii/S2666389923001307.

acceptable tradeoff between false positives and false negatives — even the best tool tested still produced an erroneous classification approximately 25% of the time.[13]

d. 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?

State rights of publicity are the most important legal rights for dealing with name or likeness. However, they differ significantly between states and may not apply well to all AI-generated materials. For example, less than half of all states protect vocal likeness.[14]

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?

There is no evidence of a need to create a new federal right of publicity, whether AI-specific or general.

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?

Such protections may raise concerns at the intersection of copyright and the First Amendment. It also does not fit well with existing right of publicity approaches, which generally protect against commercial exploitation of an artist's public persona, not against imitation of their style in other artistic works.

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?

[13] Debora Weber-Wulff, et al., *Testing of Detection Tools for AI-Generated Text*, arXiv (June 2023), https://browse.arxiv.org/pdf/2306.15666.pdf.

[14] *Cf.* INTA, *Right of Publicity State of the Law Survey* (2019), https://www.inta.org/wp-content/uploads/public-files/advocacy/committee-reports/INTA_2019_rop_survey.pdf.

In certain instances, especially when a "soundalike" recording is used in advertising, the courts have determined that vocal imitations can violate rights of publicity.[15] However, in general, "soundalikes" do not constitute copyright infringement. This perspective is supported by the legislative history of 17 U.S.C. § 114(b), which clarifies that "the mere imitation of a recorded performance would not constitute a copyright infringement, even if one performer intentionally aims to replicate another's performance as closely as possible."[16] Section 114(b) is critical to protect a performer who might want to re-record a song where the copyright in the sound recording belongs to the record producer. Any new federal legislation addressing the copying of a person's name, image, likeness, or style must be carefully drafted to protect the person against exploitation by an entity to which those rights may have been transferred.

* * *

CCIA appreciates the opportunity to comment on these important issues and would be happy to provide any additional assistance that might be useful to the Office as it prepares its report.

Respectfully submitted,

Ali Sternburg, Vice President, Information Policy
Josh Landau, Senior Counsel, Innovation Policy
Erin Sakalis, Law Clerk
Computer & Communications Industry Association (CCIA)
asternburg@ccianet.org

October 30, 2023

[15] *See, e.g.*, *Midler v. Ford Motor Co.*, 849 F.2d 460 (9th Cir. 1988).
[16] H.R. Rep. No. 94-1476, at 106 (1976).