Before the
COPYRIGHT OFFICE
LIBRARY OF CONGRESS
Washington, DC

In the Matter of	)	
	)	
	)	**Docket No. 2023-6**
Artificial Intelligence and Copyright	)	
	)	
	)	

COMMENTS OF THE AMERICAN SOCIETY OF COMPOSERS, AUTHORS AND PUBLISHERS ON ARTIFICIAL INTELLIGENCE AND COPYRIGHT

The American Society of Composers, Authors and Publishers ("ASCAP") hereby submits these comments pursuant to the Notice of Inquiry (the "Notice") issued by the Copyright Office on August 30, 2023, 88 Fed. Reg. 59,942 for written comments on copyright law and policy issues raised by artificial intelligence ("AI") systems. While each of the issues identified in the Notice would have substantial effect on ASCAP and its members and affiliates, it will only focus on those issues on which it believes it can best comment. ASCAP reserves the right to comment on those or any other issues in reply comments.

I. INTEREST OF ASCAP

ASCAP is one of the world's leading music performing rights organizations ("PRO"), currently representing over 947,000 songwriter, composer and publisher members and a repertoire of over fifteen million copyrighted musical works.

ASCAP was founded by music creators in 1914 and, pursuant to its Articles of Association, is governed by a twenty-four person Board of Directors comprised of twelve music creators (*i.e.*, songwriters, composers and lyricists) and twelve music publishers who are elected democratically by the ASCAP membership every two years.

ASCAP receives from all of its ASCAP members a global and non-exclusive grant of right to license public performance rights in the copyrighted works known as "musical compositions" on all platforms and all services, including interactive and non-interactive. It does not today license any rights other than those associated with public performances of musical compositions, such as mechanical rights and synchronization rights, and it does not license the copyright in the sound recording (a separate copyright from musical compositions). Unlike mechanical rights, public performance rights are not compulsory under U.S. law; however, ASCAP does operate under a Consent Decree with the U.S. Department of Justice.

ASCAP collectively licenses the non-dramatic public performance rights of its members' musical compositions on a non-exclusive basis to music users (*i.e.*, licensees) that publicly perform music in virtually every communications media, including internet-delivered audio-only and audiovisual video streaming (both interactive and non-interactive), terrestrial radio, broadcast television, cable and satellite, in-person performances at restaurants or concerts, as well as background music in stores, fitness centers and dance schools, and many, many more. ASCAP's primary business is offering music users "blanket licenses," meaning that in exchange for a fee, the licensee may use any amount of music in the ASCAP repertoire. The "blanket license" has long been known as the most efficient and cost-effective form of licensing at scale, and could be a reasonable form of licensing for AI.

ASCAP represents not only U.S. writers and publishers, but also hundreds of thousands of foreign writers and publishers through binding contracts that are often reciprocal in nature with approximately 100 foreign Collective Management Organizations ("CMOs") that cover nearly every country in the world. Through these agreements, ASCAP is permitted to license in the U.S. the public performing rights in hundreds of thousands of musical works controlled by non-U.S.

songwriters and composers. When the agreements are reciprocal, ASCAP likewise receives royalties from those foreign CMOs for performances of ASCAP musical works occurring overseas.

ASCAP is the only PRO in the United States that operates on a not-for-profit basis. In accordance with ASCAP's Articles of Association, all dollars collected less expenses are distributed to ASCAP members and affiliates based on performances of their copyrighted works by licensees—about 90 cents of every dollar collected goes back to ASCAP's members as royalties, meaning that ASCAP's current overhead is 10%. All of ASCAP's competitors in the U.S. operate on a for-profit basis and are owned by U.S. broadcasters, private equity and other investors.

In addition to licensing its members' work on their behalf, ASCAP enforces its members' rights when their copyrighted works are performed publicly without authorization. To that end, ASCAP brings copyright infringement lawsuits in U.S. District Courts throughout the country regularly on behalf of its members to enforce their rights to be paid.

ASCAP is thus well-positioned to comment on behalf of the nearly one million songwriter, composer, and music publisher members whose rights and livelihoods it protects. Because most ASCAP members are not sound recording artists and accordingly do not have opportunities to be paid for concert tours, merchandise, and related projects, the public performance royalties paid to these members are often their only source of income. In the event that AI platforms usurp this revenue stream, untold numbers of creators stand to lose their ability to make a living.

II. Overview of ASCAP's Comments

While generative AI has the potential to enhance human creative efforts, the unchecked use of this technology threatens to undermine the very purpose of the copyright laws by supplanting, rather than supporting, human creative work. Accordingly, the AI industry must be

held responsible under all applicable laws—including existing copyright laws and state and federal legal frameworks—to ensure that it does not unfairly and illegally exploit the work of human artists, writers, and other creators. In June 2023, the ASCAP Board of Directors unanimously approved and publicly announced that the following principles should govern the development and use of generative AI. These principles are consistent with ASCAP's collective licensing practices since its founding in 1914:

- **Humans First**: A vague appeal to "innovation" cannot justify infringing the rights of existing human creators. The copyright laws and other legal frameworks must ensure that human creativity is supported, rather than supplanted, by developing technologies, including generative AI.

- **Consent**: Developers of AI tools, creators of underlying AI models, and compilers of training datasets must be responsible for ensuring that all relevant rights have been obtained for the use of any copyrighted content included in models or training datasets.

- **Transparency**: Developers of AI tools, creators of underlying AI models, and compilers of training datasets must be responsible for collecting, maintaining and notifying rightsholders of accurate and comprehensive information regarding all copyrighted content the AI tool or dataset contains, including, where available, all standard industry metadata.

- **Compensation**: To the extent the AI industry exploits and benefits financially from the creativity and labor of human artists, writers, and other creators, it must compensate these creators fairly for the use of their works.

- **Credit**: Developers of AI tools must be responsible for labeling AI-generated musical works as such, and must provide credit to the creators whose works are utilized in such generation.

- **Global Consistency**: Because copyrights to musical works are globally administered through an interconnected structure of international agreements and licensing networks, global consistency in AI regulation is essential to enabling proper exchange of compensation and remuneration across borders. Inconsistencies in international or regional laws and regulations may not only cause operational disruptions, but create back-doors to evade legal compliance.

In accordance with these principles, voluntary collective licensing is the best way to harness the power of generative AI while preserving the livelihoods of creators. Collective

licensing has long been a staple in the music industry and has adapted to major technological developments such as an industry-wide shift to digital music consumption at the turn of the millennium. AI will be no exception.

As demonstrated by the hundreds of thousands of businesses that currently license public performance rights from ASCAP, voluntary collective licensing is practically feasible and mutually beneficial for both creators and businesses that derive value from the use of copyrighted musical works. Similarly, the AI industry can and should obtain consent and pay fair market value for copyrighted musical works before using them in the development of AI models. In fact, OpenAI reached a licensing deal with the Associated Press in July this year for the use of the latter's news archive in the development of generative AI models. Such licensing efforts need to be widely adopted across creative industries.

The main obstacle to voluntary collective licensing is the lack of willingness on the part of AI providers to come to the negotiation table with the creators. AI providers have been operating under the presumption, albeit a false one, that their use of unlicensed copyrighted works for training AI models is legal. The lack of transparency from AI providers concerning their use of copyrighted works in AI training also makes it extremely onerous for creators to enforce their rights and thus deprives them of meaningful bargaining power.

Therefore, the following measures are necessary to create the relevant incentives that can bring about meaningful voluntary licensing negotiations:

First, the Copyright Office should issue written guidance to be considered by Congress as it considers new legislation making it clear that certain AI-generated outputs may infringe the exclusive public performance rights of copyright owners under Section 106(4) of the Copyright Act.

Second, AI developers should be required to retain information on (1) all copyrighted material present in any datasets compiled by the developer or obtained from a third party; (2) all copyright material actually ingested as training input into any particular model; and (3) for each piece of utilized copyrighted material, the particular training use for which that material was employed, and whether it was used as training, testing, and/or or validation data.

Third, a federal right of publicity should be created.

Fourth, there should be a requirement that works generated by AI be clearly labeled as such, credit the creators whose works are utilized in the generation, and reference the larger database of copyrighted works on which the AI tool was trained.

On the issue of the copyrightability of AI-generated works, this should be determined on a case-by-case basis depending on the nature and degree of human involvement. When a human significantly edits, manipulates, or alters the output generated by an AI tool, the resulting material should be subject to copyright protections to the extent the work reflects human creativity.

III. COMMENTS ON SPECIFIC QUESTIONS

A. General Questions

1. As described above, generative AI systems have the ability to produce material that would be copyrightable if it were created by a human author. What are your views on the potential benefits and risks of this technology? How is the use of this technology currently affecting or likely to affect creators, copyright owners, technology developers, researchers, and the public?

Over the past few years, there has been a proliferation of generative AI tools capable of producing a wide variety of content, including music. For example, Boomy, launched in 2018, enables users to generate soundtracks by choosing from a range of styles and filters; OpenAI's ChatGPT (GPT-3), initially released in November 2022, can generate virtually any written content on demand, from poems, to travel plans, legal documents, lyrics and musical chord progressions;

GitHub's Copilot, launched in October 2021, can turn natural language prompts into coding suggestions across dozens of language. Generative AI technology continues to improve at an accelerating rate. In September 2023, OpenAI released an enhanced, multimodal version of ChatGPT (GPT-4V), which can now process and generate text, image, ***and*** audio. All of these generative AI tools have one thing in common: they are trained on massive datasets, including copyrighted content, in order to generate outputs that mimic human creations.

Like numerous technological developments in the past—ranging from drum machines to beat makers to autotune—generative AI has the potential to enhance the scope of musical creativity by supporting the creation of original music by humans. However, generative AI tools introduce a significant new risk: trained on immense datasets of original human-created musical compositions, these tools are able to generate new machine-made outputs that will compete with, and potentially displace, the human music creators whose compositions were used to train the AI tools. The ability of generative AI tools to produce—nearly instantaneously and at massively large scale—musical works qualitatively comparable to those created by humans is improving at a rapid pace, with current tools vastly outperforming those in existence even just a couple of years ago. The nightmare scenario is that AI will generate so much machine-made music so rapidly that there will soon be no longer a need for humans to create music.

Generative AI tools thus pose the foreseeable and existential risk of supplanting, rather than supporting, the creation of original music by humans. Indeed, these tools threaten to displace the very creators and artists whose creative labor they exploit as training datasets and without which they could not generate any music. As discussed in more detail in our responses to the questions below, we submit that a voluntary licensing scheme whereby the copyright holders can choose whether they wish their content to be used for AI training is the best way to harness the

power of generative AI without threatening the livelihood of creators. To that end, the Copyright Office and policymakers should provide the necessary incentives for the providers to come to the negotiating table with the rights holders, including by making it clear that a license must be obtained before using any copyrighted content for AI training.

2. Does the increasing use or distribution of AI-generated material raise any unique issues for your sector or industry as compared to other copyright stakeholders?

Unlike some other forms of protected content, a single musical composition or sound recording often reflect the creative input of numerous—potentially dozens of—contributors, including through melody composition, lyrics authorship, instrumentation, voice recording, and production. Each of these stakeholders is independently subject to encroachment by generative AI tools, and each is entitled to protection under copyright laws. Therefore, developers of AI models/tools and creators of training datasets should ensure that every aspect of the protected content used as training data has been properly licensed.

Policymakers should not be fooled by claims from technology companies that licensing requirements would hamper the development of generative AI. Their predecessors have made similar arguments for previously "new" technologies in an attempt to operate above the law in the name of progress. But time and again, new technologies have adapted to and flourished under the requirements of copyright laws. Generative AI will be no exception. AI developers and platforms do not need unlimited rights to acquire and use—by any means possible—all of the content and data available on the Internet. Nor is a broad notion of "innovation" sufficient justification for circumventing copyright rulemaking or legislation. Like every other business, AI developers and platforms should be required to identify the third-party rights implicated by their business plans, and then make the necessary investments to secure those rights. If those rights cannot be secured,

the answer cannot be to proceed in violation of the law. Instead, their business plans must be adjusted accordingly.

The widespread success of licensing in the streaming, entertainment, hospitality, and other industries—including the many hundreds of thousands of businesses currently licensed by ASCAP—illustrates that voluntary collective licensing is practically feasible and mutually beneficial for both music creators and the countless businesses that derive value from the right to publicly perform musical works. Like these other industries, the AI industry must both obtain consent and pay fair market value for the benefits they derive from human music creators; some AI developers, including Google, have already done just that. Markets are flexible, and voluntary licensing can effectively address all stakeholder concerns while upholding the purpose of the copyright laws as provided for in the United States Constitution.

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.

ASCAP commissioned a "State of the Songwriter" study earlier this year to understand music creators' sentiments about the industry and their perceptions of the biggest changes and challenges facing the music industry. Almost 2,000 music creators[1] were surveyed as part of this study. Half of the creators believe that AI is a threat to their livelihood. Seven out of ten creators want the ability to choose whether their music can be used for training AI models. Eight out of ten creators think that AI needs to be better regulated. Nine out of ten creators believe that they should be compensated for the use of their music in AI models.

[1] The songwriters include ASCAP members, members of other PROs, and unaffiliated music creators.

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?

International consistency is crucial for the licensing of musical works. Music licensing, particularly with respect to the right of public performance, is governed by numerous interconnected global data platforms used by collection societies, music publishers and creators to register copyrights, track and identify musical performances by licenses, and properly pay rights holders.

Public performance rights are generally licensed on a territorial basis by local PROs—like ASCAP in the U.S. Through the use of reciprocal agreements, PROs from across the world cross license their repertories in a manner that permits global licensing and monetization of their repertories. To enable global identification of musical performances and the proper exchange of compensation, PROs around the world rely on musical works copyright databases and matching systems, as well as certain uniform rules, including those governing repertory and the use of metadata and unique identifiers. To the extent countries adopt inconsistent regulations concerning AI, the inconsistencies can create loopholes or back doors that can disrupt music use, facilitate copyright infringement and enforcement of laws in other countries, and affect the current application of international treaties related to copyright and international trade.

The Artificial Intelligence Act in the European Union (the "EU AI Act") is the first comprehensive law regulating the use of AI. To the extent that the Act requires generative AI systems to comply with the following transparency requirements, we believe the law sets a positive and rational starting point for AI transparency legislation: (a) disclosing that content was generated by AI; (b) designing AI models in such a way as to prevent them from generating illegal content; and (c) publishing summaries of copyrighted data used for training.

Given the complexity of the international network of music licensing, it is important to note that on October 12, 2023, SACEM, the French national PRO, announced it was exercising its right, on behalf of its nearly 211,000 members and with respect to the 166 million musical works in its repertory, to "opt out" of Article L122-5-3 of the French Intellectual Property Code (which implements the EU Directive on Copyright in the Digital Single Market's provision of text and data mining as an exception to copyright protection[2] (the "EU Directive")). As a result of this "opt out", any entity seeking to feed their training databases and carry out data mining activities based on the works in SACEM's repertory will have to request prior authorization from SACEM and expressly negotiate the terms of such proposed use.

As discussed in more detail below, AI developers and providers should also obtain rights holders' consent before using their copyrighted works for AI training. ASCAP and other PROs are well-positioned to negotiate collective licenses on behalf of music creators.

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.

The existing U.S. copyright laws may be adequate to address the new challenges posed by generative AI tools. However, AI tools are dynamic technologies and judicial interpretations are evolving. As explained in more detail in Questions 6 and 8 below, under the existing copyright law, AI companies simply cannot reproduce, distribute, create derivative works of, or publicly perform a copyrighted musical work without first securing permission from the rights holders. Nevertheless, we maintain that, given prior history when new technologies have used copyrighted

[2] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market.

musical works, some clarifications or guidance by the Copyright Office may be beneficial to protect the rights of the owners of those musical works.

First, we suggest adding a clarification, through guidance issued by the Copyright Office, that the public performance of generative AI outputs that incorporate any "Musical Works Public Performance Data" (as defined below) implicates the exclusive right of public performance under Section 106(4) of the Copyright Act.

"Musical Works Public Performance Data" shall mean audio data made available on a website or web application which is used by such website or web application to publicly perform musical works to users of such website or web application (as well as any replicas or extractions of such data). While ASCAP believes that the Copyright Act and its interpretation in governing caselaw already affords such protection, this clarification will provide useful guidance to the AI industry in ensuring their compliance with U.S. copyright law.

ASCAP further maintains that a new federal right of publicity is important to adequately protect creators' rights in the context of generative AI. Today, generative AI tools exist that, when trained on samples of a person's voice, likeness, musical sound or lyrics, are able to generate wholly new images, video, sound and lyrics that are very difficult to distinguish from those of the real person. In the context of music content, this technology can be—and has been—used to generate entire songs and albums exactly mimicking a person's specific vocal, melodic or lyrical style without consent or remuneration.

When an AI user uses that person's name to generate such musical content (*i.e.*, by using the name as part of a generative AI query or prompt), there can be no doubt that the user and the AI intends to mimic the person's voice, sound or lyrics, and, accordingly, such usage should

require legal authorization. Such uses of generative AI also exploit the intellectual property of the composers and songwriters on whose creative works the generation is based.

In addition to other harms, the AI-generated content has the potential to act as a market substitute to human creations. For instance, when voice samples of the popular Singaporean singer Stefanie Sun were used, without her consent, to generate viral songs in her style, the singer commented, "[m]y fans have officially switched sides and accepted that I am indeed 'an unpopular singer' while my AI persona is the current hot property . . . . [H]ow do you fight with someone who is putting out new albums every few minutes?"[3]

While such "voice cloning" and related conduct is subject to existing state laws governing the right of publicity, the ubiquity and scale of this new technology requires a robust federal law ensuring that creators' rights are adequately protected.

B. Training

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

One major challenge facing copyright holders with respect to protection and enforcement in the generative AI context is the lack of transparency from AI companies concerning their training data.

Only a limited number of AI providers have disclosed the sources or categories of their training data, and typically only at a high level. But it is no secret that a massive amount of copyrighted works have been used to train generative AI models without consent by rights holders.

[3] Ashley King, *Singaporean Singer Stefanie Sun's Career Hijacked by AI – "My AI Person Is the Current Hot Property*, DIGITAL MUSIC NEWS (May 30, 2023), https://www.digitalmusicnews.com/2023/05/30/singaporean-singer-stefanie-sun-career-hijacked-ai/#:~:text=Singaporean%20singer%20Stefanie%20Sun's%20voice,Sun's%20voice%20has%20exploded%20online.

A number of copyright infringement lawsuits[4] have been filed against AI companies alleging the unauthorized use of various types of copyrighted works for AI training, including books, images, software codes, and song lyrics.

OpenAI's CEO has testified in Congress that OpenAI has trained its Large Language Model (LLM) (*i.e.*, GPT) on "large, publicly available datasets that include copyrighted works," and that these copyrighted works are crucial to the training of LLMs. He further admitted that if copyrighted works were not used, then it would "lead to significant reductions in model quality."[5] Given, on the one hand, the importance of copyrighted works in generative AI development, and on the other, the massive displacement potential of generative AI against creators and copyright holders whose very works were used to train generative AI models, AI providers must be required to keep records and provide disclosures of their training data. And they must start taking actions now, as the amount of training data is growing by day.

Since there is more publicly available information concerning OpenAI's generative AI products compared to those developed by other AI companies, below, we use OpenAI's ChatGPT (capable of generating lyrics and chord progressions), MuseNet (capable of generating instrumental music), and Juke Box (capable of generating songs with lyrics) as illustrative examples for the collection and curation training data for generative AI models.

---

[4] *Tremblay v. OpenAI*, 3:23-cv-03223 (N.D. Cal. June 28, 2023); *Silverman v. OpenAI*, 3:23-cv-03416 (N.D. Cal. July 7, 2023); *Kadrey v. Meta*, 3:23-cv-03417 (N.D. Cal. July 7, 2023); *Andersen v. Stability AI*, 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023); *Doe v. Github*, 3:22-cv-06823 (N.D. Cal. Nov. 3, 2023); *Getty Images v. Stability AI*, 1:23-cv-00135-UNA (D. Del. Feb. 3, 2023); *Chabon v. OpenAI , Inc.*, No. 3:23-cv-04625 (N.D. Cal. Sept. 8, 2023); *Chabon v. Meta Platforms Inc.*, No. 4:23-cv-04663 (N.D. Cal. Sep 12, 2023); *Authors Guild v. OpenAI Inc.*, No. 1:23-cv-08292 (S.D.N.Y. Sept. 19, 2023); *Concord Music Group, Inc. v. Anthropic PBC*, No. 3:23-cv-01092, (M.D. Tenn. Oct. 18, 2023).

[5] *Oversight of A.I.: Rules for Artificial Intelligence: Hearing Before the S. Judiciary Comm. Subcomm. on Privacy, Tech. and the Law*, 118th Cong. (2023) (testimony of OpenAI CEO Sam Altman), *available at* https://techpolicy.press/transcript-us-senate-judiciary-hearing-on-oversight-of-a-i/ (last accessed Sept. 19, 2023).

OpenAI's ChatGPT, a GPT-powered chatbot that conducts human-like communications in response to user queries, was trained on approximately one petabyte of text data (*i.e.*, approximately 500 billion pages of text).[6] This text data includes a combination of data copied from a public database called the Common Crawl, an organization that creates datasets from periodic scraping (*i.e.*, data extraction) of the Internet, as well as data that OpenAI scraped directly from the Internet. The Common Crawl database includes domain names that contain a wide variety of copyrighted materials (*e.g.*, books, articles, journals, essays, commentaries etc.) such as yahoo.com (Yahoo!), ox.ac.uk (Oxford University), nih.gov (National Institute of Health), justia.com (Justia), wikibooks.com (Wikibooks), mlb.com (Major League Baseball), worldbank.org (World Bank), and youtube.com (YouTube).[7] OpenAI also directly scraped similar copyrighted content from the Internet using a tool called WebText.[8] OpenAI has licensed GPT to other developers, who can fine tune the GPT model by feeding it additional training data tailored to specific use cases.

In addition to ChatGPT, OpenAI also developed a music generation tool called MuseNet, which was trained on copyrighted music and sound recordings from several databases, including but not limited to (1) Classical Archives—a classical music database; (2) BitMidi—a user generated database for early web-era music; and (3) the MAESTRO dataset—recordings from nine years of international piano competition events. OpenAI has also used other undisclosed databases

[6] *GPT-4 has more than a trillion parameters – Report*, THE DECODER (Mar. 25, 2023), https://the-decoder.com/gpt-4-has-a-trillion-parameters/.

[7] *Top-500 Registered Domains of the Latest Main Crawl*, https://commoncrawl.github.io/cc-crawl-statistics/plots/domains (last accessed Oct. 31, 2023).

[8] Open WebText, Github, https://github.com/jcpeterson/openwebtext?search=1.

in training MuseNet.[9] Similarly, Jukebox, another music generation tool by OpenAI, was trained on 1.2 million songs extracted from the Internet and corresponding lyrics and metadata from LyricWiki.[10] OpenAI has not disclosed the names of songs used to train Jukebox. It has, however, disclosed an exemplar list of the names of artists whose songs were used in the dataset, including Madonna, Frank Sinatra, the Beach Boys, Rihanna, and Elton John, to name a few.

The effectiveness of these generative AI models depends on access to a large quantity of *high-quality* data. In the context of ChatGPT, for example, its ability to simulate human creations is attributed to the large number of high-quality texts on which it was trained. High-quality texts are factually accurate texts that use sophisticated grammar, syntax, and sentence structure, such as books, research papers, and essays, which are more likely to be copyrighted works.[11] By contrast, low quality data refers to user-generated content such as social media posts. Similarly, OpenAI uses high quality music, such as classical music recordings or songs by famous artists, to train MuseNet and Juke Box.

Despite their use of a large quantity of copyrighted materials in training generative AI models, AI providers like OpenAI did not obtain prior consent from copyright holders for the use of their works as training data, and with very few exceptions, there is currently no transparency surrounding the datasets used to train generative AI models. Therefore, rights holders like ASCAP can only infer the use of their copyrighted works when they encounter blatantly infringing outputs

[9] Muse, OpenAI (Apr. 25, 2019) https://openai.com/research/musenet; Curtis Hawthorne et al., *Enabling Factorized Piano Music Modeling and Generation with the Maestro Dataset* (Jan. 17, 2019), https://arxiv.org/pdf/1810.12247.pdf.

[10] Jukebox, OpenAI (Apr. 30, 2020) https://openai.com/research/jukebox.

[11] Pablo Villalobos, et. al, *Will we run out of data? An Analysis of the limits of scaling datasets in Machine Learning* (Oct. 26, 2022), https://arxiv.org/pdf/2211.04325.pdf#:~:text=Data%20stocks%20grow%20at%20a,and%202060%20for%20vision%20data.

generated by these models or through domain names listed on the limited number of publicly available databases, such as the Common Crawl, which are known to have been used for training certain models. For example, Universal Music Group ("UMG") is reported to have discovered over 4,000 AI-generated outputs on the Internet that potentially infringes on UMG's copyrights or its artists' names or likenesses.[12] Some of the AI-generated music has been live streamed or broadcast, and thus also potentially infringes on ASCAP's public performance rights.

The lack of transparency surrounding the compilation of the training datasets makes it almost impossible for copyright holders to enforce their rights until it is too late—*i.e.*, the model has already "learned" the copyrighted works. As will be discussed in more detail in Question 7.3 below, once a model has learned a particular piece of copyrighted content, there is no effective way for it to "unlearn" the content. Therefore, it is crucial to enable copyright holders to safeguard their rights at the outset—*i.e.*, AI companies must seek their consent before the copyrighted content is ingested into the model.

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)?

As discussed under Question 6 above, with few exceptions, there is currently no transparency surrounding the datasets used to train generative AI models. Based on publicly available information, generative AI models are trained on a combination of data copied from existing third-party databases as well as data scraped or extracted directly by the developers. For example, as discussed above, OpenAI's ChatGPT was trained on both content copied from the existing Common Crawl database and content that OpenAI directly scraped from the Internet.

[12] *See* Transcript of the Copyright and AI Music and Sound Recordings Listening Session at 110 (May 31, 2023).

Similarly, OpenAI's music generative AI tools (MuseNet and Juke Box) are trained on a combination of existing user-generated datasets like BitMidi and songs OpenAI directly extracted from the Internet. However, regardless of the data sources or the methods used to obtain them, copyrighted materials have been obtained and used for AI training purposes without authorization from copyright holders.

In addition, in most cases where training data was accessed, scraped or extracted from the Internet by AI companies, it was done in blatant violations of the terms of use of the relevant websites, giving rise to potential breach of contract claims by such websites.

For example, YouTube's terms of service prohibits access to YouTube "using any automated means (such as robots, botnets or scrapers) except (a) in the case of public search engines, in accordance with YouTube's robots.txt file; or (b) with YouTube's prior written permission."[13] *The New York Times* also updated its Terms of Service in August this year to reflect the prohibition against the use of its content "in connection with: (1) the development of any software program, including, but not limited to, training a machine learning or artificial intelligence."[14] Therefore, access to, and use of content from these websites for AI training purposes breach their terms of use unless there is prior permission from the website owners.

Moreover, use of pre-existing databases for AI training does not alleviate the user's obligations to comply with the terms of use of the source websites. For example, Common Crawl's Terms of Use specifically contemplate the application of terms of use of the source content of the Common Crawl database, by providing that visitors who use its website and/or data "acknowledge

[13] *Terms of Service*, YouTube, https://www.youtube.com/static?template=terms (last visited Oct. 30, 2023).

[14] Jess Weatherbed, *The New York Times Prohibits Using Its Content To Train AI Models*, THE VERGE (Aug. 14, 2023), https://www.theverge.com/2023/8/14/23831109/the-new-york-times-ai-web-scraping-rules-terms-of-service; *Terms of Service*, N.Y. TIMES, https://help.nytimes.com/hc/en-us/articles/115014893428-Terms-of-Service.

and agree that all information, data, text, scripts, web pages, web sites, software, html page links, open data APIs, metadata or other materials . . . may be subject to separate terms of use or terms of service from the owners of such Crawled content."[15]

6.2 To what extent are copyrighted works licensed from copyright owners for use as training materials? To your knowledge, what licensing models are currently being offered and used?

Consent by the copyright holders must be obtained before their works can be used in training generative AI models. As explained in more detail in our response to Question 7 below, the development of a generative AI model requires multiple reproductions of copyrighted training materials. Once the model processes these materials, it is able to generate content substantially similar to and/or create unauthorized derivative works based on these materials. And once trained, the model cannot effectively "unlearn" any of the training materials, including copyrighted works obtained without authorization. The very purpose of these generative AI models is to produce commercial substitutes to the materials used to train them. Without licensing, creators' copyrighted works will continue to be used by generative AI to create "engines of their own destruction," with zero compensation to the copyright holders.

To date, no generative AI developer has sought or obtained a music public performance license from ASCAP in connection with the development or operation of any generative AI model. In fact, ASCAP is not aware of any music copyright license that has been issued for generative AI purposes.[16] Despite the lack of any license, as noted above, copyrighted musical works, including sound recordings, compositions, and lyrics, have been used to train generative AI models without

[15] *Full Terms of Use*, Common Crawl, https://commoncrawl.org/terms-of-use/ (last updated Aug. 21, 2023).

[16] A couple of unverified news articles concerning one deal have been published. *See* Shannon Thaler, *Warner Music Bashed For Signing Record Deal With 'Creepy' AI Pop Star With The Body Of A '12 Year Old,'* THE NEW YORK POST (Sept. 8, 2023) https://nypost.com/2023/09/08/warner-music-signs-first-ever-record-deal-with-ai-pop-star/; Anna Nicolaou, *Google And Universal Music Negotiate Deal Over AI 'Deepfakes,"* FINANCIAL TIMES (Aug. 8, 2023), https://www.ft.com/content/6f022306-2f83-4da7-8066-51386e8fe63b

authorization. And those models have generated outputs that potentially infringe on these music copyrights and threaten the livelihood of musicians and creators.[17]

Licensing is therefore necessary for the use of copyrighted works in training generative AI. It is also feasible. In fact, OpenAI recently obtained a license from the Associated Press for the use of the latter's news archive in connection with OpenAI's generative AI offerings.[18] Similar deals can be achieved in the music industry. Collective licensing agreements, negotiated and managed by collective rights organizations like ASCAP, have long been a staple in the music industry and have adapted to major technological developments such as the advent of cable television, digital music streaming services and audio-visual streaming services. Generative AI will be no exception to our ability to adapt and evolve, but new laws need to be put in place immediately. Time is of the essence.

For example, ASCAP currently negotiates, executes and manages collective licensing agreements with hundreds of thousands of businesses and identifies and matches trillions of musical performances across all types of media with the assistance of AI tools and performance files provided by its licensees who enter into a blanket license. ASCAP needs this usage information from licensees to properly track and pay its members. The benefit of a collective license from ASCAP is that it gives the licensee the right to perform as much or as little of the approximately 18 million works in the ASCAP repertory. Blanket licensing arrangements are potentially well suited for generative AI, which utilizes a tremendous amount of training data and

---

[17] *See* Transcript of the Copyright and AI Music and Sound Recordings Listening Session at 110 (May 31, 2023); Ashley King, *Singaporean Singer Stefanie Sun's Career Hijacked by AI – "My AI Person Is the Current Hot Property*, DIGITAL MUSIC NEWS (May 30, 2023), https://www.digitalmusicnews.com/2023/05/30/singaporean-singer-stefanie-sun-career-hijacked-ai/#:~:text=Singaporean%20singer%20Stefanie%20Sun's%20voice,Sun's%20voice%20has%20exploded%20online.

[18] Matt O'Brien, *Chatgpt-Maker Openai Signs Deal With AP To License News Stories*, ABC NEWS (Jul. 13, 2023), https://www.wric.com/news/u-s-world/ap-chatgpt-maker-openai-signs-deal-with-ap-to-license-news-stories/.

requires access to millions of copyrightable works. With such powerful tools as generative AI at their disposal, generative AI companies are well equipped to track and manage the use of copyrighted works just as licensees such as Apple, Amazon, Microsoft, Google and others are required to do today.

What is currently missing from the equation for collective licensing for generative AI is an incentive for AI companies to disclose the unauthorized usage of copyright material and come to the table in good faith to negotiate fair and reasonable remuneration from creators whose works they are unfairly exploiting.

As of now, AI companies do not take the view that the use of copyrighted works in generative AI development is unlawful. The Copyright Office must therefore play a critical role in the development of a rational and legitimate market for the use of copyrights in generative AI by making it clear to AI platforms that they must obtain the consent of the rightsholders of the copyrighted materials they use.

If history is prologue, the industry has seen similar behavior with respect to the Digital Millennium Copyright Act, whereby users would use copyrighted material without permission and beg for forgiveness after the fact.

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?

As discussed in Question 6 above, with few exceptions, there is currently no transparency surrounding the datasets used to train generative AI models. Given the large amount of data used for training, it is likely that both copyrighted and non-copyrighted materials were used. For example, studies have shown that some public domain books, such as *1984* and *the Adventures of*

Tom Sawyer, were used to train GPT-4.[19] The Classical Archives database that was used to train MuseNet, for example, appears to contain certain copyrighted sound recordings of classical music, but the underlying musical compositions themselves may be in the public domain. This question again goes to show the need for greater transparency surrounding what goes into the training data for generative 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.

Although ASCAP has no visibility into the storage and retention practices of generative AI developers and providers, it is reasonable to assume that training materials are necessarily retained in some form because generative AI models operate by recalling information—*i.e.*, learned patterns about the training data—it retained from the training process.[20] Generative AI models, at their core, are probabilistic mathematical models constructed by mapping the relationships of the various building blocks of content in their training data.

In the case of LLMs like ChatGPT, this means a probabilistic model that maps how letters, words, and concepts relate to one another, and how close or attenuated those relationships are.[21]

---

[19] *GPT-4 Memorizes Contents Of Copyrighted Books, And It Could Be A Cultural Issue*, THE DECODER (May 4, 2023) https://the-decoder.com/gpt-4-memorizes-contents-of-copyrighted-books/.

[20] *How Will Memory And Storage Evolve With The Generative AI Paradigm Shift*, KEARNEY (Aug. 10, 2023), https://www.kearney.com/industry/technology/article/how-will-memory-and-storage-evolve-with-the-generative-ai-paradigm-shift.

[21] Stephen Wolfram, *What is ChatGPT doing and why does it work* (Feb. 14, 2023), https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/ ("Strictly, ChatGPT does not deal with words, but rather with "tokens"—convenient linguistic units that might be whole words, or might just be pieces like "pre" or "ing" or "ized". Working with tokens makes it easier for ChatGPT to handle rare, compound and non-English words, and, sometimes, for better or worse, to invent new words.").

In the case of music generative AI models like MuseNet, this means a probabilistic model that maps chords, melodies, beats, rhythms, pitch, volume, and instrumentation.[22] During the training process, this mapping is reinforced or adjusted as the model tests itself by attempting to recreate the training content. Once trained, the model generates content by relying on the mapping it has retained from the training process. As a result, generative AI models are capable of generating, and have generated, content substantially similar to and/or are derivative works of the training materials.

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?

ASCAP's response to this Question is based on publicly available information concerning the training of generative AI models. Although there are differences among generative AI models with respect to their underlying algorithms, the types of data they process, and the technical environments in which they are hosted, their training processes are similar on a general level. All of them involve reproduction and creation of derivative works of their training data. Below, we illustrate the training processes using ChatGPT as an example, as there is greater availability of information and research concerning the development of this model.

[22] Muse, OpenAI (Apr. 25, 2019) https://openai.com/research/musenet; Curtis Hawthorne et al., *Enabling Factorized Piano Music Modeling and Generation with the Maestro Dataset* (Jan. 17, 2019), https://arxiv.org/pdf/1810.12247.pdf.

ChatGPT is a generative AI language model that is trained on vast troves of raw data and can be adapted for a wide variety of use cases. ChatGPT, for example, can accomplish tasks ranging from writing lyrics and composing musical chords, to providing fashion tips, travel planning, and recipes. The raw training data is used to develop a statistical model that can identify and predict patterns in language, such as grammar, tone, style, content, structure, and contextual relationships.[23] At a high level, the training process of the model can be broken down into three stages:

- First, data is extracted from their sources and downloaded onto servers. This process is called "scraping" and may involve the reproduction of exact copies of entire (copyrighted) works. In the case of GPT, for example, approximately 500 billion of tokens have been copied from the Internet.[24]

- Second, the scraped data is stored on servers and converted into a format that can be processed by the computers and the neural networks that are being built by the computer. This process is known as embedding and involves additional reproduction as well as preparation of derivative works of the scraped (copyrighted) works. The scraped data is then ingested into the LLM for training.

- Third, the ingested data is processed in the model's "neural networks," a computational process inspired by the human brain. Neural networks are designed to build probabilistic associations between different pieces of data. In the case of GPT, the model repeatedly analyzes the text data in their original structure, and identifies patterns from the various ways in which letters, words, phrases, and concepts relate to one another, and how close or attenuated are those relationships. The model is then programmed to repeatedly test and refine itself through trial and error by seeing how well it can re-create text from training data. For example, if an ASCAP press release is ingested as training data into the model, the model would attempt to recreate the article word by word in the training process and compare its recreation with the original article to ensure accuracy. This process therefore involves further reproduction as well as preparation of derivative works of the ingested copyrighted works.

---

[23] Jennifer Langston, *Microsoft Announces New Supercomputer, Lays Out Vision For Future AI Work*, MICROSOFT (May 19, 2020), https://news.microsoft.com/source/features/ai/openai-azure-supercomputer/.

[24] Sheera Frenkel & Stuart A., *Thompson Data Revolts Break Out Against A.I.*, N.Y. TIMES (July 15, 2023), https://www.nytimes.com/2023/07/15/technology/artificial-intelligence-models-chat-data.html. *See also* Matthias Bastian, *GPT-4 Has More than a Trillion Parameters – Report*, THE DECODER (Mar. 25, 2023), https://the-decoder.com/gpt-4-has-a-trillion-parameters/.

Generative AI models that are designed specifically for generating music, such as MuseNet and Juke Box, often leverage similar neural networks, but unlike ChatGPT, they may be trained on audio or MIDI data (which is to computers what sheet music is to human musicians), and the model processes the components of music (such as melody, beat, tempo etc.) as opposed to letters and words.[25]

Therefore, to the extent a generative AI models uses copyrighted materials as training data, absent prior consent from the relevant copyright holders, the training of the model involves at least multiple instances of unauthorized reproduction and preparation of derivative works of the copyrighted materials, all in violation of the Copyright Act. Certain usage of that data to make available music to the public would implicate the public performance right as well.

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?

Based on publicly available information, there is currently no efficient way to make a generative AI model "unlearn" inferences it gained from training on a particular piece of material.

One potential way for a model to "unlearn" is to retrain the model on a dataset that excludes the material; but the cost for doing this is prohibitively expensive.[26] Because of AI algorithms'

[25] Jade Copet, et al., *Simple and Controllable Music Generation* (June 8, 2023), https://arxiv.org/abs/2306.05284. For example, OpenAI's MuseNet was trained on hundreds of thousands of MIDI (audio) files scraped from the Internet. To prepare for training, OpenAI broke down the MIDI files into a sequence of "tokens" that each represents a discrete element of musical performance, such as notes, figurations, motives, licks, phrasing, dynamics, and instrumentation, and articulations. During neural processing, the MuseNet model learns by recognizing patterns in the basic elements of music such as figurations, motifs, timbre, pitch, rhythm, and harmony, as well as how to arrange these elements into melodies, periods, and chord progressions. Muse, OpenAI (Apr. 25, 2019) https://openai.com/research/musenet; similarly, to train Jukebox, OpenAI extracted 1.2 million songs from the Internet alongside corresponding lyrics and metadata. It then used an "autoencoder model" to compress the large, complex audio files into simpler discrete codes for the model's neural processing. Jukebox, OpenAI (Apr. 30, 2020) https://openai.com/research/jukebox.

[26] *Announcing The First Machine Unlearning Challenge*, GOOGLE (June 29, 2023), https://blog.research.google/2023/06/announcing-first-machine-unlearning.html; *Machine Unlearning: The Critical Art Of Teaching AI To Forget*, VENTURE BEAT (Aug. 12, 2023), https://venturebeat.com/ai/machine-unlearning-the-

"black-box" nature, identifying problematic (*i.e.*, infringing) data in a model and understanding its impacts on other data points in the model are not easy and affordable tasks. Even upon identification of the relevant data, there are additional challenges associated with retraining the model without sacrificing the model's utility and closely monitoring the retrained model to ensure that the data is in fact pulled out. In addition, this "unlearning" process will only get more expensive and impractical due to a constant increase in the size and complexity of datasets used to train AI models.

Absent a more successful, efficient unlearning method than retraining, AI "unlearning" is not feasible. It would also be inappropriate to require copyright holders to monitor whether such "unlearning" had occurred and been successful. Therefore, unauthorized use of copyrighted materials in generative AI model training can cause irreversible and long-lasting harm to copyright holders. This is yet another reason why it is essential for AI platforms to obtain a license from the relevant copyright holders before using their copyrighted works for 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?

As discussed in our response to Question 6 above, it is possible, but extremely difficult, to infer the use of a particular piece of work in the training dataset based on the outputs generated by the AI model. For example, if the model, when prompted to generate a song that is similar to a copyrighted work, in fact reproduced the copyrighted song or generated one song that is substantially similar to the original, then an inference could be drawn that the copyrighted song was in the training data.

critical-art-of-teaching-ai-to-forget/.

Researchers have also attempted to develop adverse inference AI models for the specific purpose of determining whether or not a particular piece of data was used to train a machine learning model without accessing the underlying database.[27] However, this reverse inference process is costly and time-consuming for copyright holders, particularly given the amount of subscription fees and data limits that AI platforms currently impose on the use of their models.

Crucially, the burden should not be placed on copyright holders to develop sophisticated technology just to query whether their copyrighted works have been used in generative AI training. Generative AI platforms should bear the responsibility in the first place to disclose usage, keep track of their training data, and obtain copyright holders' consent in advance of usage.

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.

Based on our current understanding of how generative AI models are trained and deployed, we do not believe there is any realistic scenario under which the unauthorized and non-personal use of copyrighted works to train generative AI models would constitute fair use, and therefore, consent by the copyright holders is required. Below, we provide our analysis of the four statutory fair use factors as applied to the unauthorized reproduction and/or preparation of derivative works of copyrighted works in generative AI training, again using ChatGPT as an example because of the greater availability of information concerning the training process for this model.

The **four fair use factors** include (1) the nature and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the

[27] Nicholas Carlini et al., *Membership Inference Attacks From First Principles* (Apr. 12, 2022), https://arxiv.org/pdf/2112.03570.pdf; Benjamin Zi Hao Zhao et al., *On the (In)Feasibility of Attribute Inference Attacks on Machine Learning Models* (Mar. 21, 2021), https://arxiv.org/pdf/2103.07101.pdf.

copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work. No single factor is determinative; they must be weighed together in light of the purpose of copyright. *Andy Warhol Foundation for the Visual Arts, Inc.* v. *Goldsmith*, 143 S. Ct. 1258, 1274, 1288 (2023) ("*Warhol"*). Here, the four factors weigh strongly against fair use.

1. The nature and character of the use

In *Campbell* v. *Acuff-Rose Music, Inc.*, 510 U.S. 569, 578-79 (1994), the Supreme Court explained that this factor was a question of whether the new work "merely 'supersede[s] the objects' of the original creation . . . or instead add something new, with a further purpose…[I]t asks in other words, whether and to what extent the new work is 'transformative.'" Whether an infringing use was "transformative" was at the heart of a number of cases upholding a fair use determination, such as *Authors Guild* v. *Google, Inc.*, 804 F.3d 202, 223 (2d Cir. 2015) ("*Google Books*"). However, more recently, the Supreme Court made it clear in *Warhol* that the concept of "transformative" should not detract from the separate inquiry of whether the use is commercial in nature. *Id.* ("The undisputed commercial nature of AWF's use, though not dispositive, 'tends to weigh against a finding of fair use.").

There is no doubt that the development and deployment of generative AI models have largely been for commercial use (there are some academic uses, but those are not the focus of these comments). Developing a generative AI model requires substantial investment. Microsoft invested billions of dollars in OpenAI to support the development of generative AI models. It would be difficult to justify such investment without expectation of substantial financial gain.[28] The main

[28] Jennifer Langston, *Microsoft Announces New Supercomputer, Lays Out Vision For Future AI Work*, MICROSOFT (May 19, 2020), https://news.microsoft.com/source/features/ai/openai-azure-supercomputer/.

generative AI models that are currently available—chief among which is GPT—are clearly developed and deployed for commercial purposes. They are currently being monetized through subscription fees for downloadable apps, as well as commercial licenses. For example, OpenAI charges a subscription fee for the use of its GPT-4 powered tool ChatGPT-Plus and a license fee to developers and business for integration of GPT with their applications, products and services. In fact, OpenAI has projected that ChatGPT will make $200 million in 2023 and $1 billion in 2024.[29] A number of generative music AI tools, such as Mubert, Boomy, Soundful also charge subscription fees proportional to the volume downloads. In addition, at least some of the developers or deployers of generative AI models are integrating them with search engines, and earning significant returns for that use.

The use of copyrighted materials for the development generative AI models is not transformative. Each unauthorized use of the copyrighted material during the training process is done in furtherance of a commercial purpose. As discussed above, these unauthorized uses include unauthorized reproductions (*e.g.*, in the case of GPT, reproductions of the copyrighted works in their entirety) and preparation of derivative works during scraping, embedding, ingestion, and neural processing of the model:

1. The crawling and scraping of entire works was for the purpose of analyzing those works in their entirety in furtherance of a commercial purpose;

2. The ingestion of those works was done in a manner that maintained the works (largely) intact, again for the specific purpose of using them to create a commercial tool; and

3. The internal analysis within the neural network of the copies was again in furtherance of the creation of the commercial tool.

[29] Jeffrey Dastin, *Exclusive: Chatgpt Owner Openai Projects $1 Billion In Revenue By 2023*, REUTERS (Dec. 15, 2022), https://www.reuters.com/business/chatgpt-owner-openai-projects-1-billion-revenue-by-2024-sources-2022-12-15/.

As will be discussed in more detail below, the end product is a commercial tool that replaces the very copyrighted materials used to train the AI model. A key purpose of generative AI tools like ChatGPT is to obviate the need for humans to access the underlying copyrighted materials. Replacement is simply not transformative use.

2. The nature of the copyrighted work

This factor typically favors fair use if the copyrighted work is purely informational, as opposed to works involving creative expression. *Blanch* v. *Koons*, 467 F.3d 244, 250 (2d Cir. 2006); *Andy Warhol Foundation for Visual Arts, Inc.* v. *Goldsmith*, 11 F.4th 26, 45 (2d Cir. 2021) (cited approvingly by *Warhol*, 143 S.Ct. at 1288). As discussed, since an effective generative AI model requires training on high-quality content that is more accurate, expressive, and sophisticated, this factor should weigh against fair use of copyrighted materials for generative AI training. The musical compositions of the ASCAP members are quintessential creative works. They involve the creative endeavors, and personalized expression of each artist. This factor weighs strongly against fair use.

3. The amount and substantiality of the portion used in relation to the copyrighted work as a whole

This factor also weighs strongly against fair use because the copyrighted works are copied in their entirety for generative AI training purpose. As discussed, the scraping of copyrighted works from the Internet entails copying of complete works. Importantly, the AI model also attempts to recreate these works – in the case of GPT, word by word – as part of its neural processing.

4. The effect of the use upon the potential market for or value of the copyrighted work

This factor is considered the most important of the fair use factors, and strongly weighs against fair use here. *Harper & Row Publishers, Inc.* v. *Nation Enterprises*, 471 U.S. 539, 566

(1985) (describing the fourth factor as "undoubtedly the single most important element of fair use."). The factor seeks to determine "whether, if the challenged use becomes widespread, it will adversely affect the potential market for the copyrighted work." *Bill Graham Archives,* 448 F.3d at 613; *see also Warhol,* 11 F. 4th at 48. As the Second Circuit stated in *Authors Guild* v. *Google, Inc.*, 804 F.3d 202, 223 (2d Cir. 2015) ("*Google Books*"), "[b]ecause copyright is a commercial doctrine whose objective is to stimulate creativity among potential authors by enabling them to earn money from their creations, the fourth factor is of great importance in making a fair use assessment." (citing *Harper & Row*, 471 U.S. at 566.)

Generative AI is a tool of replacement. There has been ample reporting on its use cases in every conceivable domain of content creation: from writing poems, to designing posters, composing music and lyrics, taking exams, and responding to customer queries. In contrast to the fact pattern in *Authors Guild* v. *Google,* in which copying entire works was deemed to be in furtherance of the creation of a search engine that would lead a user back to the original source material (*Google Books*, 804 F.3d at 223-225), the purpose of generative AI is to obviate the very need to access the source materials—it is intended to be the provider of the ultimate answer.

Overall, each and collectively, the four factors weigh strongly against fair use. By exploiting copyrighted works without compensation to the copyright holders and in order to create commercial products to replace those very works, generative AI undermines the very purpose of the Copyright Act.

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?

As discussed in our response to Question 8 above, the "purpose and character" (*i.e.*, the first fair use factor) of copyrighted works to train a generative AI model reflects the commercial realities of generative AI as a tool of replacement. The concept of "transformative use" is simply one, non-statutory factor to be considered in an overall balancing. The degree of the transformation must be weighed against not only the commercialism of the use, but also against the existential threat that uncompensated use presents: both to individual entities like ASCAP, but also in its systemic impact across an entire intellectual property marketplace.

Generative AI models are trained on huge volumes of raw data (including those of entities like ASCAP) in order to have the capability to accomplish a wide range of tasks (and thus a wide range of commercialization opportunities) and importantly, to obviate the need for humans to read (or, in the case of music, listen to) the original materials themselves.

Generative AI is fundamentally different from any pre-existing technology, such as search engines, that have withstood scrutiny under the copyright law. Traditional search engines, for example, respond to a user query merely by providing links to potentially responsive materials, thereby generally encouraging the user to click on the links to obtain the full answers. By contrast, generative AI models like ChatGPT do not encourage users to access the source materials; ChatGPT, for example, obviates the need to click on any links at all because it constructs answers—sourced from its training data—that are aimed at being fully responsive to user prompts. Similarly, music generative AI tools have created soundtracks in artists' voices—sourced from the artists' copyrighted music—that attracted more listeners than the artists' own music.[30] The

[30] Chloe Veltman, *When You Realize Your Favorite New Song Was Written And Performed By . . . AI*, NPR (Apr. 21, 2023), https://www.npr.org/2023/04/21/1171032649/ai-music-heart-on-my-sleeve-drake-the-weeknd.

replacement effect of these tools are succinctly captured by Singaporean singer Stefanie Sun, as we described earlier in our response to Question 5.[31]

Therefore, an analysis of the "purpose and character" of the use of copyrighted material at each step of the training process must account for the nature of generative AI as the ultimate tool of replacement. As discussed in our response to Question 8 above, each unauthorized (and non-personal) use of copyrighted materials during the training process, including scraping, embedding, ingesting, and neural processing—and regardless of whether it was done as part of pre-training of a model or fine-tuning—was done for the commercial purpose of replacing the need to access the original copyrighted materials; such use is not transformative, and weighs against fair use.

8.2 How should the analysis apply to entities that collect and distribute copyrighted material for training but may not themselves engage in the training?

Entities that collect and distribute copyrighted works for training are reproducing or distributing copyrighted material for training AI materials without authorization. If their collection is done knowingly and for the purpose of providing such unauthorized material to trainers of AI models, then they should be liable for both direct and secondary infringement. For example, a party who merely reproduced copyrighted materials for training could be liable for contributory infringement arising out of the reproduction of copyrighted materials during neural processing if the party it had knowledge that the materials would be used in the neural processing. *See Gershwin Publ'g Corp.* v. *Columbia Artists Mgmt., Inc.*, 443 F.2d 1159, 1162 (2d Cir. 1971) (A party is liable for contributory infringement if it (1) has knowledge of the infringing activity and (2) makes a material contribution to the infringement).

[31] Ashley King, *Singaporean Singer Stefanie Sun's Career Hijacked by AI – "My AI Person Is the Current Hot Property*, DIGITAL MUSIC NEWS (May 30, 2023), https://www.digitalmusicnews.com/2023/05/30/singaporean-singer-stefanie-sun-career-hijacked-ai/#:~:text=Singaporean%20singer%20Stefanie%20Sun's%20voice,Sun's%20voice%20has%20exploded%20online.

We believe that this question is essentially asking whether an entity such as Common Crawl can or should be liable for infringement. Our view is that, while there may have been a time when Common Crawl scraped content purely for academic or research purposes, these are not the primary uses for which AI model developers currently use its databases. It is now well understood that new web content scraped by Common Crawl or similar organizations are facilitating the development of commercially viable AI models. Common Crawl or similar organizations should be held responsible for knowingly facilitating infringement.

8.3 The use of copyrighted materials in a training dataset or to train generative AI models may be done for noncommercial or research purposes. How should the fair use analysis apply if AI models or datasets are later adapted for use of a commercial nature? Does it make a difference if funding for these noncommercial or research uses is provided by for-profit developers of AI systems?

As discussed in our responses to Questions 8-8.2 above, the major generative AI models currently available, such as GPT, have been deployed in commercial products. Given the significant investment required for the development of generative AI models, ASCAP is currently not aware of any effective generative AI models that have been solely developed for research or non-commercial purposes. OpenAI, for example, was originally founded in 2015 as a non-profit research organization. However, it transitioned into a capped-profit company in 2019 in order to attract investment to scale its development of generative AI models.[32] OpenAI relied substantially on Microsoft's resources in developing GPT and ChatGPT. In 2020, Microsoft and OpenAI co-developed a supercomputer for the very purpose of training GPT.[33] The supercomputer, the GPT

[32] OpenAI LP, OpenAI (Mar. 11, 2019), https://openai.com/blog/openai-lp.

[33] Jennifer Langston, *Microsoft announces new supercomputer, lays out vision for future AI work*, MICROSOFT (May 19, 2020), https://news.microsoft.com/source/features/ai/openai-azure-supercomputer/.

model, and ChatGPT, are all hosted on Microsoft's cloud computing platform Azure.[34] Microsoft has announced that it will invest an additional $10 billion in OpenAI this year.

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?

As discussed in our responses to Questions 6 and 8 above, the development of generative AI models requires training on a large quantity of high-quality content. GPT, for example, was trained on approximately one petabyte of text data (equivalent to approximately 500 billion pages); without the high volume of high-quality content, GPT would not be as successful at simulating human-like conversations. Both the quantity and quality of training data are therefore key in enabling generative AI models to potentially replace the need to access a wide range of original content, and weigh against fair use.

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?[35] 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?

As discussed in our responses to Question 8 and 8.1, given the sheer amount of content used to train generative AI models, a single model like GPT has unprecedented potential to replace a wide range of copyrighted content by numerous creators spanning various industries. Moreover, as the models cannot effectively "unlearn" any unauthorized content, any harms to the copyright holders are potentially irreversible. Therefore, any analysis for replacement effect under the fourth factor should not be limited to a single piece of copyrighted work or an artist, but also the market for that type of copyrighted work in general. To that end, disclosures by AI platforms concerning

[34] *Id.*

their training data and information concerning the types of content that have been generated through their models are highly valuable, if not necessary, for this analysis.

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

Consistent with the fundamental principles of copyright law, copyright owners' affirmative consent must be obtained before their copyrighted materials can be used for any commercial purpose, including training generative AI models. Requiring affirmative consent encourages AI platforms to work with copyright owners—or otherwise work with uncopyrighted materials as training data—which can help steer AI towards facilitating the human creation of original works rather than replacing them.

With respect to the public performance rights held by ASCAP, affirmative consent by copyright holders could be achieved under the framework of ASCAP's existing blanket license scheme. As discussed in our response to Question 6.2 above, ASCAP's current blanket license allows businesses to publicly perform the entire ASCAP repertory of more than 18 million musical works, saving time and money for every copyright owner and every business that uses music. This framework is well-suited to generative AI, which can potentially be leveraged to further enhance the implementation of the framework through more efficient data management.[36]

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

As discussed in our response to Question 8.3 above, we are currently not aware of any effective generative AI models that have been developed solely for non-commercial purposes.

[36] Michele Iurillo, *Harnessing the Power of Generative AI for Data Management and Governance: Emphasizing the Vital Role of Metadata*, LinkedIn (Aug. 28, 2023), https://www.linkedin.com/pulse/harnessing-power-generative-ai-data-management-vital-role-iurillo/.

Moreover, the Copyright Act eliminated any for-profit requirement with respect to the infringement of public performance rights. Therefore, ASCAP routinely issues blanket licenses to various noncommercial music users such as public radio broadcasters, nonprofit symphonies, museums and others. In addition, as discussed in our response to Question 7.3 above, there is currently no effective way for a generative AI model to "unlearn," thereby potentially causing long-lasting and irreversible harm to copyright holders. Given the commercial realities of generative AI and the amount of potential harm at stake here, affirmative consent should be required for all uses of copyrighted works to train generative AI models.

9.2 If an ''opt out'' approach were adopted, how would that process work for a copyright owner who objected to the use of their works for training? Are there technical tools that might facilitate this process, such as a technical flag or metadata indicating that an automated service should not collect and store a work for AI training uses?

As discussed in our responses to Questions 9 and 9.1 above, affirmative consent by copyright holders must be obtained before their copyrighted materials can be used for training generative AI models. We oppose an opt-out approach because it assumes that the copyrighted materials were obtained in a legitimate fashion to begin with; however, as discussed in our responses to Questions 8 and 8.1 above, there is nothing legitimate about scraping copyrighted content from the Internet and using it to develop a commercial tool to replace that content. The use of copyrighted materials for training generative AI is not transformative; it is not fair use; it undermines the very purpose of the Copyright Act. Therefore, consistent with copyright law, the affirmative consent of copyright holders must be obtained.

An opt-out approach condones the illegal actions of the AI platforms and undermines the rights held by copyright owners, as it effectively tells AI platforms that they can trample upon copyrights and ask for forgiveness later, when the model has already ingested, and could not

effective "unlearn," the copyrighted works. There is already information asymmetry (and resulting inequality in bargaining power) between AI platforms and creators concerning how and what training data is used; an opt-out approach empowers the AI platforms to further keep creators in the dark and leaves room for further abuse. To the extent any opt-out approach is adopted—and we strongly oppose this approach—it should at least require mandatory notification to the rightsholders identifying the copyrighted works that have been used in the training; it should also require the collection and maintenance of standard industry metadata and data identifiers for copyrighted works.

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?

The real obstacle here is the lack of willingness on the part of AI platforms to engage with copyright holders. When confronted with copyright claims, technology companies tend to overstate the obstacles to effective licensing. But ASCAP has negotiated and managed collective licensing agreements for over a hundred years, and has successfully adapted to, and leveraged, many technological developments for the protection of copyrights on the one hand, and the public enjoyment of new forms of creation on the other. ASCAP currently represents over hundreds of thousands of songwriters, composers, lyricists and music publishers across every genre, all 50 states and every stage of their career—from top hitmakers to those just starting out. ASCAP negotiates collective licensing agreements with hundreds of thousands of businesses, large and small, and tracks, with the help of AI technology, trillions of performances across all types of media. ASCAP is expert at collective licensing, and we know it can work for AI, just as it has worked for every other technology that has disrupted the music sector over the past century.

Similar collective licensing arrangements can be negotiated and executed for generative AI training.

9.4 If an objection is not honored, what remedies should be available? Are existing remedies for infringement appropriate or should there be a separate cause of action?

Remedies provided under the existing copyright law may be largely sufficient. However, because of the current lack of transparency surrounding the training data of generative AI models, we think there should be an affirmative requirement for AI platforms to notify copyright holders of the use of their copyrighted works in generative AI training before the data is ingested into the generative AI model. Penalties should be imposed for intentional and unintentional violation, unless the AI platform can make the model "unlearn" the copyrighted content within a reasonable time period after receiving notice of the violation.

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?

Since current U.S. copyright laws generally provide remedies to copyright owners, an author who does not own the copyright may not be able to object to the use of the work for generative AI training purposes. However, an author who has fully conveyed its copyright interest and right to control the work to another, and retains no ability to object to the training, might still have a state law claim for the right of publicity, for example, based on the generative AI's use of their voice or likeness. Given the lack of uniformity among states laws on the right of publicity, we support the creation of a new federal right of publicity to protect the name, voice and likeness of creators against unauthorized commercial use.

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

As discussed in our responses to Questions 6.2, 9, and 9.3, with respect to public performance rights of musical works, a similar arrangement to the current collective blanket licensing is well-suited to generative AI and can be negotiated between ASCAP (and, similarly, by other PROs) and AI platforms. What we need is for AI platforms to come to the negotiation table and engage with us in a meaningful way so that we can work out an arrangement together based on a well-tested collective licensing framework. To that end, it is crucial for the Copyright Office to make it clear that consent of the right holders must be obtained before the use of their copyrighted works for AI training. Absent such clear guidance, there is little incentive for AI platforms to come to the table.

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

As discussed in our responses to Questions 6.2, 9, 9.3 and 10, based on our experience, direct voluntary licensing is well suited for generative AI, and has worked successfully with respect to public performance rights of musical works in the U.S over the past century and through many technological developments. A similar arrangement can be replicated in other creative sectors too. Voluntary licensing—as opposed to compulsory licensing—is the most efficient solution to the copyright problems posed by generative AI. In order for it to work effectively, however, there needs to be clear guidance from the Copyright Office that consent of the rights holders must be obtained before the use of their copyrighted works for AI training.

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?

A voluntary, free market collective licensing scheme is both a feasible and desirable approach. It is feasible because, as discussed in our responses to Questions 6.2, 9, 9.3, 10, and 10.1, there is already a well-tested and efficient PRO collective blanket licensing arrangement in place that can serve as the blueprint for similar arrangements between collective rights organizations and AI platforms. Unfortunately, the main challenge we face at present is the unwillingness of AI platforms to accept the need for license for the use of copyrighted works in AI training and engage in good-faith licensing negotiations with copyright holders.

A voluntary, free market collective licensing scheme is also desirable because it satisfies the need to obtain consent, as well as the need to compensate music creators fairly and give them proper credit. Such a scheme also encourages transparency and facilitates a better process for the management of copyright information and compliance. In addition, voluntary collective licensing gives autonomy to the creators and allows market forces to allocate resources efficiently.

As a result, we oppose statutory changes concerning music licensing. In particular, we emphatically oppose any compulsory licensing requirements. Compulsory licensing in the music sector has been riddled with numerous inefficiencies and shortcomings to the detriment of music creators, resulting in price-suppression, multi-year litigations, and significant delays in payments to music creators. For example, the Copyright Royalty Board often takes more than five years to calculate and distribute to ASCAP statutory cable and satellite royalties pursuant to Section 111 and 119 of the Copyright Act. Similarly, the statutory rate for blanket mechanical phonorecords licenses pursuant to Section 115 for the period from 2018-2022 were not finalized until August of 2023. Likewise, legislated statutory licensing schemes cannot keep pace with simple technological developments, let along the monumental evolution we expect to occur with AI technologies. The victims of compulsory licensing regimes are always the songwriters and the composers.

ASCAP has consistently supported and advocated for a free market with willing buyer and willing seller standards for pricing without government intervention.

The Department of Justice has also openly opposed the expansion of the compulsory licensing in the music industry. As remarked by then Assistant Attorney General for Antitrust, Makan Delrahim in a speech on the future of ASCAP/BMI Consent Decrees: "Too often, however, it has been creators—songwriters, artists, and other rightsholders—who have received the short end of the stick under compulsory licensing, necessitating reforms like the recent Music Modernization Act, by Congress. Compulsory licensing also runs counter to the principles that form the very foundation of the free market and rights in intellectual property. Those principles hold that the best, most efficient way to allocate resources—and the most effective way to maximize consumer welfare—is through allowing parties to negotiate, to set prices based on supply demand, and available information."[37]

Generative AI is a nascent and fast-moving field. A voluntary collective licensing scheme that relies on market forces can best adapt to the evolving needs of copyright holders and AI developers alike as the technology continues to develop. A compulsory licensing scheme, on the other hand, simply cannot keep up with the pace, and risks hurting copyright holders as well as hampering the development of generative AI. Accordingly, compulsory licensing, often a solution to a broken licensing marketplace is wholly unnecessary with respect to performance of musical compositions.

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

[37] *Remarks by Assistant Attorney General Makan Delrahim on the Future of ASCAP and BMI Consent Decrees*, U.S. DEP'T OF JUST. OFF. OF PUB. AFFS. (Jan. 15, 2021), https://www.justice.gov/opa/speech/remarks-assistant-attorney-general-makan-delrahim-future-ascap-and-bmi-consent-decrees.

opt out? How should royalty rates and terms be set, allocated, reported and distributed?

As discussed in our responses to Questions 6.2, 9, 9.3, and 10-10.2, a voluntary collective licensing scheme is the best approach for solving many copyright problems posed by generative AI. We strongly oppose a government-mandated licensing scheme, which would eliminate consent and would be prone to inaccuracies and inefficiencies, and would undercompensate creators. As the story of compulsory licensing in the music sector has shown, the system is rigid, slow-moving, and hurts creators by significantly devaluing music and delaying payments. This is particularly problematic for a new technology like generative AI, which is becoming increasingly sophisticated all the time. It has barely been a year since ChatGPT was first released, and generative AI has already taken hold in every sector of the economy. By contrast, the Copyright Royalty Board routinely takes years to calculate and distribute to ASCAP statutory cable and satellite royalties pursuant to Sections 111 and 119 of the Copyright Act. A compulsory licensing will not be able to keep up with the pace of development of generative AI, and may end up hurting both copyright holders and AI developers alike.

Instead, a voluntary collective licensing scheme is ideal because PROs like ASCAP already have the requisite experience and infrastructure in place to support voluntary collective licensing arrangements, including the technical capability, assisted by AI, to track trillions of performances across all types of media. Moreover, since it relies on market forces to efficiently allocate resources, it can better adapt to the evolving needs of copyright holders and AI developers.

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

As discussed in our responses to Questions 6.2, 9, 9.3, and 10-10.3, the current PRO collective licensing system is best suited to support licensing for AI platforms. We oppose an

extended collective licensing scheme as it is similar to a compulsory licensing scheme and has many of the same problems.

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

Effective collective licensing of public performances of musical works requires a careful case-by-case consideration of key factors associated with each type of use, such as type of media, scope of territory and addressable market, music intensity, performance frequency, and applicable monetization scheme. Therefore, licensing regimes have always varied based upon the type of work and right being licensed. The applicable factors in the generative AI context are best addressed with voluntary licenses negotiated between copyright holders and AI platforms.

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)?

It depends on the specific rights to be licensed. With respect to public performance rights of musical works, we believe that whoever effectuates the performance of the works should obtain a license form the relevant PRO. This can include both the developer that trains the AI model and the company that hosts the model on its server.

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.

Appropriate algorithms may be developed to trace the sources of a particular output. With respect to ChatGPT, for example, studies have shown that it is possible to construct carefully engineered prompts to make ChatGPT complete paragraphs from copyrighted books.[38] Bing Chat,

[38] *See GPT-4 Memorizes Contents Of Copyrighted Books, And It Could Be A Cultural Issue*, THE DECODER (May 4, 2023), https://the-decoder.com/gpt-4-memorizes-contents-of-copyrighted-books/.

which is powered by a Microsoft proprietary AI model that leverages the powers of both GPT-4 and the Bing search engine, can also cite to certain website sources in its narrative answers. However, the expertise required to develop such algorithms resides with the AI platforms, and not with the Copyright Office, judges, legislators, or copyright holders.

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

The more pressing question is: what is the economic impact on human creators when licensing requirements are not respected by AI platforms and users? This impact is clearly significant and even existential for music creators. Based on our experience and conversations with our members and other music creators, most music creators believe that generative AI is a threat to their livelihood. When technology is used in a way that takes aim at human creators, ignores their rights, or devalues their work, that is not innovation. With a voluntary licensing scheme in place, AI can augment but not displace human creativity in music creation. In fact, all other digital technology sectors have been growing under voluntary licensing arrangements. A collective blanket license scheme similar to the one currently offered by ASCAP is well suited to generative AI as it offers a blanket license for the right to perform more than 18 million works in the ASCAP repertory without having to negotiate a deal with every single creator.[39] This well-tested blanket licensing scheme has successfully protected the rights of creators without hampering any technological progress over the past century.

Transparency & Recordkeeping

15. In order to allow copyright owners to determine whether their works have been used, should developers of AI models be required to collect, retain, and disclose records regarding the materials used to train their models? Should creators of training datasets have a similar obligation?

[39] *Blanket License,* ASCAP, https://www.ascap.com/help/ascap-licensing (last accessed October 30, 2023)

It is imperative that AI developers collect, retain, and make available information concerning all copyrighted material used in any stage of the development of AI tools. Retaining such records is critical to ensuring that the rights to use these materials have been duly obtained—like any other industry, the AI industry is responsible for identifying and obtaining the rights necessary to run its business.

Accordingly, AI developers should, without limitation, retain information on (1) all copyrighted material present in any datasets compiled by the developer or obtained from a third party; (2) all copyright material actually ingested as training input into any particular model; (3) for each piece of utilized copyrighted material, the particular training use for which that material was employed, and whether it was used as training, testing, and/or or validation data.

Creators of training datasets should likewise have an obligation to collect, retain, and disclose information concerning all copyrighted material contained in any training dataset. Creators of datasets should disclose information concerning copyrighted material to the AI developers with whom they share those datasets, without thereby limiting any liability of the AI developer to collect, retain, and disclose this information.

The burden to monitor the use of copyrighted content must lie with AI developers, who are best situated to know whether and to what degree a given musical work contains a copyrighted component or can be traced back to copyrighted training material. In particular (as described in our answer to Question 7 above) while technology exists that can, to a limited degree, identify input musical works by analyzing AI-generated output, this process is extremely costly and time-consuming for copyright holders, and the vast majority of the data used to train a particular AI model is unidentifiable from output alone.

15.1 What level of specificity should be required?

Where available, information concerning the use of musical works must include all standard metadata identifiers. These standard identifiers may include, without limitation, artist name, producer, label, songwriter, song title, publisher and release date. For musical works or sound recordings for which metadata is unavailable, the song title and artist (and if available, songwriter) must be included at minimum.

15.2 To whom should disclosures be made?

Because AI developers and training dataset creators must obtain licenses before using copyrighted content in their datasets or as training input into AI models, disclosure will ordinarily be made to copyright holders prior to use in a dataset or as training input. In addition, disclosures must be made to all relevant copyright holders upon request.

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

AI developers should not be permitted to use third-party models as a shield against liability for the unlicensed use of protected content. Like the third parties that develop the relevant models, AI developers incorporating such third-party models have an obligation to ensure that all copyrighted materials used in such models are properly licensed.

In addition, as stated in our answer to Question 15, both AI developers and third-party model developers must collect and retain information concerning all use of copyrighted materials in such third-party models.

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?

ASCAP does not have access to the information needed to answer this question. However, it is implausible that such recordkeeping would be either technologically or practically infeasible. Digital streaming services like Spotify and Apple Music are able to maintain sufficient data to

enable PROs like ASCAP to identify the use of protected content and compensate their members accordingly. There is no reason why AI developers and platform owners could not employ analogous measures to track the copyrighted content used in the development of their tools.

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

Because the rights to utilize copyrighted materials in either data collection or for use as training input must be obtained in advance, copyright owners should ordinarily have been notified ex ante of the use of their works. As discussed in our response to Question 9.4, we think there should be an affirmative requirement for AI platforms to notify copyright holders of the use of their copyrighted works in generative AI training before the data is ingested into the generative AI model. Of course, a blanket license would obviate the need to do this.

Generative AI Outputs

(If your comment applies only to a particular subset of generative AI technologies, please make that clear.)

ASCAP's comments below apply to any generative AI technology that produces audible sound (including songs and videos with an audio component) or content encoding or otherwise representing audible sound (*e.g.*, written music scores, waveforms, spectrograms).

Copyrightability

18. Under copyright law, are there circumstances when a human using a generative AI system should be considered the ''author'' of material produced by the system? If so, what factors are relevant to that determination? For example, is selecting what material an AI model is trained on and/or providing an iterative series of text commands or prompts sufficient to claim authorship of the resulting output?

Whether material produced using a generative AI system is copyrightable is a factual question dependent on the degree of human involvement. Where a human's involvement is limited to the simple generation of minimal queries and prompts for an AI tool, the resulting material is

not entitled to copyright protection. However, where a human significantly edits, manipulates, utilizes a sophisticated and material series of queries and prompts, or alters the output generated by an AI tool, the resulting material should be subject to copyright protections to the extent that it reflects human creativity. This is a fact-specific inquiry that must be determined on a case-by-case basis consistent with existing copyright law and applicable precedent.

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?

ASCAP believes that no revisions or clarifications are necessary to address the human authorship requirement, *i.e.*, that original creative human input is necessary. The extent to which the aid of an AI tool in human creation of an original work deprives the work of copyright protection is not as clear. At the very minimum, there needs to be clarification that insignificant use of an AI tool that is otherwise substantially created by a human does not deprive it of copyright protection. However, given the large grey area, greater clarification will be necessary of the extent of the human input required to qualify for 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?

Broad legal protections for purely AI-generated works is contrary to the purpose of copyright law, which is intended to promote creative art by human authors. In particular, broad protection for purely machine-generated works could have an adverse effect on future human creation, including by diverting royalty streams away from human creators and displacing human creators from the market. Such displacement is especially inequitable given that these tools owe

their functionality to the corpus of original human works that they exploit as training data—often without consent, credit, or compensation.

As described in our answer to Question 18, ASCAP submits that, like any potentially copyrightable work, works generated with the use of AI tools should be subject to copyright protection to the extent that they reflect human creativity. This question requires a fact-specific inquiry that must be performed in accordance with existing copyright law and applicable precedent.

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?

The framers of our Constitution sought to advance a robust collection of U.S. works made through the efforts of human creators by providing exclusive rights to those human authors. Granting broad legal protection to purely AI-generated material hinders that goal insofar as it diverts royalty streams away from human creators and displaces them from the market.

However, as described in our answer to Question 18, ASCAP submits that, like any potentially copyrightable work, works generated with the use of AI tools should be subject to copyright protection to the extent that they reflect human creativity. This question, too, requires a fact-specific inquiry that must be performed on a case-by-case basis in accordance with existing copyright law and applicable precedent.

Infringement

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

Since generative AI models operate by recalling knowledge of patterns it retained during the training process, with the right prompts, generative AI tools can and do generate outputs that are substantially similar or derivative of parts or the entirety of their copyrighted training materials,

and therefore infringe on those copyrighted materials. Certain music generative AI tools can also stream or broadcast AI-generated music, and therefore, potentially violate the performance rights of the creators of the original musical works.

23. Is the substantial similarity test adequate to address claims of infringement based on outputs from a generative AI system, or is some other standard appropriate or necessary?

The substantial similarity test may be adequate to address claims of infringement based on outputs from a generative AI model, particularly if the outputs are analyzed on an element-by-element or part-by-part basis. The adequacy of this test will depend on judicial interpretations. However, as discussed, the outputs from a generative AI model can also indicate the inclusion of certain copyrighted content in the training dataset. In the absence of transparency, reasonable inferences about the use of copyrighted materials should be legally sufficient to support separate claims of copyright infringement.

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

As discussed, some generative AI providers have disclosed their use of some existing third-party training datasets such as the Common Crawl. In those cases, copyright holders can analyze those datasets to prove access to the relevant copyrighted materials with information already in their possession. Copyright holders may also infer the existence of certain materials in the training dataset based on the outputs of the model. However, the methods of inferring training data are costly and tedious, and do not shed light on the full scope of the copyrighted materials used in training. A licensing system as discussed above would incentivize AI platforms to keep records on their training materials and facilitate collaboration among the stakeholders to develop an efficient system for keeping track of the use of those materials.

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?

The persons responsible for ingesting the original copyrighted work into the AI model should be liable for the infringing output. For example, if the copyrighted work was found to be included in the training dataset for the model, then the developer of the model would be directly liable for copyright infringement. In addition, if the copyrighted work was fed by way of a prompt by the user, then the user would be directly liable for that particular use of the copyrighted work, while the developer or the AI platform would, at a minimum, be secondarily liable in that instance.

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

Generative AI models that are offered to users for free should be subject to the same considerations with respect to output as subscription models. The open-source practice does not change the nature of generative AI models as tools of replacement. Open-source models are equally capable of generating infringing content and content that replaces the need to access the original copyrighted materials. Moreover, open-source models are often used as means to collect customer feedback or attract users for subscription models.

Labeling or Identification

28. Should the law require AI- generated material to be labeled or otherwise publicly identified as being generated by AI? If so, in what context should the requirement apply and how should it work?

AI-generated musical works should be labeled as AI-generated, and should provide credit to the creators whose musical works are utilized in such generation. In particular, credit should be given to any works or songwriter specifically referenced in the creation of the AI-generated work (*e.g.*, any songwriter or musical works referenced in any prompts leading to the creation of the

work in question), and a reference should be provided to the larger database of copyrighted works on which the AI tool was trained.

A failure to publicly identify material as being AI-generated and to credit utilized protected musical works — especially where the output incorporates a songwriter's lyrics and/or musical sounds, or is otherwise similar to existing musical works—can implicate rights under state law regimes including, without limitation, unfair competition. The requirement to disclose AI use and credit underlying works could also be incorporated into the initial licensing agreement entered into by copyright owners and AI tool developers and/or database compilers.

To the extent an author or rights holder wishes to not be associated (through labeling or other identification method) with an AI-generated work that was created via an AI system trained on that person's work, we would support a methodology to implement and administer such an option, if one is feasible.

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

The burden to accurately identify and disclose the use of AI-generated content must lie with AI developers, who are best situated to know whether and to what degree a given musical work contains an AI-generated component. It is infeasible to leave the identification of AI-generated works to downstream distributors (such as streaming services) or copyright holders (including licensors such as ASCAP) who are poorly situated to identify the use of AI-generated tools. Nor should the requirement to label a piece as AI-generated lie with the human user who facilitates such generation, who may lack the knowledge or ability to make the requisite identification. Finally, as described further in our answer to Question 29 below, while technology exists that can, to a limited degree, identify the use of AI-generated tools in musical works, this technology is imperfect and subject to obsolescence as AI tools become increasingly advanced.

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

ASCAP acknowledges that there is a spectrum of potential AI involvement in the creation of new musical works, ranging from no AI involvement at all to fully AI-generated works. While there may be difficult questions, at the margins, of determining the degree of AI involvement in the creation of a particular work, all musical works generated using AI-based tools should be subject to the labeling requirement. As described in our answer to Question 28.1, the labeling requirement should apply to both copyrighted material specifically referenced in the creation of the work in question (*e.g.*, as used in prompts), and to copyrighted material generally used as training data ingested into the model.

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?

A failure to publicly identify material as being AI-generated—especially where the output incorporates a protected musical work, or is otherwise similar to existing musical works through a query or prompt using the author's name—can create liability under state law regimes including, without limitation, unfair competition. Insofar as the requirement to label underlying works is incorporated into licensing agreements between copyright owners and AI developers, failure to label would also give rise to claims for breach of those licensing agreements.

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?

As described below, while there exist tools that can identify AI-generated materials in certain circumstances, these tools are limited in their accuracy and scalability. For instance, an AI detection tool introduced in 2023 by OpenAI was rolled back shortly thereafter due to its low

accuracy.[40] This makes it all the more important to place the burden of identifying and labeling uses of generative AI on AI developers, who are best positioned to identify the presence and degree of AI use in the works produced using their technologies.

- Identifying new instances of known AI works: automated content recognition technology can be used to identify new uses (*e.g.*, reuploads) of existing works known to be AI-generated, by identifying certain digital features of known AI-generated works and comparing them against a body of content. While these tools can detect the spread of known AI tracks, they do not distinguish AI-generated tracks from those generated by humans.

- Identifying alterations of existing works: automated content recognition technology can also be used to identify common alterations to existing works made possible by AI tools, such as modifications to pitch or speed, or swapping the vocals in an existing track for those of a different artist. Again, these tools do not directly identify AI-generated content, but infer the likely use of AI tools by identifying conspicuous similarities and differences with known tracks.

- Directly identifying AI-generated works: while there are some tools in existence and development that aim to directly detect AI-generated content by identifying common features of AI-generated works (*e.g.*, certain repetitions or recurrent patterns), these tools have significant limitations. First, developing research[41] suggests that they are easily thwarted by relatively simple manual modifications to the AI-generated content. Second, they are quickly rendered obsolete by continuing improvements to AI technology that eliminate common "tells" of AI content and enable increasingly lifelike AI-generated works. The degree to which robust AI detection is possible in the context of musical works remains an evolving technical question.

Additional Questions About Issues Related to Copyright

30. What legal rights, if any, currently apply to AI-generated material that features the name or likeness, including vocal likeness, of a particular person?

31. Should Congress establish a new federal right, similar to state law rights of publicity, that would apply to AI- generated material? If so,

[40] *Openai Shuts Down Tool To Detect AI-Written Text Due To Low Accuracy,* BUSINESS TODAY (Jul. 26, 2023) https://www.businesstoday.in/technology/news/story/openai-shuts-down-tool-to-detect-ai-written-text-due-to-low-accuracy-391269-2023-07-26.

[41] Rhiannon Williams, *AI-Text Detection Tools Are Really Easy To Fool*, MIT TECH. REV., (July 7, 2023), https://www.technologyreview.com/2023/07/07/1075982/ai-text-detection-tools-are-really-easy-to-fool?gad=1&gclid=CjwKCAjw6p-oBhAYEiwAgg2PgsPcNJuWRTS94GOh65RMWaOP2tf5tav9iW6N6YJPBndTqc6GCaDERRoC1SEQAvD_BwE.

should it preempt state laws or set a ceiling or floor for state law protections? What should be the contours of such a right?

32. Are there or should there be protections against an AI system generating outputs that imitate the artistic style of a human creator (such as an AI system producing visual works ''in the style of'' a specific artist)? Who should be eligible for such protection? What form should it take?

The below answer responds to Questions 30, 31, and 32.

Yes, the output of AI tools should be subject to the existing copyright regime and other legal frameworks, and all human creators should be eligible for such protection. To the extent an output is substantially similar to protected content or otherwise implicates exclusive rights set forth in Section 106 of the Copyright Act, that output should be subject to federal copyright law. Likewise, to the extent the output appropriates the artist's creative work or his or her name, voice, or likeness, the output should be subject to existing legal frameworks including, without limitation, those concerning publicity rights, unfair competition, and unjust enrichment.

Further, as described in our answer to Question 5, a new federal right of publicity is necessary to adequately protect the right of publicity in the context of increasingly powerful voice, melody and lyric cloning possible through generative AI systems (for example by prompting or querying an AI system to create a song in the style of a specific named songwriter).

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

We have not identified any additional issues at this time but would like to reserve the right to raise in reply to comments submitted by others.

Dated: October 30, 2023

Respectfully submitted,

The American Society of Composers, Authors and Publishers
250 West 57th Street, 14th Floor
New York, NY 10107

Clara Kim
Executive Vice President, Chief Legal & Business Affairs Officer