SUBMITTED BY DANIELLE AGUIRRE, SHANNON SORENSEN AND ERIC SUNRAY
ON BEHALF OF THE NATIONAL MUSIC PUBLISHERS' ASSOCIATION

Before the
UNITED STATES COPYRIGHT OFFICE
The Library of Congress

Artificial Intelligence and Copyright	Docket No. 2023-6

NATIONAL MUSIC PUBLISHERS' ASSOCIATION COMMENTS IN RESPONSE TO THE NOTICE OF INQUIRY

Introduction

NMPA, founded in 1917, is the principal trade association representing the interests of music publishers and their songwriting partners in the United States. NMPA's members represent the substantial majority of the market for U.S. copyrighted musical works. NMPA is the voice of both small and large music publishers and is the leading advocate for publishers and their songwriting partners in Congress and the courts.

NMPA welcomes the Office's Notice of Inquiry (88 Fed. Reg. 59,942 (Aug. 30, 2023); "Notice") on this topic of momentous significance to the creators and owners of creative works, including musical works owned or controlled by NMPA's members. NMPA has been actively engaged in the conversation around generative AI and its relationship with, and impact on, the creative economy. We are a founding member of the Human Artistry Campaign, and have participated in numerous panels, roundtables and other industry discussions, including the Office's listening sessions, on topics raised in the Notice.

At the outset, we emphasize that our membership is not "opposed to AI." There is widespread belief in the music industry that great benefits could come from generative AI systems that can assist human creators. However, the development of the generative AI marketplace is marked by breathtaking speed, size and complexity. Hindsight may well prove that there is no hyperbole in saying that generative AI is the greatest risk to the human creative class that has ever existed. Even more alarming is that we do not know how long the window is to act before it is too late. We therefore implore the Office to support proactive protections for human creators and, where there is uncertainty, to err on the side of protecting human creators. The risks between over- and under-protection are profoundly asymmetrical. There is immense danger of irreparable harm if the unauthorized exploitation of copyrighted works in the development of AI models is left unchecked. In contrast, speed bumps in the development of AI models do not implicate the same existential risks, but rather have even been advocated as desirable by many leading figures in the industry.[1]

Below we address some of the questions raised by the Office in the Notice. We are available to discuss further these and other questions that the Office may have on these important topics.

Research Papers & Studies (Question 3[2])

Summary: The AI research field is incredibly vast. Accordingly, we present several noteworthy research papers and studies organized by subject matter, including information relating to popular datasets and specific AI model architectures and their functionality.

AI research has been ongoing since the 1950s. Advances in machine learning have inspired a bevy of research into topics ranging from innovative model designs to novel training methods to information regarding the datasets themselves. In addition to the research mentioned below, we cite to several other relevant studies throughout our submission.

We note that the NMPA does not endorse the content of these papers and studies by mentioning them. Rather, we present them to highlight important developments in the AI field, as well as information relevant to musical work rightsholders.

A. Datasets

As data is a source of significant competitive advantage in the modern AI industry, there is great interest regarding dataset contents and how they are aggregated. Noteworthy research papers regarding machine learning datasets include:

- The Pile: An 800GB Dataset of Diverse Text for Language Modeling.[3]
- Disco-10M: A Large Scale Music Dataset.[4]
- Datasets that promote multimodal capabilities (*e.g.*, text-to-audio, text-to-image, etc.): LAION-5B.[5]

[1] Connie Loizos, *1,100+ notable signatories just signed an open letter asking "all AI labs to immediately pause for at least 6 months,"* TechCrunch (Mar. 29, 2023), available at https://techcrunch.com/2023/03/28/1100-notable-signatories-just-signed-an-open-letter-asking-all-ai-labs-to-immediately-pause-for-at-least-6-months/.

[2] *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.*

[3] Leo Gao et al. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv:2101.00027v1 (retrieved from https://arxiv.org/pdf/2101.00027v1.pdf).

[4] Luca A. Lanzendörfer et al. (2023). DISCO-10M: A Large-Scale Music Dataset. arXiv:2306.13512v2 (retrieved from https://arxiv.org/pdf/2306.13512.pdf).

[5] Christoph Schuhmann et al. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402v1 (retrieved from https://arxiv.org/pdf/2210.08402.pdf).

B. Model Architecture and Functionality

There is a plethora of research into novel machine learning techniques and model designs. As explained further in response to Question 6, musical works are used extensively in music models as well as large language models ("LLM"). Noteworthy research on specific models includes:

- OpenAI's *Jukebox* whitepaper, describing how OpenAI trained a generative music model on 1.2 million raw audio samples and their associated lyrics.[6]
- OpenAI's *GPT-3* whitepaper, reporting the various datasets and training methods used to build the LLM that powered earlier versions of ChatGPT.[7]
- Google's *MusicLM* whitepaper, describing how it trained a generative music model on hundreds of thousands of hours of music.[8]
- Meta's *MusicGen* whitepaper, reporting greater model performance on a smaller set of licensed data compared to competing models trained on unlicensed data.[9]
- Microsoft's Muzic Research publication database,[10] containing several papers on topics ranging from automatic songwriting[11] to singing voice synthesis using data mined from the web.[12]

Researchers also frequently publish papers on AI model functionality. Noteworthy examples include research on:

- A model's ability to memorize individual examples from its training set, which are then emitted as output.[13]
- The likelihood that LLMs will reproduce training data verbatim when trained on large, private datasets.[14]

[6] Prafulla Dhariwal et al. (2020). Jukebox: A Generative Model for Music. arXiv preprint arXiv:2005.00341 (retrieved from https://arxiv.org/pdf/2005.00341.pdf).

[7] Tom B. Brown et al. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*, Article No.: 159, pp. 1877–1901 (retrieved from https://dlnext.acm.org/doi/abs/10.5555/3495724.3495883).

[8] Andrea Agostinelli et al. (2023). MusicLM: Generating Music From Text. arXiv preprint arXiv:2301.11325. (retrieved from https://arxiv.org/pdf/2301.11325.pdf).

[9] Jade Copet et al. (2023). Simple and Controllable Music Generation. arXiv preprint arXiv:2306.05284 (retrieved from https://arxiv.org/pdf/2306.05284.pdf).

[10] *AI Music: Publications - Microsoft Research* (retrieved from https://www.microsoft.com/en-us/research/project/ai-music/publications/).

[11] Zhonghao Sheng et al. (2021). SongMASS: Automatic Song Writing with Pre-training and Alignment Constraint. Microsoft Research (retrieved from https://www.microsoft.com/en-us/research/uploads/prod/2020/12/SongMASS.pdf).

[12] Yi Ren et al. (2020). DeepSinger: Singing Voice Synthesis with Data Mined From the Web. arXiv:2007.04590 (retrieved from https://arxiv.org/pdf/2007.04590.pdf).

[13] Nicholas Carlini et al. (2022). Extracting Training Data from Diffusion Models. arXiv:2301.13188 (retrieved from https://arxiv.org/pdf/2301.13188.pdf).

[14] Nicholas Carlini et al. (2021). Extracting Training Data from Large Language Models. arXiv:2012.07805 (retrieved from https://arxiv.org/pdf/2012.07805.pdf).

- AI model reverse-engineering and tracing model outputs to the relevant training data.[15]

C. Miscellaneous

Additional noteworthy research on AI-related topics includes:

- A framework for promoting AI model transparency.[16]
- The Berkeley AI Research team's study on improving model performance by training on smaller, high-quality datasets.[17]
- Industry research studies on the modern AI music ecosystem.[18]
- A Stanford Institute for Human-Centered Artificial Intelligence study on the levels of transparency among major foundation model developers.[19]

International Approaches (Question 4[20])

Summary: The U.S. should not follow current international approaches to AI and copyright that include broad limitations or exemptions to copyright protection that harm creators and may violate international treaties. The U.S. should work with its international counterparts to promote policies that protect copyright and oppose policies that erode them, including opposing categorical exemptions for text and data mining ("TDM"), opt-out regimes or waivers of lawful access requirements.

The U.S. should continue to work with its international counterparts to support and promote policies that protect copyright and creators internationally. The U.S. should oppose any attempts to create broad and novel copyright exemptions that would violate international treaty obligations. While international protection for rightsholders is important and allows for the export of U.S. creative works, harmony in international law should never be prioritized over robust domestic copyright protections.

Only a few jurisdictions have enacted legislation that relates to AI and copyright. NMPA notes them here, but does not endorse consideration of any of these approaches in the U.S. Not only

[15] Tracing Model Outputs to the Training Data (2023), available at https://www.anthropic.com/index/influence-functions.

[16] Margaret Mitchell et al. (2019). Model Cards for Model Reporting. arXiv:1810.03993 (retrieved from https://arxiv.org/pdf/1810.03993.pdf).

[17] Xinyang Geng et al. (2023). Koala: A Dialogue Model for Academic Research. Berkeley Artificial Intelligence Research, available at https://bair.berkeley.edu/blog/2023/04/03/koala/.

[18] Stream Report Season 3: Creative AI, Water & Music (2023) (retrieved from https://stream.waterandmusic.com/#summaries).

[19] Rishi Bommasani et al. (2023). The Foundation Model Transparency Index, available at https://crfm.stanford.edu/fmti/fmti.pdf.

[20] *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?*

do they fail to provide adequate protections for copyright owners, they could also result in noncompliance with the U.S.'s treaty obligations.

- *European Union:* The European Union has two provisions for TDM exemptions. Article 3 of the European Union Digital Single Market Copyright Directive (the "EU DSM Copyright Directive") provides a TDM exemption for scientific research purposes that is conducted by non-commercial research organizations or cultural heritage institutions.[21] Copyright owners cannot opt out of TDM under Article 3. Article 4 provides for TDM exemptions for commercial uses, subject to an opt-out provision requiring a machine-readable opt-out request from the rightsholder.[22] Both Article 3 and Article 4 exemptions only apply if there was lawful access to the copyrighted works in question.

- *UK:* The UK has a TDM exemption for computational data analysis which applies only to non-commercial research purposes.[23] Lawful access is required.

- *Singapore:* Singapore's copyright law provides an exemption to reproduction rights for computational data analysis which applies to both commercial and noncommercial uses. Lawful access is required. With regard to opt-out, not only do rightsholders have no ability to opt out, but the law goes one step further to proactively void any contract provisions that contradict the exemption.[24]

- *Japan:* Japan's TDM exemption applies to all commercial and noncommercial uses so long as the purpose is not "to personally enjoy" the work and the use does not "unreasonably prejudice the interests of the copyright owner."[25] There is no opt-out provision and there is no lawful access requirement.

The U.S. should not consider or support any approaches to AI regulation that would limit the scope of copyright protections or would lead to widespread waiver or loss of copyright protections. Categorical exemptions from copyright liability violate the three-step test of the Berne Convention, which limits permissible exceptions to the reproduction right only to "certain special cases, provided that such reproduction does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author."[26] Sweeping carveouts to copyright protection, such as those noted above for TDM, go significantly beyond "special cases." Such carveouts may also stifle the developing marketplace for licensing copyrighted works for use in AI development and thereby run afoul of the prohibition against interference with the normal exploitation of the work by the owner.

[21] EU DSM Copyright Directive, Article 3.

[22] EU DSM Copyright Directive, Article 4.

[23] UK Copyright Law, § 29A.

[24] Singapore Copyright Act of 2021, Article 243, 244.

[25] Japan Copyright Act, Article 30-4.

[26] Berne Convention for the Protection of Literary and Artistic Works (1979), Article 9(2).

The U.S. should similarly oppose (and not consider) any opt-out regime. As discussed further in our response to Question 9, such systems place an immense burden on copyright owners to prevent the theft of their works, rather than appropriately placing the burden on developers of AI models to affirmatively license the content they wish to use. Opt-out regimes may also violate Article 5(2) of the Berne Convention which prohibits conditioning copyright protection upon any formality.[27] Furthermore, the U.S. should consider the impact that different jurisdictions implementing different opt-out regimes would have on creators: a patchwork of international exemptions with varying opt-out requirements would be difficult if not impossible for most rightsholders to navigate.

Legislation (Question 5[28])

Summary: New legislation regarding infringement actions appears premature at this stage. However, legislation and regulation are needed to address recordkeeping, data transparency and labeling concerns.

NMPA does not believe that, at this point in time, amendments to the Copyright Act are necessary to address infringement actions concerning the training of AI or in its outputs. Our existing copyright law framework should prove sufficiently robust and flexible to adequately address issues of infringement in the creation of datasets, training of AI models, and outputs of generative AI software. We are mindful though that, in the future, amendments to address these areas may be necessary given the dynamic nature of the AI landscape and the current lack of substantial jurisprudence applying the Copyright Act to AI uses.

However, legislation and regulation are needed now to address issues related to recordkeeping, data transparency and labeling. We discuss these matters further in response to Questions 15-16, 26, and 28-29 below.

[27] *Id.* at Article 5(2). ("The enjoyment and the exercise of these rights shall not be subject to any formality . . . ."). *See* Jane Ginsburg, "Berne-Forbidden Formalities and Mass Digitization," Boston University Law Review, 96:745, 750 (2016) ("A member state may neither condition the initial attachment of copyright on compliance with formalities nor subsequently deny coverage of particular rights to authors who fail to meet declaratory obligations.")

[28] *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.*

Training and Data Provenance (Questions 6, 7 and 12[29])

Summary: Training datasets are often built via large-scale web-scraping, either at the developer's direction or by a third-party data aggregator. Such web-scraping practices inevitably capture copyrighted works, which directly implicates the exclusive rights conferred via copyright law. Despite attempts to downplay the significance of specific works in training sets, post-training adjustments to an AI model do not account for the unauthorized use of protected works in the first instance. Moreover, it is neither practical nor consistent with copyright law to place the onus on the copyright owner to trace the use of a given work within a model from training through output. The law requires affirmative consent from the copyright owner for training use as a threshold matter.

A. AI models can presently train on any content available in digital format, and developers often obtain content in bulk by scraping the Internet.

There is virtually no limit to the kinds of copyright-protected material that can be used to train an AI model. Anything existing in digital format on the web can be taken and subsequently used for AI training.

The nature of a musical work poses unique challenges for rightsholders, as the associated copyright protects both music and lyrics. Accordingly, music publishers' works are taken from the web in a variety of formats, including but not limited to song lyrics as text, MIDI files, tablature, and digital audio and audiovisual files containing musical works (*e.g.*, sound recordings, music videos, film/tv, short-form video, etc.). Music has even been taken as images, in the form of spectrograms

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

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

6.2. To what extent are copyrighted works licensed from copyright owners for use as training materials?

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

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

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.

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

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

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?

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.

that can be generated from music audio and used to train AI image generators that can then create derivative images that can be transposed back into derivative audio musical works.[30] The breadth of ways that AI models can exploit protected music is remarkable and still evolving.

With respect to the question of training sets based on materials collected by third-party entities rather than developers themselves, we note first that distinctions based on which corporate entity collected data should never be a shield from liability for an AI developer that uses protected content without authorization. AI developers must remain responsible for their use of content, and pirated content can never be laundered into legitimacy by passing between hands.

The use of pre-existing datasets containing copyrighted material is widespread and common. The fact that many of these datasets are open source and not covered by any warranties or indemnification from the dataset provider speaks to a troubling casualness among many AI developers in utilizing content for training without regard to copyright protection. Noteworthy examples of third-party datasets include the Common Crawl ("CC") corpus, which is a publicly available collection of large-scale web data used as the primary training corpus for most major LLMs. In fact, OpenAI's whitepaper introducing its GPT-3 model reported that 60% of its data was drawn from the CC corpus.[31] Similarly, the non-profit research group Eleuther AI compiled the WebText2 dataset, which reportedly constitutes 22% of GPT-3's data, and the Pile dataset (discussed in response to Question 3).

There are also massive online communities of data scientists and machine learning enthusiasts who routinely publish datasets for public use.[32] To be clear, copying protected works to create a machine learning dataset constitutes its own distinct infringement, even if the dataset is never used to train a model (such use would constitute a separate infringement). Publishing an infringing dataset online for others to view and download also constitutes an infringement of the copyright owner's exclusive distribution right and may further implicate the public display and/or performance rights. But importantly, none of these infringements by dataset collectors can diminish the liability of an AI developer that takes the datasets and exploits them to train its systems.

Large-scale web aggregators like CC crawl the web to identify as diverse an array of content as possible. Developers then sift through these massive datasets to further curate them by deduplicating and filtering out unwanted (but not unlicensed) content. Accordingly, websites that host significant amounts of high-quality, human-generated content, like the Internet Archive and Reddit, are popular crawl targets, despite already being replete with unlicensed material.

Datasets can vary in size, from bespoke collections designed for narrow use-cases to petabytes worth of diverse material.[33] Foundation models, such as LLMs, typically use larger datasets so they

[30] Michael Kan, *AI Image Generator Can Also Produce Music (With Otherworldly Results)*, PC World (December 15, 2022), available at https://www.pcmag.com/news/ai-image-generator-can-also-produce-music-with-otherworldly-results.

[31] Tom B. Brown et al. (2020). Language models are few-shot learners: Supplemental Material, at 5 (retrieved from https://dlnext.acm.org/action/downloadSupplement?doi=10.5555%2F3495724.3495883&file=3495724.3495883_supp.pdf).

[32] *See, e.g.*, Kaggle.com and Huggingface.co.

[33] *See, e.g.*, List of datasets for machine-learning research, Wikipedia, available at https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research.

can later be applied to a variety of downstream tasks. Such models are adapted for these downstream tasks via a process called "fine-tuning," which entails further conditioning the pre-trained model on a smaller dataset supplied by the fine-tuning developer. This practice allows the fine-tuning developer to benefit from the initial developer's infringement.[34] Thus, even where a secondary developer may opt to use a smaller dataset of licensed and/or public domain material, the underlying foundation model's vast training corpus could nonetheless contain unlicensed copyrighted works.

B. Several direct licensing agreements for training material already exist in the market.

Many AI companies have announced licensing deals with copyright owners for use of their works. Noteworthy publicly known examples include:

- OpenAI and The Associated Press ("AP") to train future GPT-N models on AP's news stories.[35]
- Meta and Shutterstock to train its generative music platform—MusicGen—on Shutterstock's audio library.[36]
- Stability AI and commercial music library Audiosparx to train Stability AI's most recent generative music model, Stable Audio.[37]

C. Copyrighted works are valuable training materials.

Notably, large-scale web-scraping is generally pervasive and indiscriminate; thus larger crawls inevitably yield much protected content as well as some public domain content. Open AI itself made this clear, citing its experience developing AI, "including by the use of large, publicly available datasets that include copyrighted works," and noting that "[a]s copyright protection arises automatically when an author creates an original work and fixes it in a tangible medium, see 17 U.S.C. § 102, the vast majority of content posted online is protected by U.S. copyright laws."[38]

The same features that make many copyrighted works valuable in general—such as the richness, contemporary relevance and general quality of the content borne from an investment of time and resources made by creators in reliance on the incentives of copyright protection, as well as the structure, integrity and reliability, formatting consistency, reduced noise, and other enhancements that come from the pre-vetting of much commercial copyrighted content—also make copyrighted works particularly valuable for AI developers, as these qualities help AI developers accomplish their goal of

[34] Liability and transparency in this context are discussed further in response to Question 15.

[35] Matt O'Brien, *ChatGPT-maker OpenAI signs deal with AP to license news stories*, AP (July 13, 2023), available at https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

[36] Kyle Wiggers, *Meta open sources an AI-powered music generator*, TechCrunch (June 12, 2023), available at https://techcrunch.com/2023/06/12/meta-open-sources-an-ai-powered-music-generator.

[37] Stability AI, *Stable Audio: Fast Timing-Conditioned Latent Audio Diffusion*, stability.ai (Sept. 13, 2023), https://stability.ai/research/stable-audio-efficient-timing-latent-diffusion.

[38] OpenAI, Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation, U.S. Patent and Trademark Office Dkt. No. PTO-C-2019-0038 (2019) at 1 & n.1 (retrieved from https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf) ("OpenAI USPTO Comment").

generating content that modern users will pay for. This value drives the growing market for licensing copyrighted content for use in training sets.

D. The training process implicates the exclusive reproduction and derivative work rights.

Training methods are diverse, which makes generalizing the AI training process exceedingly difficult. The approach taken by a given developer will depend on the model's intended use case. At a high-level, AI models can be "rules-based," which means they are programmed to follow a set of pre-determined rules to solve a given task. However, training generative AI models on pre-existing works is increasingly common and constitutes a significant concern to rightsholders.

When a pre-existing work is used to train an AI model, it is analyzed in its entirety. For some models, developers will compress each training example into a compact representation and then cause the developing model to predictively reconstruct it.[39] Others might introduce distortive "noise" to the training material, aiming to have the model reproduce the original training example from this noise.[40] In each case the model's parameters are adjusted (explained below) to correct for errors in the model's generative process. Alternatively, some AI models break down each work from the training set into "tokens." These are granular, bite-sized components of the overall work, such as individual syllables in a line of text. The model then gauges each token's significance within the training, allowing it to predict the likelihood of that token's appearance in new contexts. In some cases, AI systems may employ multiple models that utilize a combination of the above training methods.

To aid the model in the training process, AI developers may label elements of a dataset with certain identifying tags to assist the inferences the model draws from the training material. This training method is broadly known as supervised learning, which by its nature requires significant curatorial discretion. Conversely, unsupervised learning occurs when a model is given free rein to identify patterns and dependencies within a training set on its own. This approach is less controllable but nonetheless yields insights into the training material that may otherwise be undetectable to the human eye. In general, the training process can take anywhere from hours to months depending on various factors such as the amount of training material used, the number of parameters in the model, and the developer's computational capacity.

Regarding the reproduction right, many of the most common training methods deliberately seek to reconstruct training material after it has been intentionally compressed or distorted. This is straightforward copying. Further, exposing training material to human annotators for supervised learning necessarily entails making a copy.

Regarding the right to make and authorize derivative works, the Copyright Act defines a derivative work as a work based on one or more pre-existing works in "any . . . form in which [the]

[39] *E.g.*, autoencoder-based models process data by compressing it into a lower-dimensional space where the model creates an encoded representation of patterns and relationships identified among the data. Then the data is reconstructed via a decoder programmed to interpret the encoded representation, with the goal of perfectly mirroring the input.

[40] *E.g.*, diffusion models process data by progressively adding "noise" (static) to the training sample, with the goal of having the model independently "denoise" the data, thereby *reproducing* the input in its original form. Errors in the denoising process reflect an imbalance in the model's parameters, which are subsequently adjusted via a feedback process called backpropagation.

work may be recast, transformed, or adapted."[41] As discussed further below, the training material's expressive qualities are abstracted and recast within the model's parameters. AI researchers have themselves explained that, in the training process, training material is "*transformed* and modeled in a very different representation of weights and biases . . . . *[I]t is derivative work*[.]"[42] (emphasis added). Ultimately, the model becomes an abstract agglomeration of its training material capable of generating (*i.e.*, communicating) verbatim copies of works within the training set, many of which are copyrighted. Such qualities fall squarely within the Copyright Act's definition of a derivative work.

E. Retention of training data is irrelevant to the infringement inquiry.

To the extent AI developers argue that they cannot be liable for infringement because they do not retain copies of their training material once training is complete, this justification is inconsistent with clearly established legal standards. Reproducing a work in copies or phonorecords without authorization violates Section 106(1) of the U.S. Copyright Act, and copies and phonorecords are material objects in which works are "fixed." 17 U.S.C. § 101 (definitions of "copies" and "phonorecords"). A work is "fixed" if it is embodied in a copy or phonorecord "for a period of more than transitory duration." *Id.* (definition of "fixed"). Accordingly, whether a work is retained after training is largely irrelevant to the infringement analysis. If a copy is made at any stage of AI development for longer than a transitory duration, there is infringement.

Moreover, as further discussed below, discarding training materials once training is complete does not account for the copy that remains embodied within the model's parameters. Thus, regardless of whether training data is "stored" in a traditional sense, each training example has an essential, inextricable influence on the model's parameters, which subsists within the model in perpetuity.

F. Inferences gained from the training process are often represented within the model's parameters.

As previously noted, the diverse array of AI architectures and training methodologies makes it difficult to generalize with respect to AI models and how they are designed. However, inferences gained by models developed via deep learning and other related approaches are often reflected in the neural network's parameters. These parameters are calculations that instruct the model on whether it should give greater or lesser weight to a given element from the training set when called upon to generate output. Large models can contain billions, if not trillions, of parameters.

As discussed above, the works ingested during training may, in some cases, ultimately be discarded after they've been processed by the model, leaving only the algorithmic representations of how works of a given type are constructed. This novel storage scheme has led AI developers to argue that "copies" of protectable expression are not made during the training process, but that is misleading. Such reasoning merely seeks to justify the unauthorized use of protected works by sidestepping traditional understandings of copying, despite the AI model serving essentially the same fundamental purpose as a traditional storage device.

[41] 17 U.S.C. 101 (definition of "derivative work").

[42] Sharon Goldman, *The data that trains AI is under the spotlight — and even I'm weirded out*, VentureBeat (Apr. 24, 2023), available at https://venturebeat.com/ai/the-data-that-trains-ai-is-under-the-spotlight-and-even-im-weirded-out-the-ai-beat/ (emphasis added).

Moreover, this reasoning is inconsistent with how the Copyright Act defines "copies." "Copies" are defined as "material objects . . . from which the work can be perceived, reproduced, or otherwise communicated." 17 U.S.C. § 101 (definition of "copies"). AI can generate substantially similar—if not verbatim—copies of its training material. Thus, protectable expression clearly resides within the model, regardless of how it is represented.

G. It is technically possible for AI models to "unlearn" inferences gained from training, but such approaches have inherent limitations.

While it may be technically possible to adjust a model's parameters to diminish a specific data point's influence on the model, such approaches have inherent limitations and would not account for use of the material in the first instance. Researchers have proposed various "machine unlearning" methods to surgically deconstruct AI models. One such technique is "exact unlearning," which seeks to completely remove the influence of specific training material such that the model behaves as if it had never been used at all. However, there are several limitations to this approach:

1) As noted above, AI models may contain billions of parameters, which makes it difficult to determine how a given piece of training material influences the model's learned patterns and dependencies.
2) Surgically removing training material can enable bad actors to infer information about the model, thereby leaving it vulnerable to adversarial attacks.
3) The parameter adjustments required for exact unlearning entail high computational costs.

Alternatively, "approximate unlearning" methods can diminish a given piece of training material's influence on the model. Such methods may be more computationally efficient and cost-effective than exact unlearning, but they do not remove the training material's influence entirely.

H. It may be possible to infer use of a given work by inspecting the model's output, but this should not be the copyright owner's only means of accessing the underlying dataset.

It may be possible to infer a model's use of a given piece of training material by inspecting system output.[43] AI systems can reproduce verbatim copies and/or detailed summaries of copyrighted works, reflecting that the model was trained on those works.

However, the rapid rate of technological innovation in this space may make it increasingly difficult to draw such inferences from the model's output. For example, retrieval augmented generation ("RAG") is a novel AI feature that enables LLMs to generate contextually relevant output by querying remote data sources, thereby eliminating the need to rely solely on the model's underlying training material.

Moreover, reliance on an AI system's output as a forensic tool to detect training-based infringement is in tension with the need for implementing safeguards against output-based infringement, like content filters that prevent the display of copyrighted song lyrics. The more such

[43] *See, e.g.*, Complaint ¶¶ 61, 66-69, 73-74, 78-79 *Concord Music Grp., Inc. v. Anthropic PBC*, Case 3:23-cv-01092 (M.D. Tenn. Oct. 18, 2023) (alleging "the fact that Anthropic's AI models respond to user prompts by generating identical or near-identical copies of Publishers' copyrighted lyrics makes clear that Anthropic fed the models copies of those lyrics when developing the programs," and providing examples of such "identical or near-identical" output).

protective measures are implemented effectively, the more they may inadvertently conceal evidence of training-related infringement.

I. It is burdensome, impractical, and not the rightsholder's responsibility to investigate how AI models use their works to generate output.

To the extent a model is engaged to generate output that reproduces some or all of a given work verbatim (*e.g.*, a copyrighted song lyric), such work's contribution to the output is apparent. For AI systems capable of generating output that is influenced by many different training examples, identifying the influence of any particular work on the output may be more difficult. However, there is an entire field of research into AI model reverse-engineering known as "mechanistic interpretability."

Mechanistic interpretability asks whether examining the connective circuits within a neural network can illuminate how a model reacts when called upon to generate a given output. For example, researchers are presently attempting to trace AI output back to the model's most relevant training material by analyzing how the model's parameters (and, by extension, its outputs) change if a given piece of training material is copied and added to the training corpus. Thus, by examining the model's behavior in response to these marginal adjustments, researchers can infer how particular training material influences a model's output.

Unfortunately, current mechanistic interpretability methods can at best only approximate how a large model's connective circuitry functions at scale. Accordingly, it is unduly burdensome for rightsholders to enforce their rights where implicated, and unreasonable to require causal proof of a given output from owners of the associated training material. Moreover, the notion that rightsholders may be able to ascertain the source of a given output should not be treated as a solution to the generative "black box" issue. These investigative measures are antithetical to the law as it is currently written, which requires affirmative consent from rightsholders for use of their works.

Training and Fair Use (Question 8[44])

Summary: The fair use analysis is highly fact specific and no type of generative AI system use would be amenable to a categorical determination of fair use. However, on examining the current marketplace, all four fair use factors would likely weigh against a finding of fair use in the context of AI training activities.

A. Fair use determinations are highly fact specific.

Fair use is an affirmative defense to infringement of one or more of the exclusive rights granted to a copyright owner in Sections 106 and 106A of the Copyright Act. Whether a use is

[44] *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.*

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?

considered a fair use such that infringement may be justified is a highly fact-specific inquiry, to be determined by analyzing the specific factual circumstances of the use under the four non-exhaustive fair use factors set forth in Section 107 of the Copyright Act. For that reason, rulemaking is ill-suited to a determination that any particular "use" in general or in the abstract—including the "unauthorized use of copyrighted works to train AI models"—could constitute a "fair use." *See, e.g.*, *Campbell v. Acuff-Rose Music, Inc.,* 510 U.S. 569, 577 (1994) (the statute "calls for case-by-case analysis" and "is not to be simplified with bright line rules").

Instead, where the unauthorized use of a copyrighted work implicates an exclusive right granted to copyright owners, Section 107 provides the framework for individualized fair use analysis, including the four non-exclusive factors courts must consider. Courts are very capable of deciding fair use issues and can, in determining whether any particular unauthorized use of copyrighted materials in the context of AI systems constitutes fair use, draw upon decades of fair use jurisprudence in making such determination.

With that said (and recognizing that each user will likely claim that its particular use is factually distinct), with respect to the types of unauthorized uses of copyrighted works currently being made by when training AI models, the four fair use factors are likely to weigh against a finding of fair use.[45]

B. Current uses of copyrighted works to train AI models are commercial, non-transformative uses under the first factor.

The first fair use factor analyzes "the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes." 17 U.S.C. § 107(1). In the current AI landscape, the training of AI models fuels a multi-billion dollar commercial industry. The developers of AI models exploit copyrighted works in training, using them to create commercial services that can provide competing works. Neither the developers nor the models take in "ideas." Rather, they take creative expression from copyrighted content for the purpose of generating other content that serves the same purpose as the content from which they take. They train on expressive works to generate other expressive works. They copy expression for expression's sake. That is the purpose and character of their use. OpenAI has made this point clearly in explaining the industry's business model:

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?

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?

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?

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?

[45] The NMPA primarily discusses these fair use questions in the context of unauthorized reproductions of musical works that are made in the process of developing AI systems. Any reference to reproduction or to musical works is not intended as a limitation on other exclusive rights or types of works which may be infringed, and to which these principles should also apply.

> By analyzing large corpora (which necessarily involves first **making copies of the data** to be analyzed), AI systems can learn patterns inherent in human-generated data and then **use those patterns to synthesize similar data** which yield increasingly compelling novel media . . . . **OpenAI's MuseNet music generation model**, which uses a similar underlying algorithm as GPT-2, **was trained on thousands of MIDI audio files**. **It can output unique MIDI files meant to sound like a specific genre or artist** . . . .[46]

Claims that training is transformative fair use cannot escape the reality—explained by the industry leader and obvious to any user of AI models—that the purpose of training is to create systems that *are able* "to synthesize similar data," for example creating musical works that "sound like" works in the training set. As this indicates, AI model training involves using exact copies of entire copyrighted works for the purpose of "synthesizing" "similar" works that are "compelling . . . media." Nothing in the training process involves commenting on or criticizing the original work; the focus is admittedly on replication and imitation. Further, the law is clear that any ultimate fair use that an end user might later make of an AI model cannot insulate the developer from liability for the unauthorized copying in creating the commercial service sold to the end user that is designed to be able to yield compelling media that is similar to the creative works copied without authorization.[47]

Neither are these unauthorized reproductions transformative in purpose, as the Supreme Court's recent decision in *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258 (2023) ("*Warhol*") confirms. In *Warhol,* the Court took a pragmatic approach to the first factor analysis, comparing how the works were monetized and in which markets, and returning the focus of the "purpose and character" test to the purpose and character of the secondary *use*, rather than of the secondary *work*. "The use of an original work to achieve a purpose that is the same as, or highly similar to, that of the original work is more likely to substitute for, or 'supplan[t],' the work . . . ." *Warhol*, 143 S. Ct. at 1274. If the secondary use "is of a commercial nature," that "tends to weigh against a finding of fair use." *Id* at 1279-80, quoting *Harper & Row Publishers, Inc. v. Nation Enterps.*, 471 U.S. 539, 562 (1985). If an original work and a secondary use share the same or highly similar purposes, and the secondary use is commercial, the first factor is likely to weigh against fair use. *Id.* at 1276-80. The first factor "considers the reasons for, and nature of, the copier's use of the original work." *Id.* at 1274. "In that way, the first factor relates to the problem of substitution—copyright's bête noire. The use of an original work to achieve a purpose that is the same as, or highly similar to, that of the original work is more likely to substitute for, or supplant, the work." *Id.*, quoting *Harper & Row,* 471 U.S. at 562 (cleaned up).

The Court held that the transformative use doctrine—a judicial gloss on the first factor—cannot swallow the copyright owner's exclusive right to prepare derivative works. *Id.* at 1282. "To

[46] OpenAI USPTO Comment, at 2 (emphasis supplied).

[47] *See, e.g.*, *Los Angeles News Service v. Tullo*, 973 F.2d 791, 797 (9th Cir. 1992); *Princeton University Press v. Michigan Document Services, Inc.*, 99 F.3d 1381, 1389 (6th Cir. 1996) (*en banc*) ("[t]he courts have . . . properly rejected attempts by for-profit users to stand in the shoes of their customers making nonprofit or noncommercial uses"), quoting W. Patry, Fair Use in Copyright Law, at 420 n.34; H.R. Rep. No. 1476, 94th Cong., 2d Sess. at 74 (1976) ("[I]t would not be possible for a non-profit institution, by means of contractual arrangements with a commercial copying enterprise, to authorize the enterprise to carry out copying and distribution functions that would be exempt if conducted by the non-profit institution itself.").

hold otherwise would potentially authorize a range of commercial copying of [such works], to be used for purposes that are substantially the same as those of the originals." *Id.* at 1285. Moreover, a "single-minded focus on the value of copying ignores the value of original works." *Id.* at 1286.

Here, the use of musical works to train AI models is done for the purpose of creating new musical works that serve purposes that are substantially the same as those of the originals.[48] If such a purpose could be considered "transformative" it would make the copying of any musical work to create a new musical work a "transformative" fair use, a notion the Supreme Court rejected. *Id.* at 1275 ("[A]n overbroad concept of transformative use, one that includes any further purpose, or any different character, would narrow the copyright owner's exclusive right to create derivative works. To preserve that right, the degree of transformation required to make 'transformative' use of an original must go beyond that required to qualify as a derivative."); *id.* at 1282 ("The first fair use factor would not weigh in favor of a commercial remix of Prince's 'Purple Rain' just because the remix added new expression or had a different aesthetic."); *id.* at 1286 (that AWF's "[c]opying might have been helpful to convey a new meaning or message" did not "distinguish AWF from a long list of would-be fair users," including "a musician who finds it helpful to sample another artist's song to make his own").

While much discussion has been directed to certain theoretical questions, none of these questions should distract from the fact that the purpose of using copyrighted works to train AI models is to create new works that compete with the works that were copied. AI models are trained using complete reproductions of existing works in order that the AI model can better simulate the expression contained in those works, and upon prompting by an end user may replicate or create derivatives of those works and use those works for any chosen purpose. AI models have been designed to be capable of generating expressive works that are comparable to and functionally equivalent to the ingested expressive works. Simply put, developers ingest copies of existing works into AI systems *so that* they can generate works that could fulfill a similar purpose or be put to similar use.[49]

It has been argued that the type of transformative purpose found in a case like *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015) may be analogous to the use made of copyrighted works in training AI. However, analogy to that case does not hold up to scrutiny. There, Google made digital copies of books solely to provide a search function. The Circuit court focused on the fact that Google took substantial measures to ensure that the public could not use its tool as a substitute for the books themselves. *Id.* at 222 ("Google has constructed the snippet feature in a manner that substantially protects against its serving as an effectively competing substitute for Plaintiffs' books."); *id.* at 226 ("The program does not allow access in any substantial way to a book's expressive content."). The purpose of the copying was "to make available significant information *about those books,*" so as to "identify[] books of interest to the searcher." *Id.* at 217-18 (emphasis in original). "Its purpose is not to communicate copyrighted expression, but rather, by revealing to the searcher a tiny segment surrounding the searched term, to give some minimal contextual information to help the searcher

[48] The purposes are the same regardless of whether the copying of expressive, copyrighted material occurs during training, "pre-training" or "fine-tuning." *See* NOI Question 8.1.

[49] *Compare Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 561 (S.D.N.Y. 2013) (computer program that scraped news articles on the web and provided excerpts of those stories to end users was not transformative nor a fair use).

learn whether the book's use of that term will be of interest to her." *Id.* at 227. No new books were created by Google and, in fact, users were directed by Google to the original books. In contrast, a fundamental use of AI models is to generate new, expressive works that can emulate and compete with the works that were copied and from which expression was taken.

Similarly, attempts to analogize to the Court's fair use finding in *Google, LLC v. Oracle America, Inc.*, 141 S. Ct. 1183 (2021) are not persuasive. That decision was also highly specific to the facts at issue in that case and does not support the argument that the unauthorized use of copyrighted works to train AI models is a "transformative" use that tilts the first factor to the secondary user. In *Google v. Oracle*, the nature of both the use (the copying) and the purpose of the use were highly distinguishable from the copying of musical or other expressive works, in their entirety, to train AI models to generate competing, substitutional works. There, as the *Warhol* Court recognized, Google "copied Sun's code, which was 'created for use in desktop and laptop computers,' 'only insofar as needed to include tasks that would be useful in smartphone[s].'" *Warhol*, 143 S. Ct. at 1277 n. 8 (quoting *Google v. Oracle*, 141 S. Ct. at 1203). The "use was justified in that context because 'shared interfaces are necessary for different programs to speak to each other' and because 'reimplementation of interfaces is necessary if programmers are to be able to use their acquired skills." *Id.* The *Google v. Oracle* Court favorably cited *Lexmark, Int'l, Inc. v. Static Control Components, Inc.*, 387 F.3d 522, 544 (6th Cir. 2004), which noted that "where a subsequent user copied a computer program to foster functionality, it was not exploiting the program's 'commercial value as *a copyrighted work*'." *Id.* at 1208 (emphasis in original).

Moreover, the original work at issue there was "declaring" computer code, which the Court found was, "if copyrightable at all, further than most computer programs (such as the implementing code) from the core of copyright." *Id.* at 1202; *see also id.* (such code was "inherently bound together with uncopyrightable ideas"); *id.* at 1198 ("Generically speaking, computer programs differ from books, films, and many other 'literary works' in that such programs almost always serve functional purposes.").

As the *Warhol* Court emphasized in limiting the holding of *Google v. Oracle*, the latter case "did not hold that any secondary use that is innovative, in some sense, or that a judge or Justice considers to be creative progress consistent with the constitutional objective of copyright, is thereby transformative. The Court instead emphasized that Google used Sun's code in a 'distinct and different' context, and 'only insofar as needed' or 'necessary' to achieve Google's new purpose." *Id.* at 1283 n. 18. Like AWF, the developer of an AI model bears the burden of justifying its unauthorized use of copyrighted works in their entirety "with some reason other than, 'I can make it better.'" *Id.* at 1285 n. 21.

Finally, and relatedly, we note that the Copyright Office assumes in Question 8.3 that the training of AI models "may be done for noncommercial or research purposes." But the training of AI models is fundamentally a commercial endeavor, especially in the case of generative AI. It is not "research" to copy works to build a "better" AI model so that it can create new works that are more likely to substitute for or supplant the works used to build the model. And any justification that copyrighted works are being copied in AI model training without a license for "noncommercial" or "research" purposes should be viewed skeptically. The reality of the market makes these labels too easy to abuse and nearly useless in the first factor analysis. For example, OpenAI began as a nonprofit research laboratory, and is now a multibillion-dollar for-profit company, powering Microsoft products.

Any commercial entity could front-end their AI systems endeavors with "research" arms or outsource such "research" to "noncommercial" entities, engage in massive unauthorized copying under the guise of fair use, and then shift entirely to commercial exploitation, leaving the creators of the copied works with no compensation. "The crux of the profit/nonprofit distinction is not whether the sole motive of the use is monetary gain but whether the user stands to profit from exploitation of the copyrighted material without paying the customary price." *Harper & Row,* 471 U.S. at 562. Furthermore, we caution against categorizing "noncommercial" uses as inherently fair use—to the contrary, while commercial use is one element of the first factor, "the mere fact that a use is educational and not for profit does not insulate it from a finding of infringement." *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 584 (1994).

C. <u>The second factor likely weighs against a finding of fair use where expressive works are used to train AI models.</u>

The second fair use factor considers the nature of the copyrighted work. Where AI models are trained on expressive works, such as musical compositions or sound recordings, this factor will almost always weigh against a finding of fair use. As noted above, when AI models are trained on expressive works, this is done precisely to reproduce and repurpose expressive content in those original works so that new, competing expressive works can be generated (including, in some cases, new works purportedly in "the style of" the original author or artist).

D. <u>The third factor likely weighs against a finding of fair use under current AI model training practices.</u>

The third fair use factor considers the amount and substantiality of the portion of the copyrighted work used in relation to that work as a whole. Current understanding is that AI model training processes ingest works in their entirety,[50] although there is no inherent need for that. The copying of entire creative works, rather than merely excerpts for uses such as comment or criticism, weighs against a finding of fair use.

In the *Authors Guild v. Google* case, copying the works in their entirety was necessary to achieve the purpose of the search function. *Id.* at 221 ("If Google copied less than the totality of the originals, its search function could not advise searchers reliably whether their searched term appears in a book (or how many times)"). Copying whole works was the difference between accurate search results and inaccurate ones. In contrast, AI model training has no such binary requirement, but rather operates on a simple "more is better" philosophy. OpenAI has explained that, "AI systems perform best when they are trained on larger amounts of data. Increasing the amount of training data available to the system increases the output system's accuracy and therefore utility."[51] Even if copying more portions of more works results in an AI model that is incrementally more commercially competitive, that is very different from the binary necessity for making complete copies in the *Authors Guild v. Google* case. On the contrary, OpenAI's comment reflects a desire that its model *take more*, so that it can *have more*—

[50] OpenAI USPTO Comment, at 6 ("Corpora used in training AI systems sometimes contain nearly all content of sampled works.").

[51] OpenAI USPTO Comment, at 7. The line chart provided alongside this quote shows an almost straight-line relationship whereby incremental additional training data provides incremental additional accuracy.

when it is plain that it could utilize licensed works with only *incremental* effects on its product. This approach is not consonant with fair use.

E. The fourth factor will almost always weigh against a finding of fair use.

While all factors must be weighed together, the Supreme Court has held that the "single most important element of fair use" is the fourth factor: "the effect of the use upon the potential market for or value of the copyrighted work." *Harper & Row*, 471 U.S. at 566; *see also Warhol*, 143 S. Ct. at 1290. In analyzing this factor, courts must consider whether "unrestricted and widespread conduct of the sort engaged in by [the secondary user] would result in a substantially adverse impact on the potential market" for the original work. *Campbell*, 510 U.S. at 590; *see also American Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 930 (2d Cir. 1994)("[A]n impact on potential licensing revenues for traditional, reasonable, or likely to be developed markets should be legally cognizable when evaluating a secondary use's 'effect upon the potential market for or value of the copyrighted work.'").

It is critical that a market to license copyrighted works for use in training AI models exists and is growing. The market is not merely a "potential" or theoretical market the existence or feasibility of which is open to debate; it is an actual market, with great potential for growth. Music companies are currently licensing works for use in training AI models. If generally allowed under a too-broad interpretation of fair use, the widespread unlicensed use of these works would eviscerate that market.[52]

Further, the ultimate result of a well-trained generative AI model is its ability to draw upon content included in its training set to produce works that directly compete with the works copied in training. Thus, licensing control and revenue are especially critical in connection with training AI models, since AI can be used to generate works that compete in the marketplace with the copied works, thereby reducing revenue from existing licensing markets as well.[53]

F. It is not a fair use to collect and distribute copyrighted material for AI training.

A business of "collecting and distributing" copyrighted materials without authorization is simply piracy. Entities engaging in these acts are stealing intellectual property and using it for commercial purposes, in direct competition with the copyright owners, who are themselves entitled to exploit the market for licensing their works in order to train AI models, or at their option, exclude

[52] *See, e.g.*, *id.*; *see also Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 174, 181 (2d Cir. 2018) (finding no fair use despite concluding that the defendants' technology "serve[d] a transformative purpose," in part because the technology "usurped a function for which [the plaintiff was] entitled to demand compensation under a licensing agreement").

[53] Question 8.5 asks "how should the effect on the potential market for or value of a copyrighted work used to train an AI model be measured?" But it then shifts to questions concerning the outputs of the AI model. The outputs should not factor into an analysis of the copying for training, which is the thrust of all or most of the other questions in Section 8 of the Notice. The issue is whether the copying of expressive works for training an AI model has an effect on the potential market for or value of those works used in the training. As discussed above, it demonstrably does. Note too that it does not matter if the works copied for training are not ultimately copied in any particular final output, as intermediate copying is still copying and is infringing. *See Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1519 (9th Cir. 1992); *Walker v. Univ. Books, Inc.*, 602 F.2d 859, 864 (9th Cir. 1979).

their works from AI training datasets.[54] And, as discussed above, the law is clear that potential fair uses of a downstream customer of an infringer cannot shield the infringer from liability (*supra*, n. 47).

Opt-Out (Question 9)[55]

Summary: U.S. copyright law is an opt-in system. Implementing an opt-out system would require a change to the Copyright Act. Opt-out regimes fundamentally undermine copyright protections by shifting the burden to obtain a license away from users. Further, opt-out regimes are profoundly unworkable in practice; there is no version of an opt-out system that would provide adequate protection for copyright owners. Affirmative consent is required for all uses of copyrighted works; attempts to characterize noncommercial uses as categorically fair are misguided.

U.S. copyright law is an opt-in system. Copyright owners have the choice to license or not to license their works to third parties, and anyone who seeks to use a copyrighted work must affirmatively obtain a license before doing so. Implementing an opt-out system would require an amendment to the Copyright Act, which the NMPA would oppose. It would also represent a fundamental erosion in the intellectual property rights of creators and an improper shifting of the affirmative responsibility to license away from developers of AI systems. NMPA strongly opposes consideration of such a measure.

An opt-out system does not work—and it is not needed. As discussed further in our response to Questions 10, 11 and 13, the large-scale licensing of copyrighted works already happens in the marketplace today. There are a number of generative AI platforms that were trained on entirely licensed and/or public domain content. For example, Stability AI reportedly trained its audio model,

[54] The existence of the market for pirated works is of course further evidence of the existence of the market for licensing those same works so that copyright owners can control and receive proper compensation for use of their creative works.

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

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

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 issues?

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?

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?

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?

Stable Audio, on a licensed music library,[56] as did Meta for its text-to-audio model, MusicGen.[57] While we believe large-scale licensing is necessary and practicable, NMPA respectfully cautions against taking as a "given" (per Question 9.3) the need for a large volume of works for AI training. While it is certainly the case that some AI models are trained on large amounts of material, it is not true of all AI models and should not be presumed to be needed or justified in all situations. In fact, research suggests that for some models, better performance can be achieved with smaller, carefully curated training datasets as compared to larger datasets.[58]

The Office has already observed how regimes that place a disproportionate burden on creators as compared to service providers ultimately serve to disadvantage creators in its study of Section 512 of the DMCA.[59] Already, rightsholders bear an enormous burden when it comes to enforcement against online piracy and implementing an opt-out system would only exacerbate those problems.

There is no version of an opt-out system that would provide adequate protections for copyright owners. An opt-out scheme that requires rightsholders to opt out on an AI company-by-AI company or application-by-application basis would not be feasible given the sheer volume of AI companies and applications; it is nearly a full-time job to keep up with developments in the AI marketplace (as any reader of this comment is likely acutely aware), let alone to opt out from all of them. Copyright owners, particularly individual creators and small businesses could not possibly meet such a burden.[60] Nor would an opt-out system that requires rightsholders to opt out particular websites from AI training (such as by using the Robots Exclusion Protocol, for example) work in practice for the simple reason that rightsholders do not control most of the websites where their works appear. Music publishers and songwriters license their works for use by online streaming services (in some cases on a compulsory basis), lyric websites and social media platforms; they do not control the code of any of those websites. Additionally, given the proliferation of online piracy, copyright owners are rarely aware of *all* of the locations of copies of their works online and so could not possibly protect against the scraping of all of those copies.

[56] *Announcing Stable Audio, a product for music & sound generation*, Stability.ai (Sept. 13, 2023), available at https://stability.ai/blog/stable-audio-using-ai-to-generate-music.

[57] *Introducing AudioCraft: A Generative AI Tool for Audio and Music*, Meta (Aug. 2, 2023), available at https://about.fb.com/news/2023/08/audiocraft-generative-ai-for-music-and-audio/ ("MusicGen . . . was trained with Meta-owned and specifically licensed music.").

[58] Xinyang Geng et al., *Koala: A Dialogue Model for Academic Research*, Berkley Artificial Intelligence Research (April 3, 2023), available at https://bair.berkeley.edu/blog/2023/04/03/koala/ ("[M]odels that are small enough to be run locally can capture much of the performance of their larger cousins if trained on carefully sourced data. This might imply, for example, that the community should put more effort into curating high-quality datasets, as this might do more to enable safer, more factual, and more capable models than simply increasing the size of existing systems.").

[59] U.S. Copyright Office, *Section 512 of Title 17* (2020), at 33. ("[T]he volume of notices demonstrates that the notice-and-takedown system does not effectively remove infringing content from the internet; it is, at best, a game of whack-a-mole.").

[60] Ginsburg, *supra* note 27, at 768 ("[L]arge and/or sophisticated copyright owners may understand the need systematically to opt out of exceptions and might have the means to undertake the necessary declarations. Smaller copyright owners and individual authors may not understand the opt-out regime (nor, depending on how it was implemented, be in a position to assume its burdens). The opt-out therefore would perpetuate, and aggravate, the disparate impact that formalities systems already wreak on individual creators.").

With regard to Question 9.1 ("Should consent of the copyright owner be required for all uses of copyrighted works to train AI models or only commercial uses?"), we emphasize that consent of the copyright owner is *already* required for all uses of copyrighted works that implicate the exclusive rights granted under Section 106 of the Copyright Act. Commerciality is not the appropriate (or legal) standard when considering the circumstances under which consent is or should be required. The Copyright Office should take care not to falsely equivocate "noncommercial" uses with fair use under the Copyright Act. While Section 107 does consider the commercial or noncommercial nature of a use, it is merely one part of one factor of the fair use analysis; it is not even dispositive of the first factor, and it is certainly not dispositive of the entire fair use analysis. [61]

With regard to Question 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?"), the copyright owner of a work has the exclusive right to object to the use of their work. The ownership of the work is determined by the Copyright Act as well as any relevant terms and conditions of contractual agreements between the parties in situations where the author is not the copyright owner.

Licensing in Connection with Training (Questions 10, 11 and 13[62])

Summary: Licensing musical works *before* training use is both *required* and *practicable*. Many digital platforms have successfully licensed musical works on a large scale in the free market in recent years. Digital platforms represent a substantial sector for the music industry, and there are well-developed processes in place for technology ventures to obtain free market licenses on a large scale.

Authorization from copyright owners is required *before* training AI models on copyrighted works. Training implicates the exclusive rights of copyright owners, and training without

[61] *Campbell,* 510 U.S. at 584 ("the mere fact that a use is educational and not for profit does not insulate it from a finding of infringement"); *Warhol*, 143 S. Ct. at 1276 ("[T]he fact that a use is commercial as opposed to nonprofit is an additional 'element of the first factor.' The commercial nature of the use is not dispositive.") (citation omitted).

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

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

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

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

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

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

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)?

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

authorization is infringement. In addition to the undisputed implication of the reproduction right,[63] a generative AI system may also implicate the derivative works right, public performance right, distribution right and public display right.

AI system developers seeking to train on musical works can license directly from copyright owners or their agents, and this licensing process is already underway.[64] Voluntary licensing to AI system developers is just the latest example of the music industry providing technology ventures large-scale access to musical works and sound recordings through voluntary licensing. Numerous audio-only and audiovisual streaming platforms, including YouTube, TikTok, Facebook, Instagram, Peloton, Snapchat, Twitch, Triller, and many others, have been able to obtain licenses covering the vast majority of the musical works and sound recording markets directly from music publishers and record labels, respectively, often using straightforward model agreements.[65] There is absolutely no reason why AI developers cannot do the same.

The current novelty of generative AI systems does not excuse a lack of content licenses, nor do claims that large volumes of works are needed by AI model developers.[66] Every new technology

[63] Open AI has explained that "Modern AI systems require large amounts of data. For certain tasks, that data is derived from existing publicly accessible 'corpora' (singular: 'corpus') of **data that include copyrighted works**. By analyzing large corpora (**which necessarily involves first making copies of the data** to be analyzed), AI systems can learn patterns inherent in human-generated data and then **use those patterns to synthesize similar data** . . . ." OpenAI USPTO Comment, at 2 (emphasis supplied).

[64] Hibaq Farah, *Google and Universal Music working on licensing voices for AI-generated songs*, The Guardian (Aug. 9, 2023), available at https://www.theguardian.com/technology/2023/aug/09/google-and-universal-music-working-on-licensing-voices-for-ai-generated-songs ("Google and Universal Music are negotiating a deal on how to license the voices and melodies of artists for artificial intelligence-generated songs.").

[65] *See, e.g.*, *NMPA and TikTok Announce Global Multi-Year Partnership Agreement*, NMPA (July 23, 2020), available at https://www.nmpa.org/nmpa-and-tiktok-announce-global-multi-year-partnership-agreement (describing a partnership with TikTok giving "NMPA members the ability to opt-in to a licensing framework that allows them to benefit from their works included on TikTok"); Tatiana Cirisano, *Facebook Strikes Deals With Major Labels to License Music On Its Gaming App*, Billboard (Sept. 14, 2020), available at https://www.billboard.com/pro/facebook-deals-major-labels-license-music-gaming-app/ ("Facebook has entered a series of new music licensing deals with labels and publishers for its Facebook Gaming platform, letting livestreamers who play video games for the platform's community of 200 million monthly viewers legally add songs from a vast catalogue of popular music to their videos."); Glenn Peoples, *Pandora and Sony/ATV Reach Multi-Year Agreement*, Billboard (Nov. 5, 2015), available at https://www.billboard.com/music/music-news/pandora-sony-atv-multi-year-licensing-agreement-6753769/; *Warner/Chappell Music and Pandora Sign Licensing Agreement*, Warner Music Group (Dec. 15, 2015), available at https://www.wmg.com/news/warnerchappell-music-and-pandora-sign-licensing-agreement-21071; U.S. Copyright Office, *Copyright and the Music Marketplace* (Feb. 2015) ("CMM"), at 56 ("The licensing of music for inclusion in audiovisual works . . . occurs in the free market for both musical works and sound recordings.").

[66] Such excuses have been expressly advanced by AI model developers. *See, e.g.*, Rob Salkowitz, *Midjourney Founder David Holz On The Impact Of AI On Art, Imagination And The Creative Economy*, Forbes (Sept. 16, 2022), available at https://www.forbes.com/sites/robsalkowitz/2022/09/16/midjourney-founder-david-holz-on-the-impact-of-ai-on-art-imagination-and-the-creative-economy/?sh=5e9027f2d2b8 ("[Q.] Did you seek consent from living artists or work still under copyright?" "[A.] No. There isn't really a way to get a hundred million images and know where they're coming from."). Such self-serving excuses ignore the reality that many successful digital platforms license works. *See, e.g.*, Dawn Chmielewski & Stephen Nellis, *Adobe, Nvidia AI Imagery Systems Aim to Resolve Copyright Questions*, Reuters (Mar. 21, 2023), available at https://www.reuters.com/technology/adobe-nvidia-ai-imagery-systems-aim-resolve-copyright-questions-2023-03-21/ ("Nvidia trained the technology on images licensed from Getty Images, Shutterstock Inc (SSTK.N), and Adobe, and plans to pay royalties."); Matt Growcoot, *New AI Image Generator Startup Takes a 'Responsible' Approach to Copyright*, PetaPixel (Sept. 8, 2023), available at https://petapixel.com/2023/09/08/new-ai-image-generator-startup-takes-a-responsible-approach-to-copyright/ ("Bria AI says that not only has it licensed visual content from large image libraries,

is novel at the start, and most companies believe that they require massive scale in order to succeed. Happily for technology platforms, the music industry has a fully-developed market for licensing works in bulk.

Given the existence of a fully functional free market that is being successfully used by many technology platforms to license across the full market for musical works, there is no basis for a government-imposed collective licensing scheme. Compulsory licensing is an extreme remedy that deprives copyright owners of their right to contract freely in the market, and takes away their ability to choose whom they do business with, how their works are used, and how much they are paid. Songwriters and music publishers are already subject to the Section 115 compulsory license, despite the lack of evidence that it is necessary. As the Office has noted, "Viewed in the abstract, it is almost hard to believe that the U.S. government sets prices for music. In today's world, there is virtually no equivalent for this type of federal intervention—at least outside of the copyright arena . . . . [C]ompulsory licensing removes choice and control from copyright owners who seek to protect and maximize the value of their assets."[67]

While specific terms will vary between licensees, the process of obtaining voluntary licenses to musical works would not be fundamentally different for AI model developers than it is for the many other digital platforms that license music in the free market. Indeed, many of the major companies in the AI model development space, including Alphabet, Amazon, Apple and Meta, have significant experience negotiating voluntary licenses for music on an industry-wide basis for their other digital services.

but has also collaborated with 'boutique stock photo agencies around the globe, individual artists, and photographers.'"). For any platform, the choice to pirate creative works instead of license them is just that: a choice, motivated not by noble intention, but by the selfish desire to make more money by taking from others without consent.

[67] CMM, at 145, 148.

Transparency (Questions 15-17[68])

Summary: Transparency and recordkeeping requirements are needed to disincentivize infringing activity and to support enforcement activities by copyright owners. As their purpose is to enable effective enforcement, such requirements must not be treated as establishing a defense to or a safe harbor for infringement. The primary parties who should be required to maintain records under a transparency and recordkeeping scheme are developers of AI models, those who use existing AI models to develop new AI tools and those who broker datasets for use in AI training. As the market for AI develops, there may be additional players who should also be subject to transparency and recordkeeping requirements, in anticipation of which any laws and regulations should be drafted in a manner allowing for adaptation to a shifting marketplace.

A. Recordkeeping Obligations: AI Model Developers

Experience to date indicates that unless the law compels AI model developers to disclose the contents of training datasets, they often will not do so.[69] It is of great concern to our members that AI model developers may eventually rely on their own lack of recordkeeping and transparency to

[68] *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?*

15.1. What level of specificity should be required? 10.1. Is direct voluntary licensing feasible in some or all creative sectors?

15.2. To whom should disclosures be made?

15.3. What obligations, if any, should be placed on developers of AI systems that incorporate models from third parties? Should creators of training datasets have a similar obligation?

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

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

17. Outside of copyright law, are there existing U.S. laws that could require developers of AI models or systems to retain or disclose records about the materials they used for training?

[69] Katherine Miller, *Introducing the Foundation Model Transparency Index*, Stanford University (Oct. 18, 2023), available at https://hai.stanford.edu/news/introducing-foundation-model-transparency-index ("Most companies also do not disclose the extent to which copyrighted material is used as training data."); Alistair Barr, *Llama copyright drama: Meta stops disclosing what data it uses to train the company's giant AI models*, Business Insider (Jul. 18, 2023), available at https://www.businessinsider.com/meta-llama-2-data-train-ai-models-2023-7 ("A major battle is brewing over generative AI and copyright. Publishers want to be paid if their work has been used to train large language models. Big tech companies would rather not pay. One way to avoid the issue is to just not tell anyone what data you used to train your AI model. Meta seems to be trying that tactic."); James Vincent, *OpenAI co-founder on company's past approach to openly sharing research: 'We were wrong,'* The Verge (Mar. 15, 2023), available at https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview ("OpenAI has shared plenty of benchmark and test results for GPT-4, as well as some intriguing demos, but has offered essentially no information on the data used to train the system . . . . [One] reason suggested by some for OpenAI to hide details of GPT-4's construction is legal liability. AI language models are trained on huge text datasets, with many (including earlier GPT systems) scraping information from the web — a source that likely includes material protected by copyright.").

deprive copyright owners of their just remedies for copyright infringement and illegal use of proprietary works.[70]

To ensure that rightsholders are not deprived of their remedies and AI model developers cannot evade accountability by destroying the evidence, developers of AI models must be required to collect and retain complete and detailed records of the contents of datasets collected and/or used to train AI models. The retention requirement should be for a period of at least 10 years given current interpretations of the law concerning statute of limitations accrual.[71]

Moreover, compliance with recordkeeping and/or transparency requirements must not become a defense to or a safe harbor against claims of infringement. These measures should not be a substitute for complying with the law and obtaining licenses to use proprietary works—rather, these measures should be a companion to such obligations, by preventing destruction of the evidence that could vindicate the claims of rightsholders.

And great care must be taken to avoid leaving any loopholes in these requirements, which could rapidly be exploited by sophisticated developers. The requirements must apply to datasets used by any type of entity, in any way, at any stage of training (including initial training and retraining), as well as for refining and testing. Thus, although these comments, to aid the reader's understanding, discuss and include distinctions between terms such as "models," "platforms," "developers" and "aggregators," the requirements being proposed should not be limited to such technology and actors but must instead be applied broadly.

The detailed records of datasets maintained by AI developers must be sufficient to allow rightsholders to identify the works used in a training set and the source from which such works were acquired. At a minimum, such records should include the following information:

(1) Information sufficient to identify each specific work used in training, retraining, refining or testing the AI model, or any similar use;

(2) The portion of each work used;

(3) Available metadata associated with each work (*e.g.*, title, author, owner(s), performers, etc.);

(4) The purposes for which each work has been used;

(5) Any copyright management information associated with each work;

[70] *See* Defendant Midjourney, Inc.'s Notice of Motion and Motion to Dismiss Plaintiffs' Complaint and to Strike Class Claims, *Andersen et al. v. Stability AI Ltd. et al.*, 3:23-cv-00201 (N.D. Cal.) (ECF No. 52), at 1, 9 (motion to dismiss arguing "the Complaint does not identify a single work by any Plaintiff that Midjourney supposedly used as training data" and "nowhere does the Complaint plead what was in this 'subset' of training material, or that it included even one work by *Plaintiffs*").

[71] *See, e.g., Petrella v. Metro-Goldwyn-Mayer, Inc.*, 572 U.S. 663, 671 (2014) (explaining that under the separate-accrual rule, "when a defendant commits successive violations, the statute of limitations runs separately from each violation. Each time an infringing work is reproduced or distributed, the infringer commits a new wrong. Each wrong gives rise to a discrete 'claim' that 'accrue[s]' at the time the wrong occurs. In short, each infringing act starts a new limitations period.").

(6) Each work for which no license has been identified, including the reason the work has not been licensed (such as whether the AI developer relied on a determination that such work is in the public domain);

(7) Each work for which a license has been obtained, including the identity of the licensor;

(8) Information sufficient to identify the immediate source from which each work was obtained, as well as its full provenance (*i.e.*, the chain of individuals or entities who ultimately provided such work to the immediate source), to the extent available;

(9) What security measures have been implemented to ensure the works will not be leaked to or accessed by third parties, and when they were implemented; and

(10) Any third parties to whom the works have been disseminated or who have otherwise gained access to the works.

B. Recordkeeping Obligations: AI Developers Using Existing AI Models

Developers of AI platforms incorporating other developers' AI models should be required to maintain all records provided by the AI model developer reflecting any of the categories of information set forth in response to Questions 15 and 15.1 above. For clarity, the existence of such records in the AI model developer's possession or the public availability of such records should not discharge AI platform developers of this duty. The AI model developer should also be required to provide the AI platform developer sufficient information to understand each dataset aggregator from which works were acquired and the number of works acquired and used in training; the number of works used through direct licenses with rightsholders; and the works for which the AI model developer determined a license was not required, whether because they are in the public domain or otherwise.

Further, AI platform developers should maintain the same records identified in subsection (A) above, in connection with (1) any further use of the works used to train the AI models for additional training, retraining, refining or testing the AI platform, and (2) any additional works used in connection with the AI platform.

C. Recordkeeping Obligations: Dataset Aggregators

Dataset aggregators should be required to maintain the same records identified in subsection (A) above.[72]

D. Disclosure Requirements

The law must provide copyright owners with reasonable and effective means to obtain the records called for in subsections (A) through (C) above. The availability of such records cannot be merely through recourse to the courts. Court discovery procedures are insufficient; they generally

[72] Certain of the information categories set forth in response to Questions 15 and 15.1 may be inapplicable as to some dataset aggregators; for example, a dataset aggregator who does not engage in developing AI may lack records reflecting specific uses of works in connection with training AI.

require the initiation of a proceeding, may vary by jurisdiction, and are often prohibitively expensive, especially for small companies and individual creators.

Anticipated costs to developers of AI systems would include the costs of employing compliance personnel or third-party service providers to ensure collection and maintenance of the required records and compliance with disclosure requirements. Such costs should be minor compared with the investments required to build AI models. On the other hand, recordkeeping would encourage direct licensing, resulting in fewer disputes, less infringement and greater compensation to owners of copyrighted works. It is essential to keep in mind the impact of *not* requiring recordkeeping: hobbling rightsholders' ability to enforce their rights against infringement by AI companies.

Because copyrighted works should never be used to train AI models without specific authorization from the owners of such works, separate notice is not necessary. Licensors may provide for certain notice requirements in the course of their negotiations, for example by requiring dataset aggregators to notify them of each AI developer to whom their works were transferred or requiring AI developers to notify them at the time their works have been used for training. However, a general rule requiring notification could potentially call into question the requirement that AI developers and dataset aggregators obtain permission from copyright owners before using their works. Recordkeeping and transparency requirements are more appropriate in this context.

E. Existing Laws

The NMPA is not aware of an existing, comprehensive legal regime requiring the type of recordkeeping being proposed herein in connection with musical works. To be sure, in many cases developers of AI models and systems must maintain some of the information that is part of the NMPA's proposal overall. For example, to the extent developers maintain copies of the works used to train AI, they may not remove copyright management information from those copies,[73] and developers who are facing or involved with litigation concerning their training practice have an obligation to preserve relevant records. But these existing obligations are insufficient to meet the moment. The law should affirmatively require all developers to maintain in the ordinary course all types of information that copyright owners may require to enforce their rights.[74]

[73] 17 U.S.C. § 1202(b).

[74] Detailed business recordkeeping requirements have been established in other industries, such as in healthcare and finance. *See, e.g.*, 45 C.F.R. § 164.316 (requiring covered healthcare entities to maintain documentation of security measures protecting electronic protected health information); 10 NYCRR § 405.10 (requiring detailed recordkeeping for hospitals in New York State, to be maintained for at least six years); 31 C.F.R. §§ 1010.401, 1010.410 (requiring detailed recordkeeping for financial institutions, because such records will "have a high degree of usefulness in criminal, tax, or regulatory investigations or proceedings"). The AI industry is, almost by definition, owned and operated by companies that are highly sophisticated in terms of information processing and storage, who are more capable even than the healthcare and finance industries to satisfy the types of recordkeeping requirements being called for here, at relatively minimal inconvenience to their business operations.

Copyrightability (Questions 18-21[75])

Summary: Human authorship is required for a work to be copyrightable. Copyright policies should continue to incentivize and protect human creativity by granting protection only to original human authorship. Whether protection is merited in any given case is a fact-intensive question that will likely need to be addressed by the courts. Copyright Office registration policies should not disincentivize creators with a legitimate claim to a work from registering due to vague or burdensome disclaimer requirements and should treat disclaimer of AI-generated material the same as disclaimer of public domain material.

A. Copyrightability

Copyright protects, and should continue to only protect, original human authorship. It is not necessary to revise the Copyright Act to clarify or modify this requirement nor to provide additional standards to determine when the requisite level of human authorship for copyrightability has been met. Courts can, and in fact have already begun to, address questions of authorship related to works that include AI-generated material by applying existing legal standards regarding authorship and originality.[76]

As a policy matter, copyright law should never protect purely AI-generated content that does not represent human expression. Existing copyright law rightfully incentivizes human creativity by granting protection to the "the fruits of intellectual labor" that "are founded in the creative powers of the mind."[77] The importance of distinguishing between purely AI-generated content and original authorship of humans for the purposes of eligibility for copyright protection cannot be overstated. Policies surrounding copyrightability of AI-generated content will either serve to incentivize continued investment and effort into human creativity or to disincentivize it. The law should never be such that human creators stand to gain more from repeatedly clicking a button to generate massive amounts of AI-produced materials than from putting their hearts, souls, experiences, skills, talents, and emotions into expressive works of art.

[75] *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?*

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?

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?

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?

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?

[76] *Thaler v. Perlmutter*, 1:22-CV-01564 (D.D.C. Aug. 18, 2023) (ECF NO.18).

[77] *Trade-Mark Cases*, 100 U.S. 82, 94 (1879).

A fundamental purpose of copyright law is to motivate humans to develop artistic skills and to engage in the demanding and rewarding work of creating art, and not to entirely hand off the work of creation to a machine. Copyright protection is a reward for real creativity. This is not to say that creators should be discouraged from using new methods or new technologies to express themselves—to the contrary, creators have always found ways to use new technologies to assist them in creating art—but copyright only properly applies to human authorship, in whatever form that authorship may take. Where creators use AI technology as a tool in the creative process to make works that represent their original authorship, their works should be entitled to protection under copyright law. Copyright law also rightfully protects sufficiently creative human selection and arrangement of otherwise unprotectable elements,[78] including those that may be generated by AI, as the Office has made clear. Creators that use AI to refine, recast, or modify, or to create new derivative works based on their preexisting works may also have legitimate claims of authorship over the resulting work in some circumstances.

The ways in which a human may engage with AI vary widely and will certainly continue to evolve as the technology develops. Analyzing the protectability of works that include AI-generated material may entail highly fact-intensive and case-by-case questions about how the user has interacted with the program and with its output, and what the role or influence was of preexisting materials. Courts have overwhelmingly been the primary venue where matters of dispute on copyrightability have been resolved, and are best positioned to continue to be so, given the broad diversity of implementations of AI that currently exist, and the countless others that will almost certainly exist in the future.

Content that is infringing should never be copyrightable. In the case of generative AI that has been trained on preexisting third-party works, to the extent that a program's output includes elements of such preexisting works or constitutes unauthorized derivatives thereof, those elements can never be owned by the program's developer or user, and their creation and reproduction may constitute copyright infringement.[79] Where an output includes a reproduction of a preexisting work under copyright, the human author of the preexisting work may be considered an owner of the portions of the output that include her original work.

B. Registration-Related Considerations

[78] *See, e.g., Skidmore v. Led Zeppelin,* 952 F.2d 1501, 1074-75 (9th Cir. 2020).

[79] 17 U.S.C. §§ 103(a) ("…[P]rotection for a work employing preexisting material in which copyright subsists does not extend to any part of the work in which such material has been used unlawfully."), 103(b) ("The copyright in a compilation or derivative work extends only to the material contributed by the author of such work, as distinguished from the preexisting material employed in the work, and does not imply any exclusive right in the preexisting material."); *Circular 14, Copyright in Derivative Works and Compilations*, United States Copyright Office, at 2 ("Protection does not extend to any preexisting material, that is, previously published or previously registered works or works in the public domain or owned by a third party. … In any case where a copyrighted work is used without the permission of the copyright owner, copyright protection will not extend to any part of the work in which such material has been used unlawfully."); Compendium of U.S. Copyright Office Practices (Third Ed.) § 313.6(B) ("[T]he Office may refuse registration if the [unauthorized] preexisting material is inseparably intertwined with the compilation or the derivative work, such as … an unauthorized arrangement of a song.").

Where human creators have a legitimate claim to copyright in a work, they should not be disincentivized from applying for registration, and copyright registration rules and guidance must be clear and easy to follow.

Some applicants may face confusion under the current guidelines that require explicit disclaimer of "AI-generated content that is more than de minimis."[80] The concept of artificial intelligence is dynamic and difficult to define and apply. There is no industry-accepted definition of AI or generative AI, and the current understandings of these terms may change over time, as may the technology itself. Particularly as the copyright registration process is meant to be accessible to unrepresented applicants, registration rules that penalize applicants for not knowing the technical operations of the creative tools they use or misunderstanding this evolving, nuanced field should be avoided. Additionally, an overbroad or hyper technical disclaimer requirement would imperil existing and future registrations and complicate legitimate enforcement efforts. We anticipate that lawsuits to enforce copyrights will become more complex, unpredictable, and expensive under the Office's current policy, due to AI-based Section 411(b) challenges.[81]

The Office should apply the same rules to applications for works that include some component of purely AI-generated material as they do applications for works that include some amount of public domain material. When registering a work that contains an "appreciable" amount of public domain material, the applicant must identify it in the "Material Excluded" field of her application.[82] On the other hand, the Office's articulated rule regarding AI-generated content is that "AI-generated content that is more than de minimis should be explicitly excluded from the application."[83] The differing language here raises questions as to "appreciable" and "more than de minimis" have the same or different meanings. Accordingly, we suggest that the Office clarify the applicable standard and use the same terminology in both situations.

[80] 88 Fed. Reg. 16190, 16193 (Mar. 16, 2023).

[81] As the Office has acknowledged, non-compliance with current disclaimer policies may lead to the courts "disregard[ing] a registration in an infringement action." *See* 88 Fed. Reg. 16190, 16193-94 (Mar. 16, 2023).

[82] Copyright Office Circular 50: Copyright Registration for Musical Compositions, page 3. https://www.copyright.gov/circs/circ50.pdf

[83] 88 Fed. Reg. 16190, 16193 (Mar. 16, 2023).

Copyright Infringement and Liability (Questions 22-25[84])

Summary: Courts have used and can continue to use the language of the Copyright Act and legal precedent, including (with respect to the output of AI programs) the substantial similarity test, to make determinations as to infringement and liability. No "AI-specific" rules or exceptions are necessary. However, it may be difficult in the AI context for copyright owners to detect infringement and to prove access and copying, and thus the detailed recordkeeping discussed herein in response to Questions 15–17 is required.

Where output from AI programs results from training on pre-existing copyrighted works (in whole or in part) and is substantially similar to those works, the copyright owner's exclusive rights of reproduction, adaptation, publication, performance or digital transmission, display, and preparation of derivative works may be implicated.

The substantial similarity test is adequate to address claims of infringement based on the creation or use of AI-generated output. To prove copyright infringement, a plaintiff must show that the defendant had access to plaintiff's work, and that the works are substantially similar. The challenge of addressing infringing output generated using AI lies less in the substantial similarity test than in the difficulty plaintiffs may face in proving access in some cases.

As discussed in the NMPA's responses to Questions 15–17 herein, detailed recordkeeping and appropriate transparency will assist plaintiffs in identifying and proving their works were used in the training of AI models.[85] Where a work generated using AI is substantially similar to an existing copyrighted work, the AI developer's and/or dataset aggregator's failure to maintain and disclose complete records as required should result in a presumption that the AI model and/or program was trained on the copyrighted work.

Further, where a developer of an AI model or program allows members of the public to train the model using potentially copyrighted works, the developer should adopt a method for collecting

[84] *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?*

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?

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

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?

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

Are existing civil discovery rules sufficient to address this situation?

[85] For clarity, these requirements are necessary *both* (1) to identify infringement by way of unauthorized copying in the process of training, as discussed above, and (2) to prove copying as an element of infringement, regardless of whether the initial training itself was authorized, where output is substantially similar to a work on which the AI was trained, as discussed in this section.

and maintaining records of the works used by those members in this additional training and ensure the public only train the model on works that have been authorized for such a purpose.

As with all cases alleging copyright infringement, the answer to questions about infringement involving AI programs is highly fact-dependent and must be determined on a case-by-case basis. Existing doctrines of direct and secondary liability can be applied to determine liability as to a particular party involved in the development and use of generative AI models and programs. Each of the parties identified in Question 25 may be liable together or on their own, for direct, vicarious and/or contributory infringement. There is no categorical answer as to the type of party that should be found liable, or for what type of infringement, without factual context.

Copyright Management Information, Labeling and Identification (Questions 26, 28-29[86])

Summary: Further to the need for greater transparency and to disincentivize infringement, all relevant CMI should be required to be maintained. With regard to labeling, output generated using AI should identify, in its metadata, the AI system used to generate it, the developer of that system, the models used by the system and the datasets the model was trained on. However, neither labeling nor identification tools sufficiently disincentivize or resolve infringing uses; the other proposed protections discussed in these responses are also necessary.

AI developers must obtain a license in order to use copyrighted works for training purposes, and as proposed herein, recordkeeping requirements should be imposed. Dataset aggregators must maintain all copyright management information ("CMI") for works they collect or transfer; and AI developers must maintain all CMI for protected works that are used in connection with their models and platforms. As with other recordkeeping requirements proposed herein, maintaining CMI will disincentivize infringement, as it can assist rightsholders in discovering and proving both access and infringement. Section 1202(b) prohibits intentional removal or alteration of CMI for the same reason.

To ensure that CMI is maintained throughout the transfer and training process, each dataset aggregator transferring works to AI developers, and each AI model developer making its model available to AI platform developers, should be required to represent that it is not aware of any instance of CMI having been intentionally removed from the works intended for or used in training AI, other

[86] *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?*

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?

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

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

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

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?

than with the authority of the copyright owner, and that it has employed reasonable measures to ensure that such has not occurred.

Where a developer fails to maintain CMI for material in its training data, or fails to obtain a representation from the creator of the dataset or developer of the AI model that no CMI has been altered or removed, and where the copyright owner can show that CMI was altered or removed, courts should apply a presumption that the developer of the AI system knew or had reasonable grounds to know that CMI was removed or altered without authorization and would induce, enable, facilitate, or conceal an infringement of copyright under 1202(b).

With regard to labeling, output generated using AI should identify, in its metadata, the AI system used to generate it, the developer of that system, the models used by the system and the datasets the model was trained on. This requirement should apply to all digital output generated using an AI system. The developer of the AI system that generates the output should be responsible for providing the requisite identification.[87]

The NMPA reserves comment on the appropriateness of consequences for failing to label or for removing a label, such as fines or suspension of licenses, until the nature and details of the legal obligations are known.[88]

With regard to detection, the NMPA understands that several tools have been developed that can detect songs generated entirely by AI and that represent that they perform with a high level of accuracy, including Matchtune (which reports a 100% match rate); Deezer's Radar; and Tunecore's Believe. However, the mere identifying of works generated by AI models and platforms does little to resolve concerns regarding the unauthorized use of copyrighted works to train AI programs for the purpose of generating similar and competing works. Identification may have many societal benefits, including informing consumer choice; and between labeling requirements and third-party identification tools, labeling is preferred. Neither labeling nor identification tools, however, sufficiently disincentivizes infringing uses; the other proposed protections discussed in these responses are also necessary.

Dated: October 30, 2023

Respectfully submitted,

NATIONAL MUSIC PUBLISHERS' ASSOCIATION

[87] Underscoring the technical feasibility of a labeling requirement, some generative AI products already in the market appear to have already voluntarily adopted metadata labeling for generated output. *See Content Credentials*, Adobe (Oct. 10, 2023), available at https://helpx.adobe.com/creative-cloud/help/content-credentials.html ("Content Credentials indicating the use of generative AI tools will be included with all content generated with Adobe Firefly to help promote transparency around the use of generative AI. In the future, Content Credentials from other Adobe apps will also support indicating that generative AI was used in the creative process.").

[88] The NMPA notes that the DEEPFAKES Accountability Act bill, H.R. 5586, 118th Congress (2023), proposes civil and, in some cases, criminal penalties for failure to meet specified labeling requirements in cases of "false personation," as well as a private right of action.

Danielle Aguirre
Shannon Sorensen
Eric Sunray

1900 N St. NW
Suite 500
Washington, D.C. 20036