1818 N Street, NW
Suite 410
Washington, DC 20036

Oct. 30, 2023

United States Copyright Office
Artificial Intelligence and Copyright Request for Information

Comments of Public Knowledge

Nicholas P. Garcia
Policy Counsel
nick@publicknowledge.org

Meredith Filak Rose
Senior Policy Counsel
mrose@publicknowledge.org

I. General Questions (Q.1, 2, 5)

1. [G]enerative 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?

Public Knowledge works to promote free expression, an open internet, and access to creative works and tools. We fight for a more creative and connected future for all, and there is tremendous potential for GAI to move us towards this vision.

At the same time, we must acknowledge the challenges posed by the rapid and large-scale automation of creative and intellectual work. But AI's effect on the creative labor market is different from, and broader than, its intersection with copyright law. Creative work is economically precarious, with higher-than-average unemployment rates and lower-than-average incomes.[1] The employers of creative workers, on the other hand, tend to be large, concentrated, and monopsonistic.[2] Copyright law alone cannot address these dynamics or the profound economic implications raised by AI systems.

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?

As described above, Public Knowledge recognizes a range of potential benefits and risks posed by GAI technology. Additionally, as a public interest advocacy organization that works on a range of technology policy issues, we are also keenly aware of how GAI may specifically impact advocacy work and public processes.

First, the rapid adoption and use of GAI creates a threat of societal backlash, which could result in changes to technology policy and copyright law that imperil the openness and accessibility of the internet, throw up barriers to creativity, and rentrench the power of the biggest publishers, distributors, and content creators while leaving small and independent creators to struggle with economic precarity and disruption.

---

[1] Randy Cohen, Artists in the U.S. Workforce 2006-2020, Americans for the Arts (Mar 2021), https://www.americansforthearts.org/by-program/reports-and-data/legislation-policy/naappd/artists-in-the-us-workforce-2006-2020.

[2] *See* United States v Bertelsmann SE & Co., et al, Civil Action No. 21-2886-FYP (D.D.C. Nov 14, 2022).

Second, while GAI could lower barriers to participation in democratic and public processes—such as open comment periods like this—there is also the prospect that bad actors could employ GAI tools to create false narratives and engagement. Even without the benefit of GAI this has happened before. In 2017, the FCC was flooded with fraudulent automated public comments in support of repealing net neutrality rules during the "Restoring Internet Freedom" proceeding.[3] While those efforts were identified due to the rudimentary methods available to perpetrate such fraud at scale, GAI may make it more difficult to recognize and evaluate authentic participation in public processes in the future.

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.

While our copyright system is far from perfect, and GAI is forcing society to grapple with difficult questions about the nature of creativity and the economic inequalities in our creative industries, our existing copyright laws and doctrines are generally suited to address the challenges posed by GAI.

The perceived risks to creativity posed by GAI are, by and large, not substantively novel. Artists have learned from (and mimicked) one another's styles since time immemorial and celebrity impersonators have existed as long as we have had celebrities. By and large, the law is already equipped to deal with these harms. While GAI may increase the speed and reach of these practices, it has not changed the underlying substance in a way that requires new legislation.

Nevertheless, in light of the potentially disruptive effect of GAI and the significant antipathy towards the large tech firms and the Silicon Valley ethos that currently dominates the AI sector, there may be a temptation to deviate from the established copyright canon and wield copyright as a bludgeon against the technology as a whole. As the Copyright Office already determined when it recommended against creating special ancillary copyrights for news content,[4] even when there are serious

---

[3] Jon Brodkin, "Up to 9.5 million net neutrality comments were made with stolen identities" ArsTechnica (Oct. 17, 2018), https://arstechnica.com/tech-policy/2018/10/up-to-9-5-million-net-neutrality-comments-were-made-with-stolen-identities/; Ryan Singel, FILTERING OUT THE BOTS: WHAT AMERICANS ACTUALLY TOLD THE FCC ABOUT NET NEUTRALITY REPEAL (Oct. 15, 2018), https://cyberlaw.stanford.edu/blog/2018/10/filtering-out-bots-what-americans-actually-told-fcc-about-net-neutrality-repeal.

[4] US Copyright Office, Copyright Protections for Press Publishers (June 2022, available at https://www.copyright.gov/policy/publishersprotections/202206-Publishers-Protections-Study.pdf

concerns about the future of important human activities, expansions of copyright aren't necessarily the effective solution—and would necessarily mean limiting fair use and free expression. New copyright legislation, or significant changes in existing copyright doctrine in the courts, to attempt to redress larger issues of economic dysfunction or realign intellectual property rights to be less permissive of learning, sharing, and fair use would be a mistake.

This should not be misconstrued as calling for a hands-off approach to AI oversight more generally. Public Knowledge has filed public comments emphasizing the need for: expert and adaptable regulation and oversight of AI, ideally through an expert AI regulator; government investment in AI capability, including development of public sector AI; a "both/and" approach to dealing with speculative AI safety risks while still addressing existing and imminent AI harms; and robust protections and accountability processes to mitigate bias, eliminate discrimination, and protect privacy.

Additionally, it is important to shore up existing laws and regulations where necessary. The application of AI does not suddenly bypass existing safeguards and protections, however in some areas updates or clarifications may be invaluable to ensuring proper accountability.

One such area of concern relevant to the USCO are "right of publicity" or "name, image, and likeness" (NIL) laws. While these rights of action are not exactly copyrights or intellectual property rights, their overlap with issues of creativity, expression, and the control of creative works will impact copyright. Currently, right of publicity and NIL laws exist as a patchwork of state-level regulations. A federal right of publicity law could be a strong tool to combat both economic (unfairly capitalizing on a public figure's likeness) and disinformation harms posed by GAI technologies like deep fakes and voice models. There is also growing consensus among creative stakeholders around the need for such a federal law; as such, we encourage the USCO to closely examine how any proposed legislation might impact our copyright system. We discuss the contours of such legislation in response to Questions 30 and 31 below.

Finally, Congress needs to address the real underlying issues: economic inequality driven by substantial corporate consolidation in creative industries and the systemically reinforced economic precarity of creative work. Additionally, GAI's impacts are likely to spread beyond creative industries to the workforce more generally, and thus policymakers should be prepared to address these broad economic concerns using tools more suited to addressing them such as competition policy, labor law, and investment in a robust social safety net. Copyright cannot, and has never been able to, secure a living wage for the vast majority of creators. There is

no reason to believe that this core problem can be fixed (or even meaningfully ameliorated) by simply adding "more copyright." Congress must examine more holistic and substantial responses rather than making copyright law carry the water for the vast, systemic economic dysfunction that leaves creators economically vulnerable.

II. Understanding AI Training (Q.6-7)

Policy decisions about new and emerging technologies should be firmly rooted in informed and expert information about the technology. However, it is important to recognize that AI research and development remains a rapidly evolving and highly technical field. There are diverse approaches—and considerable secrecy—among companies at the forefront of AI technology. To become best informed, the Copyright Office should review what we hope will be the detailed and up-to-date information about AI training from the broad spectrum of companies, AI researchers, academics, and civil society organizations submitted as part of this NOI.

We include brief answers to the NOI questions on AI training to contextualize our understanding for the purpose of our analysis of the applicable copyright law and our policy recommendations.

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

Copyright-protected materials used in AI training datasets include a wide array of data including text from the open web, books, software code, photographs, digital art, music, videos, or any other media, depending on the nature and purpose of the model.

These materials are collected and curated through various channels and methods, including through web scraping, web indexing, online repositories, digital libraries, proprietary data collection, or through partnerships with content providers. Many of the most important ML training datasets are publicly accessible datasets assembled and curated by academics, universities, and nonprofit research groups.

7. To the extent that it informs your views, please briefly describe your personal knowledge of the process by which AI models are trained.

For the purpose of our copyright analysis, it is helpful to break down the AI training process into distinct components: dataset creation, model training, and model deployment. Each of these components might be, and often is, carried out by

different parties and for different purposes. Each of these components represents a point at which copyright-related questions might arise.

Dataset creation can be carried out by a developer of an AI system, but is also likely to rely in whole or in significant part, on datasets created and curated for the purpose of research by noncommercial entities. Dataset creation may involve simply indexing open web content, as in the case of the LAION-5B dataset,[5] but in other instances may involve collecting and organizing a vast quantity of potentially copyrighted content.

Model training involves the system processing a given piece of data, and using mathematical analysis or evaluation of the training data to learn from it. The actual techniques through which a system interacts with data can vary widely, and will be different for different kinds of media or depending on the purpose of the model. The work being analyzed is not reproduced or stored in the model; instead, the model uses each work to improve its overall understanding of whatever it is training on (art, photographs, writing, and so on). Most importantly, a GAI model does not contain an archive or directory of the works it trains on that the system queries, searches, or references. Based on the best explanations of how GAI training works, training a GAI system is generally analogous to reading a book, looking at a photograph, admiring a painting, or listening to music. While this may sound simple, large-scale ML training is computationally (and therefore financially) expensive.

Finally, trained models may be deployed by different parties than those that collected the dataset(s) or developed the model through training. Further potentially complicating analysis, is the fact that an existing model can be further refined, or fine-tuned, using additional rounds of training on more curated datasets. Most importantly, model deployers connect a model with software code that allows the model to accept inputs, process that input computationally using the patterns it learned in the training process, and deliver an output. This process is called inferencing.

> *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.*

There are different approaches to AI training and how training materials are handled. Most importantly, our understanding is that the non-transitory reproduction and the retention of training materials are not inherently necessary for the training process,

[5] LAION-5B FAQ, available at https://laion.ai/faq/.

though retention of the content of training data may be expedient for a number of reasons.

As previously noted, some AI training datasets are indexes pointing to URLs on the open web. These indexes allow the ML system to access the material needed for training without doing more than the kind of transitory copying inherent in the nature of any kind of digital access.

Even when training datasets are composed of content that is reproduced locally for training purposes, once a given model has trained on the material there is no need to retain the training data. ML models and GAI systems do not (indeed cannot) store or access their training data during inference; they are not indexes or databases or algorithms layered on top of an index or database.

While training data might be retained by AI developers for a variety of reasons after completing training of a model, these data handling practices do not, and should not, factor into how the resulting models are evaluated from a copyright perspective.

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

To vastly oversimplify, the inferences gained from the training process take the form of building and modifying the strength of connections between a model of datapoints, resulting in a collection of mathematical weights that define the model. The work being analyzed is not reproduced or stored in the model; instead, the model uses each example to improve its overall understanding of whatever it is training on (art, photographs, writing, and so on).

> *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?*

Generally, no. While additional training may be able to target certain behaviors of the existing model, removing the impact of a specific piece of training data from an existing model is not possible.[6] Retraining a model absent the particular piece of training material would remove the effect of that information from the model but retraining can be extremely costly, time consuming, or both.

---

[6] *See*, Tiffany Li, Algorithmic Destruction, SMU LAW REVIEW (forthcoming 2022), available at https://ssrn.com/abstract=4066845.

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?

While there are various techniques being developed to attempt to detect whether a particular piece of training data was used in a given AI model, we are unaware of any accepted, reliable means of doing so.

III. Evaluating AI Training (Q. 8)

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.

Applying our existing copyright laws to our best understanding of GAI systems leads to the conclusion that while this technology is new, and there may be flaws in design and execution that create instances of copyright infringement, its core elements and essential ideas do not run afoul of copyright law.[7]

Creating AI Training Datasets Generally Constitute Fair Use

Considering datasets separately from AI training is important because entities that create and assemble datasets most directly handle and evaluate the potential copyrighted material that lies at the heart of the AI training controversy. Datasets can be created by separate entities from AI developers that do training, are shared across AI projects, and may seem like the obvious target for copyright infringement allegations based on simple, clear cut instances of reproduction of copyrighted material.

However, the creation, curation, and function of AI datasets makes them strong contenders for fair use protection. They are transformative works, with minimal contribution of each constituent work to the overall value of the complete work, and the nature of their use is preliminary to non-infringing creative activity.

The most instructive cases are the Google Books cases and *Sega v. Accolade*.

The parallels to the Google Books cases are particularly strong. Google Books specifically came under fire from authors, publishers, artists, and photographers over the mass digitization of copyrighted content. The project was ultimately found to be a fair use that promotes copyright law's very purpose: to expand public knowledge

[7] *See generally*, Mark A. Lemley and Bryan Casey, *Fair Learning*, Texas Law Rev., Vol. 99 Issue 4 (March 2021), available at https://texaslawreview.org/fair-learning/.

and understanding.[8] The court found Google's use to be highly transformative, as Google Books provided valuable information about the book without serving as a market substitute for it. Google Books also expanded access to books, serving the public interest. Although Google is a for-profit company, the court noted that Google did not sell the scans or the snippets, nor did it show ads in the snippet view, minimizing direct commercial gain. Similarly, while Google scanned entire books, the court emphasized that users were only shown a limited number of snippets, preventing them from reading or accessing the book's entire content. Thus, even though the entire book was scanned, the actual amount displayed to users was minimal, making this use reasonable in the context of its transformative purpose. Finally, in factor four, the court determined that Google Books did not serve as a market substitute for the original works.

This same analysis applies to AI training datasets: they are massive compilations of digitized copyrighted content that has have been processed to allow for a different use than the original works (i.e. ML training) and the datasets are not serving as a means of access to the original works or as market competitors to the original works (no one is downloading AI training sets to get access to copyrighted content, especially since the datasets themselves are generally based on publicly accessible information).

The *Sega v. Accolade* case further bolsters this argument.[9] Accolade copied and reverse-engineered Sega's software to understand the functional requirements for Sega Genesis compatibility to develop their own games that were compatible with the Sega Genesis console. The court found that this was a transformative use, as the objective was to gain understanding and create new, original content, not to replicate Sega's software. The original copying was merely a necessary and non-infringing precursor to allow for better understanding of the copyrighted content, which leads to an ultimate non-infringing fair use.

AI Training Itself Does Not Violate Copyright

The AI training process itself does not violate any copyright rights, and need not even be considered through fair use analysis.

---

[8] *Public Knowledge Welcomes Sweeping Victory for Fair Use in Google Books Decision* (Oct. 15, 2015), https://publicknowledge.org/public-knowledge-welcomes-sweeping-victory-for-fair-use-in-google-books-decision/.

[9] Sega Enterprises Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992).

As a threshold matter, accessing, linking to, or interacting with digital information does not infringe any copyright. Reading a book, looking at a photograph, admiring a painting, or listening to music is not, and never should be, considered copyright infringement. This is not a "fair use" issue; the ability to use, access, or interact with a creative work is outside the bundle of specific, enumerated (and limited) rights granted by copyright.

As previously discussed, a GAI model does not contain an archive or directory of the works it trains on. The work analyzed during training is not reproduced or stored in the model; instead, the model uses each work to improve its overall understanding of whatever it is training on (art, photographs, writing, and so on) and the resulting model is a new, non-derivative construct. ML processes can be thought of as operating on the non-expressive, unprotectable ideas, abstractions, and functional elements of a given piece of training data.[10] As these elements of a work are not protected by copyright, and the works in the training data are not themselves reproduced by the training process, there is no basis for a copyright infringement claim based on the training process alone.

An alternative view is posed in *Andersen v. Stability AI*, the class action lawsuit against Midjourney and Stable Diffusion (two of the most widely used GAI art systems). In their complaint, the plaintiffs conceptualize the training process differently. They characterize the training process as a mechanism for storing the training images in a compressed state.[11] Ultimately, this interpretation will be tested in litigation—and it has already been dismissed with leave to amend by the district court[12]—but the generalization process that training images go through does not result in any conventional kind of storage. There are two facts that make this apparent.

First, in an ideal model, none of the content used in training can be reproduced by the system with any reliable fidelity. Second, the simple reality of the quantity of information these systems are trained on, compared to the size of the models themselves, should make that obvious. One can download the fully trained Stable Diffusion model weights at the size of 4 GB, while the LAION-2B dataset it is trained on — the smallest version — contains around 80,000 GB of images; no amount of

---

[10] See, *Baker v. Selden*, 101 US 99 (1880); 17 U.S.C. 102(b) ("In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.").

[11] Blake Brittain, *Judge pares down artists' AI copyright lawsuit against Midjourney, Stability AI,*Reuters (Oct. 30, 2023), https://www.reuters.com/legal/litigation/judge-pares-down-artists-ai-copyright-lawsuit-against-midjourney-stability-ai-2023-10-30/

[12] *Id.*

compression would allow for the model to contain all that information. Simply put, the model does not contain its training data within it.

When Analyzed Under Fair Use, AI Training Is a Fair Use

If one were to analyze the AI training process under the traditional fair use framework, the outcome would still favor a finding that the process constitutes a fair use of copyrighted materials.

For the first factor, AI training is a transformative process, akin to the transformative nature of the Google Books project and Accolade's reverse engineering in Sega v. Accolade. The purpose is not to replicate the copyrighted material for consumption but to derive patterns, structures, and understanding. The objective is knowledge extraction and the creation of a new entity (the trained model) that can generate or understand content based on patterns it has learned. Moreover, while some entities behind AI models may be commercial, the non-commercial research purposes behind many AI training endeavors, such as academic research or open-source projects, further bolster the fair use argument.

When it comes to the nature of the works used, AI training datasets often contain both factual and creative works. While creative works are typically given more protection, the manner in which they're used in AI training diminishes the significance of their creative nature. The focus is on extracting patterns, not on the expressive content of the works.

For the third factor, even though AI training as a process might use entire works, the essential activity does not reproduce these works but extracts general concepts into a new product that retains nothing of the training data. This is similar to how Accolade needed to initially copy Sega's code to understand it or how Google Books scanned entire books but displayed only snippets. The use of the entirety of a work is for the initial comprehensive understanding, not for replication or distribution.

Finally, under factor four, AI models do not serve as a market replacement for the copyrighted materials they train on. Instead, they can be seen as enhancing the market by providing new ways to interact with, understand, or generate content.

> *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?*

Entities collecting copyrighted material and curating it into datasets for training purposes should be evaluated based on the purpose and character of their own use.

As previously discussed, generally, the creation of datasets should be considered fair use on the basis of their transformative nature, the minimal contribution of each constituent work, and the nature of the use as preliminary to non-infringing creative activity. This analysis should be applied broadly to all entities engaged in similar conduct.

While, it is possible that commercial entities could be held to a different standard than entities that make datasets publicly available for research purposes, case law suggests that even if the end purpose is commercial and directly competitive to the copyright holder, preliminary copying for the purpose of learning is a protected fair use.

> *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?*

The shift from noncommercial to commercial use of AI models or datasets could alter a fair use analysis. Funding from for-profit developers may blur the noncommercial nature of research projects, but other factors are more significant, such as the level of accessibility of the resulting noncommercial components of the project; if AI companies are funding public accessible research that anyone can benefit from the fundamental purpose is retained, even if the companies are themselves able to use the resulting products. Fair use ought to be construed broadly and in the context that it is aligned with the core function of copyright: to promote "the Progress of Science and useful Arts." It is not a narrow exception to an absolute monopoly to be defeated by a mere shadow of economic competition.

> *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?*

GAI systems require vast amounts of training data. The volume of material used in training datasets, and in training GAI models, weighs against assigning individual copyright holders significant rights when analyzing GAI systems.

IV. Evaluating Restrictions on AI Training

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)?

It may be morally, economically, or otherwise valuable to establish non-copyright-based systems to give artists and other creators greater control over how their work is used in training GAI systems. Creators should feel empowered by publicly sharing their work online, rather than fearful about how their work might be exploited. Policies that encourage sharing build a more vibrant and creative open internet for all to enjoy.

AI companies should develop policies for honoring creators who want to "opt out" of their work being collected for AI training purposes. Any such system should provide artists and creators the ability to affirmatively express their preference to not have their work used in AI training in advance, and should also provide systems whereby they can later object to the use of their work in training, to prevent it from being further used for AI training.

An opt in, or affirmative consent, system would present an unacceptable barrier to fair use and creativity by creating a default of restriction. Given the vast amount of existing copyrighted work, and the quantity of work needed to train AI systems, any opt in system would be practically and logistically infeasible. Conversely, an opt out system requires creators to express a definitive preference to be more restrictive, creating a default of permissiveness that promotes an overall more open creative environment.

It is critical to note that any opt out system should cover only AI training. Opt outs should not limit other forms or purposes of web crawling, web scraping, or use of the work. Functions like historical web archiving and search indexing are vital for an open internet. Any system of opt outs must recognize and accommodate potential fair uses of copyrighted work.

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

No, consent should not be required for any use of copyrighted works to train AI models, whether they are commercial or noncommercial. It may be beneficial to establish systems to respect creator opt outs, however.

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?

Considerable work would be required to develop a process for administering opt outs. Technical tools, such as machine readable flags equivalent to robots.txt, would certainly be essential elements of any such system.

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?

Opt outs should not be legally mandated given that such a system would exceed the existing rights of copyright holders and burden fair use. However, a voluntary or industry standards-based system will present practical obstacles in coordinating actors in the GAI space.

There are also considerable technical challenges in ensuring machine readability, mechanisms for preserving flags while protecting privacy and free expression rights that could be burdened by onerous digital rights management solutions, and in conducting audits and assessments of datasets for compliance.

Yes, given the volume of works used in AI training an opt in system would be logistically and practically infeasible.

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?

Noncompliance with an AI training opt out regime should not give rise to copyright infringement claims because AI training does not infringe copyright.

Additionally, because of the high statutory penalties for infringement, and given the sheer size of training datasets, potential copyright infringement liability from accidental or unidentified non-compliance with opt-outs would create a risk profile that effectively requires a licensing model. For reasons discussed in response to Question 13, below, this is undesirable.

Any opt-out system should instead establish a separate system of enforcement and remedy, designed with the ultimate goal of encouraging compliance with tools and best practices for respecting opt outs, while not disincentivizing AI innovation, hampering fair use, or improperly expanding copyright to include an "access right."

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?

Maybe; an opt out system should not be tied to copyright, but adjudicating which creators have the right to restrict a work's training availability is a challenging policy question. For example, in the case of works with multiple authors or creators that may hold a creative interest in a work, like a film or song.

V. Transparency and Disclosure (Q. 15, 16)

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?

Transparency requirements regarding AI training data is an important requirement for a number of reasons including the identification of copyrighted material for the purpose of compliance with any opt out requirement regime.

Most significantly, dataset transparency enables more widespread scrutiny and understanding of bias, privacy, and other harms and risks posed by AI systems.

15.1. What level of specificity should be required?

Public disclosure of the full training data used should be the goal, and should satisfy any transparency or disclosure requirements.

Requiring AI developers to publish reports, summaries, and evaluations of training datasets could serve as a compliance pitfall for open-source, non-commercial, or competitive AI developers with more limited resources and team. Simple full disclosure of the full training datasets should therefore be the ideal, to ensure that compliance is as simple and straightforward as possible.

15.2. To whom should disclosures be made?

Disclosures should be public. Public disclosure enables independent auditing, evaluation, and assessment of AI training.

However, some models may use training data that is sensitive, contains personal information, or other data that should not be exposed publicly. For systems that use training data that should not be exposed publicly, reports or summaries regarding training data that protect privacy should instead be made available. The additional compliance burden that this might create is commensurate with the risks inherent

in using more sensitive training data. In other words, if a developer cannot afford to implement proper data reporting practices for handling sensitive training data, the regulatory burden is serving as an effective deterrent against an insufficiently resourced or dedicated developer working with sensitive data.

Additionally, an expert regulatory agency with clear jurisdiction over the development and deployment of AI systems would be able to conduct confidential audits of sensitive or private training data and could set adaptable standards for reporting requirements.

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 not be any notification requirement regarding AI training.

Training sets can be extremely large and the owner of all copyrighted material therein may be impossible to determine. Requiring notification would severely restrict the potential sources of AI training data, which would have considerable downstream negative effects on innovation, competition, model robustness, and prevalence of bias.

VI. Copyrightability and Authorship

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?

“How much human involvement is required to make a work eligible for copyright protection” is a question with both a clear doctrinal answer, and an unworkable practical answer. Doctrinally, there appears to be an uncharacteristic amount of consensus around those cases at either end of the spectrum: a work made entirely by artificial intelligence, with minimal human input, would not be eligible for copyright protection; a work made by a human author with minimal or non-substantial AI assistance would be. Most cases, however, will fall somewhere in the middle, with no clear answer in doctrine or practice. The Copyright Office has taken the position that any elements of a work generated by artificial intelligence are

unprotected by copyright.[13] While doctrinally sound, this raises endless enforcement questions, many of which are enumerated in the current NOI.

This is further complicated by the fact that GAI has already been in use across a number of creative fields for years. The Massive AI software suite has been used to generate large-scale melee battles for film since 2003.[14] Real-time augmented reality filters, such as those available on Snapchat, FaceTime, and Instagram, are powered by artificial intelligence and have been available for a decade. Music in particular has had a long relationship with GAI: voice-synthesizer Vocaloid, released in 2007, is so popular that the app's AI-powered cartoon avatar has gone on world tours annually since 2014;[15] David Bowie used a custom-built "Verbasizer" app to generate song lyrics as early as 1995;[16] and Adaptiverb, which uses machine learning to craft the perfect bespoke reverb tail has been commercially available since 2016.[17]

Rather than attempt to solve this on purely doctrinal grounds, we encourage Congress and the Copyright Office to examine the role of both registrability, as well as alternative non-copyright systems, as policy tools.

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?

The Copyright Act and existing case law clearly and sufficiently delineate the need for human authorship. The fact that they do so in the abstract, however, leaves how to implement such a standard open for debate. The most appropriate option would be to require applicants to "show their work" through process documentation. Some creative software (such as ProCreate, a visual art application for iOS) record process videos by default; however, the landscape of creative software is vast and varied, and imposing such a requirement may be beyond the technical reach of many applicants at this time.

---

[13] Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 88 Fed. Reg. 16,190 (16 Mar 2023) (37 CFR pt 202), https://www.copyright.gov/ai/ai_policy_guidance.pdf.

[14] Erin Carson, *How 'Lord of the Rings' Used AI to Change Big-Screen Battles Forever,* CNET (Sep. 4, 2020), https://www.cnet.com/culture/entertainment/features/how-lord-of-the-rings-used-ai-to-change-big-screen-battles-forever/. .

[15] https://mikuexpo.com/history

[16] Matthew Braga, The Verbasizer was David Bowie's 1995 Lyric-Writing Mac App, Vice (Jan. 11 2016) https://www.vice.com/en/article/xygxpn/the-verbasizer-was-david-bowies-1995-lyric-writing-mac-app.

[17] Apaptiverb Overview, https://www.zynaptiq.com/adaptiverb/adaptiverb-overview/.

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?

Protection of GAI-enabled work presents a conundrum. While it is, in the hands of individual creators, a tool like any other, it also holds the potential to displace significant swaths of the creative ecosystem if deployed irresponsibly. Sound policy must balance the potential upsides of these new tools against the economic precarity of existing creative workers. Registration cannot carry all the weight of developing balanced AI policy, but it is an important tool in the policy toolbox.

The policy realities are stark. Creative workers already face an unstable and unfavorable labor market. The market for creative work tends toward monopsony; buyers and publishers are highly concentrated, and wield significant price-setting power against a large (and largely diffuse) sea of independent suppliers.[18] In many creative industries—including fiction writing and music—these workers are considered independent contractors by default, with no ability to organize or collectively bargain. They experience higher-than-average unemployment rates and lower-than-average incomes.[19] To these workers, generative AI tools pose a significant risk of economic displacement, as certain kinds of complex creative work—such as video game asset design[20] and book cover art[21]—are being "streamlined" by AI tools in lieu of human creators.

In the short- to medium-term, generative AI provides the greatest benefit to larger, concentrated purchasers of creative work. One need look no further than this Office's roundtables; while major trade organizations offered uncharacteristically mild statements about AI and its role in the creative process, independent creators repeatedly raised the alarm about potential displacement. In addition to enhancing productivity, GAI allows major labels, publishers, and other major entertainment

---

[18] *See* United States v Bertelsmann SE & Co., et al, Civil Action No. 21-2886-FYP (D.D.C. Nov 14, 2022).
[19] Randy Cohen, Artists in the U.S. Workforce 2006-2020, Americans for the Arts (Mar 2021), https://www.americansforthearts.org/by-program/reports-and-data/legislation-policy/naappd/artists-in-the-us-workforce-2006-2020 .
[20] Shannon Liao, *A.I. May Help Design Your Favorite Video Game Character*, NYTimes (22 May 2023), https://www.nytimes.com/2023/05/22/arts/blizzard-diffusion-ai-video-games.html.
[21] Daniel Piper, *This AI-generated book cover is causing controversy*, Creative Bloq (16 May 2023), https://www.creativebloq.com/news/ai-book-cover; Lidna Codega, *Tor Tried to Hide AI Art on a Book Cover, and It Is a Mess*, Gizmodo (16 Dec 2022), https://gizmodo.com/tor-book-ai-art-cover-christopher-paolini-fractalverse-1849904058.

companies the ability to flood an already-crowded market with low-cost alternatives;[22] "sure thing" bets, such "new" music from deceased artists;[23] or cheap sequel scripts to existing franchises.[24] Any regulations designed at intervening in the market must ensure that those productivity gains are enjoyed primarily by the creative workers, rather than being used to solely enhance the balance sheets of employers, as has happened in the past.[25]

> *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?*

We believe that the idea of an alternative, sui generis right for GAI-enabled work is worth exploring. The benefits of such a system may include faster and cheaper registration, and a lowered standard of documentation to illustrate which parts are attributable to AI, and (potentially) provenance of the work's AI components, if those components are subsequently the subject of litigation. In exchange, such a regime would ideally provide protection for a limited term (such as 5 or 10 years from the day of publication, potentially with an option to renew) and access to the Copyright Claims Board for limited damage awards.

The benefits of such a system would, we believe, extend well beyond GAI-enabled work. Copyright is, in many ways, a system designed to default to the maximum: it grants all creators, regardless of their needs or desires, a sweeping array of rights and penalties that last well beyond the bounds of the creators' own lifetime. In addition to serving as a practical middle ground for the protection of GAI-enabled work, an alternative protection system with reasonably limited damages, term, and scope--as well as (ideally) faster and cheaper registration procedures--would more closely align with the needs of many creators. This acknowledgment of variable needs underlaid

---

[22] Amanda Hoover, *AI-Generated Music Is About to Flood Streaming Platforms*, Wired (17 Apr 2023), https://www.wired.com/story/ai-generated-music-streaming-services-copyright/.
[23] Mark Savage, *Sir Paul McCartney says artificial intelligence has enabled a 'final' Beatles song*, BBC (13 Jun 2023), https://www.bbc.com/news/entertainment-arts-65881813.
[24] Jake Coyle, *A.I. is one of the main reasons that Hollywood writers are on strike: 'Too many people are using it against us and using it to create mediocrity'*, Fortune (5 May 2023), https://fortune.com/2023/05/05/writers-strike-hollywood-ai-scripts/.
[25] *See, e.g.*, Lawrence Mishel, *Growing inequalities, reflecting growing employer power, have generated a productivity–pay gap since 1979*, Economic Policy Institute (2 Sep 2021), https://www.epi.org/blog/growing-inequalities-reflecting-growing-employer-power-have-generated-a-productivity-pay-gap-since-1979-productivity-has-grown-3-5-times-as-much-as-pay-for-the-typical-worker/; Michael Brill et al, *Understanding the labor productivity and compensation gap*, Beyond the Numbers: Productivity, vol. 6, no. 6 (U.S. Bureau of Labor Statistics, June 2017), https://www.bls.gov/opub/btn/volume-6/understanding-the-labor-productivity-and-compensation-gap.htm.

the creation of the Copyright Claims Board, and we believe could serve as the foundation of an alternative protection schema as well.

VII. 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?

Currently, AI-generated material featuring the name, likeness, or vocal likeness of a particular person may be subject to liability under state name/image/likeness (NIL) or right of publicity (ROP) laws. These are, however, subject to two major limitations. First, these protections are generally only available to individuals whose likeness holds independent commercial value, and thus do not provide protection for the vast majority of individuals.[26] Second, these rights exist only as an uneven patchwork, with mismatched duration and protection.

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?

We broadly support the creation of an individual right against being digitally replicated via GAI. The devil is, as always, in the details—particularly regarding the assignability, contours, and duration of the right, as well as any potential secondary liability issues that may implicate digital platforms to which such material is uploaded. Moreover, any universally-available right needs to adequately address the "digital doppelganger" problem—namely, ways of dealing with situations in which an AI-generated work, by pure mathematical chance, looks or sounds like an otherwise unknown individual. Such instances should not give rise to liability, or trigger a rabbit hole of provenance questions about the training data of the GAI system that generated the accidental lookalike.

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?

---

[26] A notable exception to this is New York's right of publicity statute, which provides a cause of action against anyone who "discloses, disseminates or publishes sexually explicit material[s]" that includes "computer-generated nude body parts as the nude body parts of the depicted individual or the depicted individual engaging in sexual conduct … in which the depicted individual did not engage." NY S.B. 5959, available at https://legislation.nysenate.gov/pdf/bills/2019/S5959D.

No. Human artists have always mimicked the style and substance of one another's work, be it for satire, homage, or simply to further their own artistic skill. Permissible stylistic mimicry is not spontaneously rendered illegal because it is a computer, and not a human, doing the mimicking. Moreover, passing off a mimickry as the artist's own work already violates existing law. Protecting "style" would prove unworkably broad, severely curtail freedom of expression, and be fundamentally un-administrable either by the courts or by the current registration apparatus.

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

Although it is beyond the policy ambit of the Copyright Office specifically, we believe it bears repeating: collective labor has a significant role to play in shaping how market forces utilize AI, both in creative fields and more generally across the market. The strike by Writers' Guild of America West is an useful illustration of how workers concerned about strategic use of GAI can impact the parameters for appropriate development, deployment, and use of GAI within an existing industry. By negotiating contract terms that clearly delimit the scope, scale, and context of AI usage by both studios and writers, WGA has sought to craft a solution that will allow creative workers to share in the benefits of AI, while protecting them from the worst of its potential harms.

AI is, as noted above, a wildly diverse tool with a range of applications we cannot fully imagine today. Its use and role in the economy will continue to grow and evolve. The government has a critical and central role to play; it is rarely, however, the "first to know" of new developments. Because of this, we believe that the best "first responders" to the evolving role and use of AI are the workers whose jobs are directly impacted. We believe that organization among workers should be widely encouraged not only for its own sake, but as a way to check irresponsible growth and mismanagement of a tool that has the ability to fundamentally reshape the economy.