Before the
UNITED STATES COPYRIGHT OFFICE
Library of Congress

Notice of Inquiry

Artificial Intelligence and Copyright: Notice and Request for Public Comment

Docket No. 2023-6

COMMENTS OF UNIVERSAL MUSIC GROUP

Jeffrey Harleston
General Counsel and EVP, Business & Legal Affairs
Alasdair McMullan
Head of Litigation and EVP, Business and Legal Affairs
Universal Music Group
2220 Colorado Avenue
Santa Monica, California 90404

David Kokakis
Chief Counsel
Universal Music Publishing Group
2105 Colorado Avenue
Santa Monica, California 90404

Universal Music Group ("UMG") respectfully submits these comments in response to the Copyright Office's Notice of Inquiry ("NOI") concerning artificial intelligence and copyright.

UMG is the world's leading music company.[1] Among its record labels are such iconic imprints as Blue Note, Capitol, Def Jam, Deutsche Grammophon, Interscope, Island, Motown, Republic, UMG Nashville, Universal Music Latin Entertainment, and Virgin. UMG owns the most extensive catalog of recordings in the industry, covering the last hundred years of many of the world's most popular artists. UMG is also home to Universal Music Publishing Group ("UMPG"), one of the industry's premier music publishing operations that publishes and administers some of the most important musical compositions in modern history. Collectively, UMG owns or controls a catalog of sound recordings and musical compositions of incalculable artistic, cultural, and economic value.

UMG has historically been at the forefront of confronting new technologies, embracing innovations that serve the artistic community and curtailing those aspects of new technologies that undermine creative effort and output. UMG has helped develop the protocols and market-based solutions for sea changes in the technological consumption of music, such as streaming. UMG believes it can play a similar role with generative AI ("AI"), which offers tremendous opportunity but poses extraordinary risks to the creative community, if used irresponsibly, illegally, or without respect for the creative community.

Copyright is the key protection that will allow UMG to take advantage of the opportunities that AI presents and ensure that our artists, songwriters and creators are rewarded

[1] www.universalmusic.com

for their work. It is the foundation upon which our industry is built, and strong copyright protections are vitally important to ensure creativity continues to be incentivized and rewarded.

While this Notice of Inquiry covers a great deal of important ground and we look forward to a continuing dialogue with the Copyright Office and other policymakers on all of the issues it raises, we believe that there are two fundamental principles that must be maintained. First, copyright protection can and should only extend to expressive content created by human beings. This principle is fundamental to the constitutional underpinning of copyright law. To be sure, technology can assist human authorship, but it should not and cannot replace it. Second, use of copyrighted content by AI, whether for training or other purposes, requires permission from the copyright owners, via negotiated licenses. The entire music industry operates in this fashion and on a scale that few appreciate. UMG has an extraordinary record of success in developing bespoke free-market business partnerships with hundreds of digital partners. These relationships respect the rights of artists and songwriters, while fostering innovation and unparalleled consumer experiences and access.

We welcome the opportunity to offer our views and perspective on the important role of copyright in cultivating the potential benefits and addressing the inherent challenges presented by generative AI.

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?

Like many historically disruptive technologies, generative AI offers both great promise and poses enormous risk for the music industry. When it assists and promotes the human creative process, generative AI can be a potent tool the music industry should embrace and cultivate in collaboration with the AI development community. Indeed, as described below, UMG is already exploring the benefits this new technology offers to artists, songwriters and the industry that supports them. But in its nascent state of explosive growth and seemingly limitless copying and exploitation of millions of works of authorship without license, generative AI imposes potentially existential artistic, cultural, and legal threats to the music industry. We believe these risks can be managed, particularly if the fundamental tenets of copyright law are applied and responsible, fair "rules of the road" establishing AI trustworthiness are established and heeded. Companies like UMG have a wealth of experience in creating sensible, lawful, and mutually beneficial licensing models for the use of music. The same path forward is available in a world of artificial intelligence, but the effort must start now, with thoughtful application of existing law and, where appropriate, greater legal clarity on the obligations of those who create, market, and distribute generative AI. We set forth below both the benefits we are seeing and foresee for generative AI, followed by our real-world assessment of the risks generative AI poses for the music community.

A. Benefits

UMG has benefited from the use of AI for several years and continues to explore its potential at all levels of the production and creative process of making music. The unifying principle underlying these efforts is that AI should assist human creativity (not supplant it) and afford content owners and creators control over how their artistic works are used by generative AI.

For example, AI has long played a role in the recording studio. Apple Logic Pro X generates drum tracks, Captain Plugins generates chord progressions, and AI assists in creating Dolby Atmos mixes to provide an immersive audio music experience. UMG employed an AI technology to remix and restore audio content for a Beatles documentary ("The Beatles: Get Back"). These technologies enhance the technical recording and production process that is driven, in the first instance, by human creativity.

UMG artists and producers likewise use AI technology not only to improve recordings but also reconfigure and reimagine them in creative ways. For example, AI facilitated the creation of a new sound recording of a previously unreleased Beatles song ("Now and Then") that laid dormant in fragmented demo format. Peter Jackson's audio team used machine assisted learning (AI) to extract John Lennon's clean vocal from a demo recording. That vocal track was used in combination with George Harrison's acoustic and electric guitar performances from a later recording and with new performances by Paul McCartney and Ringo Starr. This allowed the group to create a complete, integrated recording of a composition that the four members of the band were never able to record all together.[2] In another instance, one of UMG's distributed

[2] See https://www.thebeatles.com/announcement

artists used a generative AI system named "Supertone"[3] to train on recordings of his own vocals. Using the technology, the artist was able to release a new single in six languages, in his own voice, on the same day. In both instances, ethically managed AI, used with the authorization and at the direction of the artists and rightsholders, provided critical technology to realize a human artistic vision.

The potential for these kinds of applications is unlimited. For example, artists, songwriters, and producers could create personalized recordings for fans, stream live performances in dozens of languages simultaneously, or pursue previously unattainable artistic collaborations. And concomitant with those efforts, AI also can serve as a powerful tool for recording artists and songwriters to expand their reach. One UMG company, Ingrooves, owns three patents in AI to assist with marketing independent artists. AI likewise can help songwriters pitch new songs to recording artists, build an artist's tour with greater efficiency, identify new potential fan bases, and monetize music in unprecedented ways. We applaud applications of AI that support artists in this fashion.

More broadly, we are working with generative AI developers to explore how their technology can promote and enhance the creative process. For example, UMG has partnered with Endel,[4] a generative AI tool that derives functional soundscapes to facilitate sleep, focus and meditation, driven by scientific insights into how music affects our mind-state. The Endel relationship will allow artists to create new or derivative recordings built from their own stems or sound recordings, all subject to the artists' control and approval. We've also recently

[3] https://supertone.ai
[4] https://endel.io/

announced an expansive, industry-first strategic relationship with BandLab,[5] the world's largest social music creation platform, to foster market-led solutions with pro-creator standards to ensure generative AI technologies serve the creative community effectively and ethically.

And at the highest level, UMG recently announced a collaboration with YouTube to experiment with generative AI tools that are predicated on human creativity and take into account the important interests of creators and copyright holders.[6] This arrangement combines the resources of UMG with YouTube, a leading music platform that hosts a staggering quantity of musical content and serves as a launchpad and platform for both new and emerging artists and established stars. YouTube and the music industry have a track record of confronting disruptive technologies that imperil intellectual property rights and finding market-based solutions. For example, when YouTube's launch prompted a flood of user generated content that jeopardized copyrights at an unprecedented level, YouTube and the music industry collaborated to create the first legal framework and compensation system, Content ID, that allowed rightsholders to control and monetize the use of their creative output by users. The arrangement spurred a cottage industry of content identification solutions (AudibleMagic, ACR Cloud, etc.) that are made available to companies who cannot afford the development of their own solutions.

The UMG/YouTube collaboration incorporates two key components. First, YouTube has committed to a set of AI principles to responsibly (a) develop appropriate protections and safeguards for rights owners, (b) acknowledge the need for creators and rightsowners to control the use of their creative works in generative AI; and (c) build AI tools and develop content

[5] https://www.bandlab.com/
[6] *YouTube and Universal Music Group Team Up to Experiment with AI*, Los Angeles Times, August 21, 2023

policies predicated on trust, safety, and the lawful and authorized use of intellectual property rights. Second, UMG and YouTube are jointly launching the Music AI Incubator, which has brought together a working group of leading UMG artists, songwriters and producers, Google DeepMind technologists, and YouTube and UMG experts to explore, experiment, and offer feedback on AI-related musical tools and products. Rather than simply respond *post hoc* to technologies that have already been released, the premise of this experiment is to influence the development of AI tools such that they serve recording artists and songwriters and the industry in which they operate *from inception*. This groundbreaking model – participation of the creators and rightsholders from the ground up – is the best way to explore and identify possible pathways to responsible AI technology used in service of artists, the music ecosystem, technology interests and consumers alike.

B. Risks

As noted above, generative AI that collaborates with and obtains input and licenses from content creators and owners can be a remarkable boon for the music ecosystem, because it supports human creativity. In contrast, generative AI that simply steals the fruits of human creativity to build multi-billion-dollar businesses and generate hollow substitutes for genuine human creative expression is the antithesis of what the law does and should allow. In those latter uses, it is simply a tool of infringement, and a powerfully vast and uncontrolled one at that. Popular LLMs can produce infringing lyrics in seconds, text-to-image AI systems can generate visual works that infringe copyrighted artwork, and voice cloning AI models can learn to mimic a specific person's voice using *only 3 seconds* of source audio.[7]

[7] *Microsoft's New AI Can Simulate Anyone's Voice with 3 Seconds of Audio*, Benj Edwards, Arstechnica, January 1, 2023

In response to other questions below, we will address more specifically the unlawful use of intellectual property in the development and implementation of generative AI. For present purposes, UMG notes the high-level, indisputable violations built into unlicensed AI models to spotlight the deleterious effects of such systems. Broadly, many generative AI models violate intellectual property rights at a number of levels:

- They train on vast troves of copyrighted materials scraped from the internet and other sources without the necessary licenses and permissions required to exploit those materials;

- They reproduce, manipulate, and process those copyrighted materials to perfect their systems and attract customers, justify subscription fees, and raise investment capital, again without license or permission from and compensation to the owners of those materials;

- They generate outputs, prompted by users, that are substantially and sometimes strikingly similar to copyrighted works, such as word-for-word copies of famous song lyrics; and

- They clone artists' images, likenesses, and voices to create audio and visual works that are disturbingly similar to the artists themselves, in violation of rights of publicity and laws prohibiting unfair competition and consumer deception.

The effects of these infringements and violations are far-reaching for artists, songwriters, the music industry, the creative community, and the U.S. economy.

Consider the training stages of generative AI. The music industry is predicated on the production and *licensing* of copyrighted musical and related content. Record labels and music publishers license copyrighted music for countless artistic and commercial uses, including physical recordings, streaming services, music videos, film and television scores, fitness services, social media, user-generated content, video games, smart phone apps, music in commercials and advertising, samples of sound recordings used in other sound recordings, interpolations of musical compositions used in other compositions, consumer goods ranging

from children's toys to greeting cards - even FDA-approved medical devices.[8] These bespoke arrangements not only span a vast range of consumer offerings, but they've been negotiated with an array of entities – from fledgling start-ups to some of the biggest companies in the world. In a free-market, licensing is always evolving as media matures and evolves technologically – witness the transition from CDs to downloads to streaming. The flexibility inherent in free market negotiations has allowed for nimble and bespoke arrangements spurring innovative new consumer offerings. The use of copyrighted materials for AI training falls squarely within this category of a newly developed market for licensing musical content.

Sadly, a significant portion of the AI industry has chosen to ignore its legal obligation to obtain the licenses necessary to reproduce and exploit our copyrighted content for training purposes, even though these training materials are the essential foundation upon which the AI industry is building giant commercial enterprises. No other business that is so dependent upon copyrighted materials fails to even seek, let alone obtain valid licenses. And when AI ingests massive quantities of musical or lyrical content indiscriminately and without license, it has an immediate and devastating impact on artists, songwriters and the companies that support them. It drastically interferes with UMG's ability to license musical content and generate income that both directly supports artists and their creative efforts, as well as the infrastructure that produces and markets the fruits of those efforts. It diminishes our ability to invest in and support new artists, leading to fewer opportunities for artists and songwriters, and fewer jobs for producers, studios, instrument and equipment manufacturers, musicians, managers, stylists, directors,

[8] See press release: https://www.universalmusic.com/medrhythms-partners-with-universal-music-group-for-pioneering-prescription-music-platform-to-treat-neurologic-injury-and-disease/

graphic designers, choreographers, video production crews, caterers, and the entire music ecosystem that supports the industry. Fewer artists result in fewer concert tours supporting fewer venues – with a corresponding economic impact on the businesses around those venues (hotels, retail, restaurants, transportation, etc.).[9] Market rate, negotiated licenses of the type prevalent in all other sectors of the music industry would address these harms, but the generative AI community is largely oblivious to – or willfully disregards – their obligation and the need to obtain those licenses.

These harms, of course, flow through to the output AI generates on the backs of the copyrighted content we own and control. Different types of outputs cause different kinds of damage. At one end of the spectrum, generative AI is increasingly capable of producing outputs that closely resemble the works on which it was trained. For example, it is a simple matter to reproduce exact replicas of copyrighted song lyrics with a prompt asking for them or even asking about certain subjects that copyrighted lyrics cover (*e.g.*, a prompt requesting a song about the death of Buddy Holly returns the lyrics to "American Pie"). Such strikingly similar outputs are without question infringing, but they also threaten to replace the legitimate, established market for licensing authorized lyrics to a variety of online services, which deprives songwriters and publishers of important income. Generative AI is also capable of creating high-quality deep fakes or voice clones.[10] Beyond violating the rights of publicity of the mimicked artists, these imitations confuse consumers and compete unfairly against human-authored copyrighted works

[9] See https://www.oxfordeconomics.com/resource/livemusic/ or https://www.ustravel.org/news/taylor-swift-impact-5-months-and-5-billion

[10] *AI voice clones mimic politicians and celebrities, reshaping reality*, Pranshu Verma and Will Oremus, *Washington Post*, October 15, 2023 and *Major Record Companies Hate Ai Voice-Cloning Platforms That Don't Pay. The One They Hate Most Was Created By A 20-Year-Old Uk Student*, Daniel Tencer, Music Business Worldwide, October 12, 2023

and the artists and companies that create or support them, potentially causing irreparable damage to an artist's reputation, brand, and livelihood. Imagine a believable clone of a famous artist singing or saying something defamatory, offensive, or philosophically or politically at odds with that artist's own convictions.

It is hard to overstate the scope of these problems. On a daily basis, UMG reviews a constantly growing flood of AI-generated works that infringe our intellectual property or our artists' rights in their name, image, likeness or voice. Since August alone, the number of identified AI generated uploads to UGC platforms implicating our rights has grown 175%. Roughly 47% of the notices we've sent thus far seeking removal of violative AI content were triggered because a UMG sound recording was detectible in the underlying vocal or instrumental; the remaining were violative of a musical work copyright, trademark, or a right of publicity. As of this writing, we face a backlog of over 10,000 URLs to be individually reviewed and processed.

As a global company, we have had to develop the expertise to engage in protecting our and our artists' rights in this way – as well as through litigation. However: (1) we are effectively using a thimble to empty an ever-growing ocean of infringements, and (2) most creators and music businesses do not (and cannot realistically be expected to) engage in such an expensive and time-consuming effort. The burden on rightsholders is substantial while the ease of AI generation in an age of no guardrails continually increases - and at an astounding rate.

Moreover, there are parallel and equally threatening harms even when the output does not resemble the protected materials on which the AI was trained. Approximately 120,000 new recordings are uploaded to streaming services daily, and if that rate continues "over 43 million new tracks will have been uploaded to Spotify and other music streaming services this

year."[11] As AI-generated music becomes increasingly easy to create, it saturates this already dense marketplace, competing unfairly with genuine human artistry, distorting digital platform algorithms and driving "cheap content oversupply" - generic content diluting human creators' royalties. Consumers may in turn be misled into streaming or patronizing computer-generated music they believe was created by a real person.

This generic content, increasingly the result of generative AI, also robs creators of their rights of integrity and control. They never get to choose whether or not their work is used to create new works. And those works may compete with genuine, human-created music and/or contain material that is defamatory, violent, racist, or offensive to the artist's beliefs. In a sense, artists and songwriters are made to become unwitting participants in undermining their own market or creating content that does violence to their beliefs, artistic aesthetics, and reputation.

The harms we identify above extend far beyond the music industry. In 2021, the value added by the total copyright industries to the GDP, of which we are a part, exceeded $2.9 trillion, accounting for 12.52% of the U.S. economy. That same year, the total copyright industries employed nearly 16.1 million workers, accounting for 8.14% of all U.S. employment. The vast, unregulated capture and exploitation of musical content by generative AI imperils these sources of revenue and employment for the whole economy. Music is also one of the US's great cultural exports, so cheapening and devaluing it with imitations or generically competitive music takes a tremendous artistic and cultural toll on the U.S. as well.

[11] *There are 120,000 New Tracks Hitting Streaming Services Each Day*, Murry Stassen, Music Business Worldwide, May 25, 2023

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?

The music industry faces particularly acute threats from generative AI for a number of reasons. First, our industry encompasses multiple types of intellectual property used across a wide spectrum of contexts. UMG owns or manages rights in copyrighted sound recordings and musical compositions, visual works of art in the form of album artwork, and audiovisual works in the form of music videos, documentaries, music-themed scripted television and movies, and other performances. Our artists own rights of publicity in their names, likenesses, voices, and personas, and we are often granted the right to monetize and enforce those rights on their behalf. We own numerous trademarks for our own products and services; we own or serve as the exclusive licensee for various trademarks and rights of publicity; and we exercise those rights for use in the creation of music merchandise or other goods and materials. We exercise significant care and strategy to determine how these properties are experienced by consumers, released, and marketed. Generative AI currently robs UMG, our artists, and songwriters, and ultimately the entire creative musical community of that control.

Second, as described above, the music business is a licensing business. For example, UMPG's and other music publishers' primary income stream derives from licensing musical compositions and lyrics to third parties for myriad uses, ranging from commercial and advertising purposes, to independent documentaries, to interpolations in songs controlled by other companies, to licensed lyrics services. Generative AI bypasses that licensing model and ingests our content wholesale, without license or permission. That massive appropriation of properties we license threatens our core economic models.

Third, music is widely and freely available in ways that other expressive works are not. If a consumer wants to read a book, she purchases a hard copy or downloads it to her Kindle. If that same consumer wants to hear a particular recording, she need only go to Spotify or YouTube or any number of authorized streaming sites. The easy access to music and its sheer ubiquity on the internet makes it acutely vulnerable to scraping by unscrupulous AI models, leading to the harms we describe above. While licensed music services may use various forms of technology to protect their services from unauthorized access, web scrapers often bypass this technology in order to obtain our music.

And finally, the nature of how music is consumed exacerbates the risks of generative AI. In a few short years, our industry has transitioned from selling physical copies of music, to a download model, to the now-dominant streaming model. We have navigated these fundamental changes successfully and developed fair and sensible protocols for how these technologies can serve the music community without robbing it. But our industry is inherently "substitutional" – a choice to enjoy a particular recording is also a choice to not consume every other recording at that point in time – and time is, of course, limited. As noted above, the streaming services are already besieged by non-human content, some replicating real artistry and other offering a cheap, lifeless alternative. Generative AI can only exacerbate that unhappy reality.

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.

UMG directs the office to papers and studies identified by the Recording Industry Association of America (RIAA), the National Music Publishers Association (NMPA) and the Copyright Alliance.

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? For example, several jurisdictions have adopted copyright exceptions for text and data mining that could permit use of copyrighted material to train AI systems. Separately, the European Parliament passed its version of the Artificial Intelligence Act on June 14, 2023, which includes a requirement that providers of generative AI systems publish "a sufficiently detailed summary of the use of training data protected under copyright law." See Artificial Intelligence Act, amend. 399, art. 28b(4)(c), EUR. PARL. DOC. P9_TA (2023)0236 (2023), https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

The development of public policy concerning AI is in its infancy internationally, and the U.S. is positioned to be a world leader in this effort. UMG has observed both salutary and counterproductive approaches under discussion in other countries as follows.

We acknowledge and endorse commitments made by the G7, including the Hiroshima AI Guiding Principles and Codes of Conduct issued on Monday, October 30.[12] But also the commitments in the May 2023 Hiroshima Leaders' Communiqué (paragraph 38)[13] and the April

[12] See principle 11 in the Code of Conduct: "Organizations are encouraged to implement appropriate safeguards, to respect rights related to privacy and intellectual property, including copyright-protected content." https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems

[13] https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/20/g7-hiroshima-leaders-communique/

2023 Ministerial Declaration of the Digital and Tech Ministers' Meeting.[14] Among other things, the Communiqué and Declaration broadly emphasize (1) "multi-stakeholder" participation in the development of AI standards that prioritizes fairness, transparency, and adherence to existing laws; (2) commitment to "human-centric and trustworthy AI"; and (3) continued discussion and analysis of how best to safeguard intellectual property rights, including copyrights.

UMG further believes that draft text of the European Union's Artificial Intelligence Act includes helpful proposals on government review of generative AI models before release, continued assessment of those models, recordkeeping provisions, transparency, labeling obligations, and more. As one of the proposed bill's key recitals reads:

> …it is appropriate for the Commission and the AI Office to monitor and periodically assess the legislative and governance framework of such models and in particular of generative AI systems based on such models, which raise significant questions related to the generation of content in breach of Union law, copyright rules, and potential misuse.

The EU Act also commits to many of the broad principles announced by the G7, as noted above, including a focus on creator led, fair, and transparent AI standards that, among other things, respect intellectual property rights. The final version of the EU Act is now being discussed in tripartite negotiations ("trialogue") and UMG hopes that the EU Parliament's position on bookkeeping and transparency will be adopted.

On the other hand, there are some policies, including ones that were adopted years ago before the rise of generative AI, we believe the U.S. should avoid. Whatever their historical merit, generative AI poses threats that render them obsolete and damaging for the creative community, the music industry, and the general integrity of intellectual property law. One

[14] Ministerial Declaration The G7 Digital and Tech Ministers' Meeting 30 April 2023 (see paragraphs 42, 45 and 47)

conspicuous example is Text and Data Mining ("TDM") exceptions to copyright law as enacted in 2021 by Singapore. In addition, the law in Japan, last amended in 2018, also includes an overly broad TDM exception. Although the exception is somewhat limited and includes some protections for rightsholders, it has the potential to cause confusion and harm. Exceptions of this nature, in a world where generative AI swallows vast amounts of data in an uncontrolled fashion, run contrary to basic principles of fairness and the purpose of copyright law to reward creative effort. UMG is pleased that the United Kingdom explicitly rejected such policies last year in recognition of the irrevocable harm it would inflict upon its creative industries.

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.

UMG believes that basic principles of existing copyright law supply the appropriate general framework to govern how generative AI uses copyrighted works. Those principles include, among other things, the requirements of human authorship for copyrightability; the prohibitions against copying or violating any of the other Section 106 rights without authorization from the copyright owner; the standards of substantial and striking similarity; and the prohibitions against altering or removing copyright management information.

However, because of the novelty of generative AI, its methods of copying vast quantities of copyrighted works for training purposes, its parallel capacity to generate an endless supply and variety of outputs in seconds, and the opacity of its systems, generative AI raises issues and threats that copyright law has not previously had to confront. The constant announcement of new lawsuits challenging different aspects of generative AI forecasts continued uncertainty and delay that can best be addressed with certain focused legislation in the areas of recordkeeping and

transparency, copyright law, rights of publicity, and labeling regulations. Details of these proposals are also discussed in further depth in response to questions that specifically call for their discussions, but the key areas requiring attention and the policy reasons for addressing them are set forth in this answer.

A. Copyright – The Reproduction Right and Training

UMG believes that the exclusive right of reproduction reserved for copyright owners in 17 U.S.C. Section 106(1) forbids the unauthorized copying of copyrighted works in training datasets. Ingesting copyrighted works for training purposes can be accomplished in various ways – for example, downloads of content directly scraped from the internet or downloads of third-party datasets – but it necessarily entails a reproduction requiring authorization from the copyright owner. To this extent, existing copyright law should govern the infringing use of copyrighted materials in training.

However, it has been UMG's experience that emerging technologies often speciously try to recharacterize digital copying as either falling outside the statutory language of reproduction or otherwise entitled to statutory exemptions that do not genuinely apply. For example, in *Capitol Records, LLC v. ReDigi, Inc.*, 910 F.3d 649 (2d Cir. 2018), where ReDigi operated a service that permitted the "resale" of downloaded music files, ReDigi protested that it never made an actual reproduction of those files in phonorecords ("material objects in which the work is fixed"), because it merely "migrated" digital data to its own servers, with no trace of the original file left on the seller's device. While the Second Circuit rejected that characterization and found the reproduction right to have been violated, the assertion of that defense resulted in protracted litigation, extensive motion practice, and an appeal. In *Capitol Records, LLC v. BlueBeat, Inc.*, 765 F. Supp. 2d 1198 (C.D. Cal. 2010), BlueBeat defended its reproduction of

sound recordings as ephemeral copying permitted by 17 U.S.C. Section 112, because it claimed to make only momentary copies to create "pure simulations" for streaming. Here too, while the Court found that Bluebeat did not satisfy the statutory conditions for the ephemeral copies exception, the assertion of the defense resulted in needless litigation over a clearly infringing service and allowed that illegal service to continue to inflict harm during the pendency of the case.

While reproduction of copyrighted works for the purposes of training is not genuinely in doubt, UMG is concerned that owners or developers of generative AI who are unwilling to license training data may assert even more creative versions of these specious kinds of ostensible defenses. Particularly given the billions of dollars in resources they enjoy, and their intimate familiarity with technologies that are largely opaque to the outside world, some companies may be particularly adept at facially recharacterizing acts of reproduction as something other than copying or conjuring up reasons why they are entitled to other defenses. Indeed, AI developers appear already to be anticipating copyright challenges; some protest publicly, for example, that their large language models do not "copy or store" training materials or that the training materials are not "accessible" to the AI model after training is complete.

The very fact that generative AI models can produce identical copies of copyrighted lyrics or replicas of other copyrighted works means that they have reproduced the underlying works in some format. When an infringer makes an unauthorized copy, the particular format chosen - whether WAV files, MP3 files, PDFs, or others - does not affect whether an infringement has occurred. An infringing copy is still an infringing copy no matter what format an infringer chooses. Generative AI is not a magical medium and would not be able to replicate a work in its output if it had not copied and retained a reproduction of that work in some digital

form in the first instance. Nevertheless, the specter of endless metaphysical legal battles over whether training materials are "reproduced" in "copies" will not be a sensible, predictable means of curtailing infringement on an unprecedented level.

Moreover, as the technology matures, it could theoretically outpace statutory language that was adopted when generative AI was not even imaginable. For example, if AI someday could ingest copyrighted works for training purposes somehow without technically meeting the requirements of reproduction, including embodiment in "material objects" and fixation of sufficient duration and stability, the law should still provide a remedy.

For these reasons, UMG proposes an amendment to Section 106 of the Copyright Act that clarifies that the use of copyrighted works for training of generative AI is an exclusive right of the copyright owner. This right could be added to Section 106 as a new species of exclusive right or built into the reproduction right with language that clarifies that uses for training are deemed *per se* to implicate the reproduction right. In either case, such an amendment would provide content owners and generative AI developers with a bright line rule that would avoid needless uncertainty and litigation.

B. Copyright – Sound Recordings

The rights of reproduction and creation of derivative works are more limited for sound recordings than other types of copyrighted words. Whereas all other types of works are protected against copying that results in something "substantially similar," 17 U.S.C. Section 114(b) limits the reproduction and derivative work rights for sound recordings to copies that "directly or indirectly recapture the actual sounds fixed in the recording." The same provision doubles down on this limitation by providing that the rights of reproduction and creation of derivative works do "not extend to the making or duplication of another sound recording that consists entirely of an

independent fixation of other sounds, even though such sounds imitate or simulate those in the copyrighted sound recording." This limitation appears to have evolved at a time when physical record piracy was the principal concern attracting legislative attention.

As we have stated throughout this response, when generative AI trains on *any* copyrighted works, whether sound recordings or not, it makes an initial unauthorized reproduction of those works (and, as noted above, we also advocate for clarifying legislatively that using copyrighted works for training should itself be identified as a Section 106 right). In the case of sound recordings, such reproduction necessarily entails capturing the actual sounds embodied in the sound recording. Once captured, those actual sounds may end up incorporated in AI-generated sound recordings.

Nevertheless, the unique features and capacities of generative AI warrant an expansion of the sound recording reproduction and derivative work rights, because the "actual sounds" limitation no longer makes sense in the AI context. Whatever the justification of permitting "imitations" or "simulations" of sound recordings by human beings, generative AI has already evolved to the point where it can instantly produce countless recordings that are indistinguishable from all of (or aspects of) the sound recordings on which they were trained. AI has already proven capable of creating recordings that sound exactly like a given artist – a singer or even an instrumentalist – performing songs they have previously recorded, songs they have never recorded, and perhaps even songs whose lyrics or style would be offensive to their aesthetics or values. An AI-generated soundalike is inherently different from a human imitation, both because of its incredible powers of instant imitation and because it is enabled by copying of copyrighted sound recordings in the training process. These kinds of outputs that sound so much like recordings on which they were trained should accordingly be deemed infringing separate

and apart from whether they incorporate actual sounds from a sound recording on which the AI system was trained. There is no public benefit from these fakes, which have been flooding streaming platforms and which not only steal from the efforts of artists, songwriters and their partner labels and publishers, but mislead the public about the origin and legitimacy of what otherwise appear to be "the real thing."

For this reason, UMG suggests amending the Copyright Act so that a sound recording is protected not only against AI-outputs that recapture the actual sounds from the sound recordings on which an AI system was trained, but also against outputs that are substantially similar to sound recordings on which the system was trained. Without the change, sound recording creators and rightsowners will be vulnerable to abuse by generative AI companies that may try to argue disingenuously that their soundalikes somehow circumvent the actual sounds limitation of Section 114(b). UMG does not believe that such arguments are plausible, but as noted above, its experience with other newly developed technologies raises the concern that it may face such specious arguments. The Section 114(b) exemption for ordinary, human-created imitations/simulations need not be otherwise altered and will continue to protect human beings who wish to make cover recordings that sound like the originals that inspired them. It is rather the unique efficiency, accuracy and breadth of AI-generated sound recordings that pose sufficient threat to require this legislative correction. Moreover, this protection can work in concert with a federal right of publicity, discussed below.

C. Federal Right Of Publicity

As set forth in greater detail in response to Questions 30-33 below, UMG believes that Congress should enact a federal right of publicity statute directed particularly to generative AI. A recording artist's voice, name, image, and likeness represent among the most valuable aspects of

their artistic identities and should be protected from rank imitation or abuse by generative AI. Moreover, deep-fake, and other imitations of an artist's voice, name, image, or likeness generated by AI lead to consumer confusion, unfair competition against actual artists, market dilution and damage to the artist's reputation and brand – potentially irreparably harming their career.

D. Transparency and Recordkeeping

Generative AI developers and deployers must be completely transparent in the data they use to train their systems. They must keep and make readily available to the public detailed and accurate records of the data they have used for training, the provenance of that data, and the existence of any licenses authorizing the use of that data. However, as recent research has illuminated, the lack of transparency is pervasive throughout the generative AI industry.[15] UMG provides greater detail on best practices for this recordkeeping in response to Question 15 below but uses this answer to explain why it is imperative that the U.S. adopt legislation requiring AI developers to maintain such records and make them publicly available.

First, transparency is critical for rightsholders to be able to protect their intellectual property, because of the often-undetectable nature of infringement in the AI context. In virtually all other contexts, infringement is apparent from works circulating in the marketplace – a composer hears the melody of her song used in another song or a record company discovers an unauthorized sample of its recording. Indeed, the law recognizes this reality, because it permits a copyright holder to establish copying circumstantially by proving either striking similarity or

[15] See e.g., https://crfm.stanford.edu/fmti/

access and substantial similarity. Sometimes, generative AI falls into this category, as in the case of the AI-generated content we discover that is substantially or strikingly similar to copyrights we own or control.

However, the use of copyrighted works in the training of generative AI may also be completely invisible to rightsholders, particularly where it is not feasible to search for prompts that would reveal the presence of rightsholder data or test all potential outputs made possible by the training data. Regardless of the nature of the outputs, the unlicensed reproduction and use of copyrighted materials for training purposes is itself a threshold infringement that must be reasonably detectable to rightsholders.

Without a comprehensive system of publicly available recordkeeping, rightsholders face an untenable "Catch-22." When AI training sets are opaque, rightsholders must initiate expensive litigation and pursue discovery to determine whether their content was used in data training sets, and yet without such evidence, rightsholders cannot even initiate litigation in the first place. Legislation requiring that records of materials used to train AI algorithms be made available to the public would remedy this unreasonable burden on enforcing one's copyright.

Second, as we stress throughout, the use of copyrighted and other protected materials in training must be licensed to comply with copyright law. A clear requirement of appropriate recordkeeping will require AI developers to be thoughtful and selective in the content they use for training purposes and will encourage them to seek appropriate authorization in the first instance. Making recordkeeping a fixed requirement will not only deter and reduce infringement but will help promote a robust marketplace of licensing free from litigation and uncertainty. It will, in fact, serve copyright law's basic purpose of rewarding creative authorship.

Finally, recordkeeping requirements will serve consumer interests. Consumers availing themselves of copyrighted works – by listening to a song or purchasing an art book – have a right to know the source of those works. If they are fans of a particular songwriter or performing artist, they want to know that what they are hearing was created by that writer or artist. By extension, consumers should also be able to determine what data was fed into the generative AI they are using. They should not have to guess whether the particular AI system they are using has trained on legitimate, authorized copyrighted works rather than infringements or fakes.

For these reasons, UMG proposes a statutory requirement that AI developers be required to maintain accurate records of all materials and information used to train their model. Those records should be sufficiently detailed and understandable to a layperson: source URLs should be attached to specific and detailed descriptors of individual ingested material (but a mere list of URLs should not be deemed acceptable). These records should be available for public inspection. Because recordkeeping is important beyond the scope of copyright considerations, a recordkeeping requirement could easily exist outside of Title 17.

E. Labeling

As set forth in greater detail in response to Question 28 below, Congress should impose a requirement that fully AI-generated content be labeled as such, particularly in those instances where the risk of consumer deception is high and/or a known human is being mimicked. Consumers have a right to know whether what they are consuming is real, created by a human being, or is the product of a computer program. When a generative AI output is *not* deceptive or when it has only been used as a tool in the creative process, labeling may not be necessary. Communication of labeling and provenance data can be largely accomplished – at least in the music space – via metadata and industry standards such as DDEX. However, in order to establish

a trustworthy and responsible AI ecosystem, Congress should ensure labeling rules must be followed by all.

TRAINING

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

Different AIs train on different material depending on their desired output and functionality. For example, music generating AIs may train on stock music libraries or recordings ripped from streaming services. The AI uses these copies as training material to "learn" all elements of the sound recordings and compositions, including instrumentation, dynamics, timing, melodies, lyrics, harmonies, rhythms, song structures and forms, and timbres. Image generating AIs may train on album cover art, artist photographs, and other visual artwork. Text generating AIs may train on song lyrics, books, and other forms of text (usually) scraped from the internet. The collection and copying of such materials are discussed further below.

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)?

AI developers obtain data for training from multiple sources, many unauthorized by the copyright owner or captured from websites whose terms of use forbid reproducing the information. Often, AI developers either do not disclose the source of their training materials or describe those sources in very general, nonspecific terms. For example, OpenAI explains that it acquires data to train ChatGPT from "three sources of information: (1) information that is publicly available on the internet, (2) information that we license from third parties, and (3)

information that our users or our human trainers provide."[16] Of course, "publicly available on the internet" does not mean copyright-free, license-free, or even "free." These generalized categories shed little light on the actual provenance and lawfulness of the training materials. And copying materials on the internet will obviously capture a wealth of copyrighted material. Based on UMG's experience in communicating with generative AI developers, AI companies acquire training materials broadly from the following sources, each of which poses enormous risk to intellectual property rights.

First, generative AI models scrape vast troves of information from the internet. OpenAI's description of materials "publicly available on the internet" would fall into this category. While AI companies also draw training materials from public domain sources as discussed below, the quality of generative AI output is directly proportional to the quality and quantity of training materials. Accordingly, as AI developers need greater access to more materials and simultaneously want their systems to produce high quality output, many resort to scraping the web for more professional and higher quality text, images, and music.

Of course, materials "available on" or "scraped" from the internet can embrace almost anything, lawful or infringing. When we license music to a digital business partner, that authorization does not allow for anyone with access to that service, site, or platform to use, copy or commercialize the music for another purpose. Our agreements with music streaming services allow streaming of content by a user for personal and non-commercial use only. UMG also licenses song lyrics to authorized lyrics aggregators, the terms of which impose restrictions on how the copyrighted works may be accessed and used. When a generative AI model simply lifts

[16] https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed

those recordings or lyrics from these services or websites, it is making an unlawful reproduction that is not permitted under our licensing agreements. Moreover, the internet is replete with pirated copies of sound recordings, unauthorized compilations of song lyrics, infringing copies of photographs and images, and countless other unauthorized reproductions of copyrighted works. Training on pirated content is no more legal than training on copyrighted content ripped without authorization from an official source. Any further reproduction or distribution of these infringements only multiplies and perpetuates unlawful conduct through an infringing chain of distribution.

Some examples: As detailed by the *Washington Post* earlier this year,[17] Google's C4 training set (used to train Google's T5 and Facebook's LLaMA, among other LLM generative AI models) was trained on "scraped" materials including "scribd.com…, a subscription-only digital library.[18] Also high on the list: b-ok.org…, a notorious market for pirated e-books that has since been seized by the U.S. Justice Department. At least 27 other sites identified [by the U.S. government] as markets for piracy and counterfeits were present in the data set." The *Post* went on to note "The copyright symbol — which denotes a work registered as intellectual property — appears more than 200 million times in the C4 data set."

AI developers also acquire training materials from third party datasets (collections of media that may be used, among other things, for training AI systems). Some AIs may be trained on public domain data, such as Librivox.[19] Others are released under permissive licenses that allow 3rd parties to do whatever they want with the licensed data, such as the CC0

[17] https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/
[18] C4 was subsequently made openly available by Google (see https://www.tensorflow.org/datasets/catalog/c4). No license is specified for this dataset.
[19] https://librivox.org/

License.[20] However, many popular datasets have a permissive license *that does not reflect the true nature of the works they contain*. For example, one such dataset is "The Pile," which is permissively licensed but known to contain pirated books. In addition, AI developers may confuse the license associated with a list of media assets with the license of the assets themselves. For example, the Podcast Index[21] list is available under a permissive license, but many of the podcasts it references have restrictive licenses and are owned by major media corporations. In other words, just because a dataset claims to be open or permissively licensed does not mean the content it contains or references may be legally used for training purposes.

Other datasets are licensed from third party providers who purport to impose some restrictions on how the data is used, such as a requirement that the data only be used for noncommercial purposes. However, there is again no assurance that use of such data is lawful. The fact that the dataset provider imposes restrictions does not mean it acquired the materials lawfully in the first place. Additionally, even where the dataset provider is authorized to sublicense the data for some purposes, the terms and conditions of those licenses may vary and may not permit the uses to which the AI developer will put the data. UMG is aware of datasets licensed only for non-commercial, scientific, or research purposes that are nevertheless used for brazenly commercial purposes in subscription-based AI models.[22] Thus, training materials "licensed from third parties" may only have been authorized for noncommercial or other limited purposes or might not even have been authorized by the copyright holders at all; in either case, those materials cannot be lawfully be used to train a commercial generative AI.

---

[20] https://creativecommons.org/publicdomain/zero/1.0/
[21] https://podcastindex.org/
[22] *AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability*, Andy Baio, *Waxy*, September 30, 2022

Some AI developers assert they train on "ethical" or "copyright-safe" sources, including public domain content, Creative Commons or similar freely available content, content obtained from stock libraries, or specifically commissioned content. However, UMG has communicated with numerous AI companies that claim that they are "ethically trained," only to learn that they are using training datasets containing infringing materials.

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?

Many of the generative AI companies initially jumped the gun by training on unlicensed and unauthorized materials from the outset. The goal was apparently to consume as much data as possible without regard for the rights of intellectual property owners. However, there is a promising trend of developing licensing models for the authorized use of content for training. Some examples include the following:

- Shutterstock has licensed its image, video and music libraries and associated metadata to OpenAI for training purposes. Shutterstock offers a Contributor Fund that compensates artists for use of their creations in training Shutterstock's generative technology and also provides ongoing royalties tied to licensing activity for newly generated assets.[23]

- Getty Images launched Generative AI by Getty Images,[24] which trains on its own library of licensed images, compensates the artists whose images were used for training, and indemnifies users of the service against all claims of infringement. Adobe has announced a similar program.[25]

- In July 2023, the Associated Press and OpenAI announced an agreement pursuant to which OpenAI will license portions of AP's text archives to develop generative AI in news products and services.[26]

[23] https://support.submit.shutterstock.com/s/article/Shutterstock-ai-and-Computer-Vision-Contributor-FAQ?language=en_US
[24] https://www.gettyimages.com/ai/generation/about
[25] https://www.computerworld.com/article/3699053/adobe-offers-copyright-indemnification-for-firefly-ai-based-image-app-users.html
[26] ChatGPT-maker Open-AI signs deal with AP to license news stories, Matt O'Brien, AP, July 13, 2023

- Stable Audio, Stability AI's music generation tool, is trained on music licensed from AudioSparx, a music library and stock audio website.[27]

UMG strongly supports efforts to ensure lawful and ethical licensing of copyrighted content for training purposes. The companies noted above maintain enormous libraries of copyrighted content, including images, news archives, music, and video. While we obviously do not know the specifics of these agreements, the fact that these companies have already established licensing and compensation models in partnership with leading AI developers proves that licensing training data is feasible, practicable, and beneficial for both content owners and the developers who license the content. These models also prove that licensing does not limit the availability of content, but rather promotes access to that content in lawful and ethical ways.

As set forth throughout this response, UMG is committed to developing models and best practices for licensing musical content to AI developers as well. Given the vast library of musical content we control, and our constructive dialogues with a number of companies, we have every confidence that we will develop fair and lawful solutions that promote the technological promise of AI without imperiling the rights or interests of the creative community.

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 noted above, some AI developers use allegedly public domain content for training purposes, but it sometimes may not be sufficient to provide adequate quantities of data for particular generative AI projects. As a result, generative AI models inevitably look to additional sources containing copyrighted content to acquire adequate training data. Moreover, relatively

[27] https://www.stableaudio.com/faqs

little recorded musical content that would be of most utility to AI is in the public domain. No sound recordings subject to full federal copyright protection (*i.e.*, those fixed after February 15, 1972) have yet entered the public domain. Moreover, pursuant to the Music Modernization Act, the vast bulk of pre-1972 sound recordings (indeed all published from 1924 forward) are still protected. And more broadly, the music of greatest popular interest, whether compositions or sound recordings, is of more recent vintage, remains in copyright, and thus requires a license.

As mentioned above, there are also cases where developers have funded AIs made for research purposes, and then later used the research work for commercial purposes, effectively "laundering" the training data.[28]

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.

The way that AI training works makes retention of considerable amounts of training materials nearly inevitable, and UMG is highly skeptical of developers' claims that they do not "store" those materials. To start, software development is an iterative, exceedingly repetitive process, and as models are developed and modified, they will need to be retrained.[29]

AI developers make continuous improvements and modifications that require some form of training, be it retraining from scratch, refining (or "fine-tuning") a base "foundational" model, or any other form of training. In all cases, the training materials need to be stored. AI developers are also constantly working on the next version of their AI, which necessitates on-going access to

[28] *AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability*, Andy Baio, *Waxy*, September 30, 2022

[29] For example, Model v1 is trained on dataset v1. Then AI developer collects more data to make dataset v2 and wants to train model v2 on dataset v2. A developer would generally *also* want to train model v2 on dataset v1, to double check that model v2 actually brings an improvement.

the training material. Retaining training data assists with benchmarking, which is good engineering practice. Sometimes datasets need to be customized to the task at hand. Moreover, training data is often manipulated to increase the amount of available data (*i.e.*, "data augmentation"). For example, images may be flipped or cropped, and music may be cut into clips or run through a process of source separation. Customized and manipulated data is derived from the "original" training set.

The only circumstance where training material would not be needed throughout this repetitious process is if generative AI developers got their initial AI design "right" on the very first try and did not build additional versions or do any associated refinements – a highly unlikely scenario.

Finally, as detailed elsewhere in this response, AI developers should be required to maintain scrupulous records of the materials they use for training purposes. Those records must, in turn, be subject to audit to ensure their accuracy and completeness. For any such audit rights to be meaningful, AI developers must be required to maintain those training materials for a sufficient period of time.

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.

As noted above, AI developers acquire training materials from various sources, which may govern how the training materials are reproduced. Whatever the mechanisms involved, it is our understanding that the training data is always reproduced.

For example, AI developers often use third party datasets that link to content online. Indeed, many of the most popular datasets are lists of URLs and supporting data, such as those located at LAION.AI[30] and Hugging Face.[31] In these cases, the dataset just points to files, so the developer must download copies of the data from third party sites, which may contain copyrighted and even pirated content and often have terms of service that forbid downloading or commercial uses.

In other instances, the training dataset may include all of the crawled data itself. In these instances, the developer either uses a script to download the portions of the data it wants or downloads the whole dataset and works locally or from its cloud instance. Regardless of whether the training materials are downloaded from third party sites or directly from the datasets, the developer must make copies of the data in order to use it for training purposes.

When AI developers scrape training data from the web, they crawl and catalog portions of the web looking for content useful for their purposes and then simply download what they want to use.

With respect to the duration of these reproductions, as described above in response to Question 6.4 above, the training process invariably requires retention of copies of training materials. Because of the iterative nature of training, the refinement of AI models, the constant development of new versions of those models, and the manipulation of data to increase the amount of available data, training data will often be stored for long periods of time.

And finally, as set forth above in response to Question 5, each of these mechanisms of copying training data is a *per se* violation of the exclusive right of reproduction. Moreover, as

[30] https://laion.ai/blog/laion-400-open-dataset/
[31] https://huggingface.co/datasets/google/MusicCaps

also explained in that response, in order to guard against the inevitable efforts to recharacterize these reproductions as falling outside traditional notions of copying, and also to guard against future technological efforts to circumvent the traditional understanding of reproduction, UMG proposes that the use of training data itself be identified as an exclusive right of copyright owners in Section 106 of the Copyright Act.

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

A generative AI algorithm will often be capable of resynthesizing its training material, and this is often described as "memorization." Some will claim that AIs only learn concepts but that is clearly not true, as LLMs are able to reproduce song lyrics[32] or passages from books, and image generating AIs can often synthesize images that are clearly identifiable as a derivative of their training material when given the right prompt.[33]

This data is encoded and stored in the AI model parameters, but developers do not understand the exact data representation mechanism yet. However, the storage or representation method should not be of concern because the important issue is that it can be proven that data is stored or represented in these systems, because it is possible to get exact or strikingly similar copies of training material as output from a trained model, once the model no longer has direct access to the training data.

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?

[32] See *Concord Music Inc., et al v. Anthropic PBC*, Case No. 3:23-cv-0192, United States District Court for the Middle District of Tennessee

[33] *AI Spits Out Exact Copies of Training Images, Real People, Logos, Researchers Find*, Chloe Xiang, Vice, February 1, 2023

Although this is an active area of research, today it is not possible for an AI model to "unlearn." Google itself launched a "Machine Unlearning Challenge", in an effort to advance the science.[34] On an encouraging note, recently Microsoft published a paper describing a method for making an AI trained on the books3 dataset (which contains unlicensed copyrighted books, including Harry Potter) forget some elements of the material on which it was trained.[35] But absent further technological developments, removing content from or "unlearning" an AI today requires retraining.[36]

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?

The clearest way to identify a model's training material is via disclosure by the developer. In some cases, developers publish research papers, data cards, or blog posts describing their training material. Outside of direct disclosure, there are some means of determining whether an AI model trained on specific materials, but they are imperfect, often anecdotal, and not capable of identifying all or even significant portions of training materials.

Most obviously, where an AI model generates outputs that are identical or substantially similar to copyrighted works, the model was trained on those works. A clear example is the ability of certain generative AI to produce nearly identical copies of copyrighted song lyrics. UMG has also observed instances where the output combines qualitatively and quantitatively significant portions of more than one song into a single output, often authored by different

[34] https://blog.research.google/2023/06/announcing-first-machine-unlearning.html
[35] https://www.microsoft.com/en-us/research/project/physics-of-agi/articles/whos-harry-potter-making-llms-forget-2/
[36] However, if a base model trained on non-infringing content, but it was fine-tuned on infringing content, only the fine-tuning step would need to redone.

songwriters and controlled by different publishers. Again, the model would not be able to accomplish this combination without having trained on the underlying copyrighted works.

This means of determining what training materials were used is completely impractical as a means of detection, however, as it would require endless entry of text prompts or blocking certain kinds of output in the hopes of generating sufficient clues about the identity of those training materials. Also, generative AI models can interfere with this method of detection by filtering text prompts or blocking the output of certain material, thus making the unauthorized reproductions invisible to users. The music community should not bear the burden of constantly testing AI models to determine whether their creative efforts have been stolen.

The research and technology communities have not yet invented a tool to identify training materials from outputs either. In other words, we cannot rely on a technical solution to determine what data was used to train a given AI.[37] The inability to detect the materials within a training set further counsels in favor of the transparency and recordkeeping measures discussed elsewhere in this response.

8. Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use? Please discuss any case law you believe relevant to this question.

The wholesale appropriation of UMG's enormous catalog of copyright-protected sound recordings and musical compositions to build multibillion commercial enterprises is anything but fair use. Such appropriation falls outside of the broad justifications that typically characterize fair

[37] There is ongoing research into ways to decipher whether a model trained on a particular dataset, including membership inference (given access to a model, can you identify if a given piece of content was used to train that model) and model inversion (extracting training content from a model). Additionally, content watermarking has provided some evidence of copyrighted materials in training datasets; for example, Getty is establishing that Stable Diffusion trained on its images because various output images display Getty Watermarks.

uses – such as parodic comment or factual news reporting – and does not satisfy the statutory standards for fair use. It is simply theft on an unprecedented scale that threatens the core of our business and our mission.

As noted above, the music industry largely operates on a licensing model. We license our content for a myriad of purposes, ranging from exploitation of our works for use in particular media (streaming or audiovisual works, for example), to using our content to develop important new technologies, such as AI-powered tools that help make recommendations to users who show preferences for certain kinds of music. In all these instances, the authorizations we grant support our existing artists and songwriters and provide the resources to discover and develop new artists and songwriters. These licenses, in other words, reward the very creativity copyright law is designed to promote. Allowing AI to swallow our content whole for training without authorization destroys that fundamental purpose. It undermines a critical market for the creative content we control and reduces the creative community's incentive to keep creating. And it does so for the benefit of companies that attract staggering investments and charge hefty subscription fees.

To make matters worse, this copied content is in turn used to train computers to generate *non-human* and uncopyrightable content that directly competes with the creative works we seek to promote. Training on our content thus both invades a licensing market critical to our existence and leads to digitally-created market substitutes for genuine human creations. This scenario creates two fatal impediments to the application of fair use. First, it does the opposite of promoting human creativity by rewarding only computer-generated creations that merit no copyright protection, and second, it creates the precise kinds of market substitutes the courts routinely refuse to treat as fair use. We can think of no precedent for finding this kind of

wholesale, commercial taking that competes directly with the copyrighted works appropriated to be fair use.

While Questions 8.1 and 8.5 below ask specifically about two of the four statutory fair use factors, we use our response to this question 8 to address all the fair use factors and provide a consolidated analysis of why using UMG's content for training cannot and should not qualify for fair use. Note also that we confine our analysis to UMG's content, with the understanding that different types of copyrighted works might warrant a different analysis.

A. Training On Copyrighted Music Does Not Fall Within The Broad Purposes of the Fair Use Doctrine

Before embarking on specific analysis of the statutory fair use factors, UMG first considers whether AI training on copyrighted music serves the broader purposes of the fair use doctrine. The *Warhol* decision reiterates that the Copyright Act balances "competing claims upon the public interest: Creative work is to be encouraged and rewarded, but private motivation must ultimately serve the cause of promoting broad public availability of literature, music, and the other arts." *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258, 1273 (2022) (citation omitted). Fair use reflects this "balancing act between creativity and availability" and avoids "rigid application of the copyright statute when, on occasion, it would stifle the very creativity which that law is designed to foster." *Id.* at 1273-74 (quoting *Stewart v. Abend*, 495 U. S. 207, 236 (1990)); *Google LLC v. Oracle America, Inc*., 141 S. Ct. 1183, 1196 (2021) (same).

When a generative AI model consumes copyrighted sound recordings, melodies, lyrics, or other musical content to train itself to create digital facsimiles of human works, it serves none of these interests. Fair use ensures public availability of "literature, music, and other arts" – all

products of human creativity – so that other humans can draw on those works to create new ones. The purpose of the doctrine is not vindicated when generative AI algorithms learn to mimic human creativity. To be sure, there are ample non-copyright incentives for AI developers to create these systems – patent law, copyright protection for computer codes, and the extraordinary investment AI is attracting – but copyright law is not the vehicle to promote those technological incentives. Copyright law and the fair use doctrine balance incentives and availability to promote human creativity.

Moreover, training on copyrighted music falls far outside the paradigmatic uses of creative works central to the fair use doctrine. *Warhol* reminds us to consider the illustrative list of purposes that are considered fair uses identified in the preamble to 17 U.S.C. §107 – "criticism, comment, news reporting, teaching **…** scholarship, or research" – as a "guide" to assessing the first fair use factor. *See Warhol*, 143 S. Ct. at 1274. *See also Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 578 (1994) (preamble serves as "guide" on first fair use factor). Similarly, the Second Circuit warned that it is error to ignore the illustrative list of potential fair uses in the preamble when assessing a fair use defense. *Ringgold v. Black Entertainment TV, Inc.*, 126 F.3d 70, 78 (2d Cir. 1997). Each of these illustrative fair uses involves borrowing the copyrighted work to create new works that draw beneficially upon the borrowed materials. Each of these uses are justified by the fact that the copyrighted work is a building block to a new expressive work, whether a book review or a newscast about a matter of public interest. Authors, in other words, stand on the shoulders of prior authors.

The unauthorized copying of copyrighted music to train AI strays far from these kinds of uses which enrich public dialogue and comment. In each of Section 107's illustrative uses, the right of a human author to comment, or engage in scholarship, for example, is enhanced by the

ability to draw on prior works. Ingesting music to create non-human replicas serves no similar purpose. It may be that ingesting some data benefits scholarship or research broadly; training that, for example, facilitates a medical discovery is something UMG applauds. But copying entire catalogs of human art to create computer-generated music serves no similar beneficial purpose. Thus, at the broadest level, training on copyrighted music is squarely at odds with the interests the fair use doctrine is intended to protect.

B. The Statutory Factors Do Not Support Fair Use for Training

i. The Purpose And Character Of The Use

Question 8.1 below asks how the first statutory fair use factor – the purpose and character of the use – should be assessed in light of recent Supreme Court precedent. It further asks us to identify which "use" should be evaluated and whether different stages of training raise different considerations. We address those inquiries here as part of the overall fair use analysis.

We start with identification of the relevant "use" which must be evaluated as either fair or not. The use in question is the reproduction of copyrighted music for training purposes. It is that use that allows AI developers to create multibillion dollar businesses that attract investors and subscription fees. It is that use that invades a robust market for licensing music. And it is that use that ultimately enables outputs that are offensive to the rights of genuine human creators. As Justice Gorsuch explained in concurrence in *Warhol*, the use to be evaluated under the first fair use factor is the "challenged use." *Warhol*, 143 S. Ct. at 1274. Here, in the first instance, UMG challenges the wholesale consumption of its catalog of music, without permission or compensation, to create and support a commercial business.

To this extent, UMG believes that the fair use analysis should disaggregate the use of our copyrighted content for training purposes from the variable output that AI is able to generate by

reason of this training. Many outputs will be facially infringing – identical copies of song lyrics, for example – while other outputs may not resemble any particular training material, such as a new set of lyrics or the creation of a generic pop tune. But all such outputs, which are enabled by training on our content, will broadly compete with our content in a way that is anything but fair. As set forth below, where a challenged use creates a potential substitute for the copyrighted work, it cannot be fair. Moreover, to the extent that any given output is fortuitously closer to a fair use – for example, if AI were to generate a true parody in the sense of *Acuff-Rose* – such outputs can be assessed on a case-by-case basis. Regardless, the initial reproduction of the entire UMG catalog to train AI is the use in question, and it is not fair by any measure.

The Supreme Court's recent holdings on fair use support this position. *Warhol* cabins analysis of the first fair use factor – "the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes," 17 U.S.C. Section 107(1) – in a number of key ways. First, it emphasizes that this factor hinges on whether the purported fair use potentially *substitutes* for the copyrighted work it borrows:

> The "central" question [of the first fair us factor] is "whether the new work merely 'supersede[s] the objects' of the original creation . . . ('supplanting' the original), or instead adds something new, with a further purpose or different character." *In that way, the first factor relates to the problem of substitution—copyright's bête noire*. The use of an original work to achieve a purpose that is the same as, or highly similar to, that of the original work is more likely to substitute for, or "'supplan[t],'" the work.

See Warhol, 143 S. Ct. at 1274 (citations omitted; emphasis added). The key concept is that the alleged fair use cannot seek to "achieve a purpose" closely akin to that of the copyrighted work it borrows.

Second, *Warhol* requires balancing the new "purpose" or "character" of a purported fair use against the commercial nature of that use:

> In sum, the first fair use factor considers whether the use of a copyrighted work has a further purpose or different character, which is a matter of degree, and the degree of difference must be balanced against the commercial nature of the use. If an original work and a secondary use share the same or highly similar purposes, and the secondary use is of a commercial nature, the first factor is likely to weigh against fair use, absent some other justification for copying.

See Warhol, 143 S. Ct. at 1277. In other words, even where the purported fair use achieves a "further purpose or different character," that transformation of purpose or character must always be weighed against the commercial nature of the use. And a commercial use whose purpose is close to that of the copyrighted work will almost never constitute fair use.

Finally, the Court warned that the "transformative" nature of an alleged fair use – the extent to which the use revealed a "further purpose or different character" – must not interfere with a copyright owner's exclusive right to prepare derivative works, defined by the Copyright Act as those in which a work is "transformed": "an overbroad concept of transformative use, one that includes any further purpose, or any different character, would narrow the copyright owner's exclusive right to create derivative works." *Id.* at 1275. The copyright owner's exclusive right to "transform" her work thus serves as yet a third check on the ability of third parties to create their own transformations without authorization.

Copying UMG's content for training purposes fails all these tests. First, it effects an unlawful "substitution—copyright's bête noire," *Warhol*, 143 S. Ct. at 1274 – of the purposes and aims of copyrighted music at several levels. Training enables the creation of market substitutes for the copyrighted works it appropriates. It either creates actual replicas that can be swapped out for the originals or more broadly musical content – lyrics, soundalikes, or even generic recordings – that compete in already saturated market for music. Whether other kinds of content may serve genuinely distinct purposes, using human music to create artificial

competitors is neither transformative nor fair. Moreover, at the industry level, training also supplants a key purpose and aim of record labels and music publishers. As noted above, the music industry operates by licensing music for countless purposes, including the development of new technologies. Training achieves the same purposes without a license.

The purported transformative nature of training on copyrighted music is also significantly outweighed by its fundamentally commercial nature. There is little doubt that generative AI companies are commercial, moneymaking enterprises. For example, most major AI players charge users subscription or per use fees. Microsoft's 365 Copilot, operating on OpenAI, charges $30 per user per month on top of existing business fees.[38] Stability AI charges a $149 per month subscription fee for nearly unlimited use and competitor Midjourney charges up to $60 a month.[39] Many voice clone models charge subscription fees, as well.[40] Numerous generative AI models train on copyrighted materials – or artists voices – precisely so that they can lure paying subscribers and investors with the promise of creating lyrics, or music, or album covers at the proverbial push of a button. In other words, UMG's content is a key building block for the commercial allure of many AI models. That direct link – using UMG's content to build a business that attracts investment and subscribers – heavily outweighs any alleged "different purpose" those models can assert.

And finally, the use of music for training is nothing short of a direct assault on the derivative work rights of the music community. Composers routinely reinvent their songs through arrangements and new versions. Their publishers license those songs to be used

[38] *Microsoft will charge you $30 per month to use A.I. tools with Office—on top of what you already pay*, Diana Bass and Bloomberg, *Fortune*, July 18, 2023
[39] *Adobe to Sell Generative AI Subscription With Copyright Assurance*, Brody Ford, *Yahoo! Finance*, June 8, 2023
[40] https://slashdot.org/software/voice-cloning/

derivatively in other compositions, in soundtracks, and countless other uses. Performing artists create remixes, combine recordings with authorized samples, and create collaborations with both live artists and posthumous recordings. Indeed, as noted above, some UMG artists are using generative AI *with their consent* to produce these derivative works. It simply cannot be a fair use for generative AI to steal those opportunities from the music community. In this way too, training on copyrighted music runs afoul of the first fair use factor as explicated by the Supreme Court.

ii. The Nature of The Copyrighted Work

Under the second fair use factor, there can be no dispute that music is "closer to the core of intended copyright protection" than less artistic and expressive works, and that accordingly, "fair use is more difficult to establish" when music is copied. *Campbell*, 510 U.S. at 586. The Supreme Court's decision in *Oracle* revived the importance of the second factor, which has sometimes been subordinated by the lower courts. The Court stressed that "copyright's protection may be stronger where the copyrighted material is fiction, not fact, where it consists of a motion picture rather than a news broadcast, or where it serves an artistic rather than a utilitarian function." *Oracle*, 141 S. Ct. at 1197. In finding fair use, it started its analysis with the finding that the computer code in question, "is, if copyrightable at all, further than are most computer programs (such as the implementing code) from the core of copyright." *Id.* at 1202. The necessary implication of this analysis is that music – the quintessential artistic, rather than functional, work – is less susceptible to fair use.

iii. The Amount And Substantiality Of The Portion Used In Relation To The Copyrighted Work As A Whole

The fact that generative AI ingests sound recordings and musical compositions in their entirety militates further against fair use. *See ReDigi*, 910 F.3d at 662 (defendants copying of "the whole of Plaintiffs' copyrighted sound recordings … tends to disfavor a finding of fair use").

iv. The Effect Of The Use Upon The Potential Market For Or Value Of The Copyrighted Work

Question 8.5 below asks how the fourth fair use factor should be measured in the case of training on copyrighted materials, and whether the inquiry should focus on outputs that compete with a particular copyrighted work, the body of works of the same author, or the market for that general class of works. UMG provides its complete analysis of this factor and these questions in this response.

The fourth fair use factor – which looks to "the effect of the use upon the potential market for or value of the copyrighted work" – has been dubbed "undoubtedly the single most important element of fair use." *Harper & Row, Publishers Inc., v. Nation Enterprises*, 471 U. S. 539, 566 (1985). *See also Stewart v. Abend*, 495 U.S. 207, 238 ("The fourth factor is the 'most important, and indeed, central fair use factor'") (quoting 3 M. Nimmer, *Nimmer on Copyright* § 13.05[A]). The key consideration under this factor is "whether defendant's utilization functions as a market substitute for plaintiff's work," and "goes beyond plaintiff's present exploitation to consider the impact of the use on potential markets plaintiff may subsequently enter." *Nimmer* §13F.08. Given the focus on market "substitution," factor four inherently overlaps to some degree with the first factor.

While Question 8.5 requests comment on how different classes of *outputs* compete with the copyrighted work, training on copyrighted music imposes a market interference at both the

input (training) and output stages. As noted above, music is licensed for a nearly unlimited range of uses. Apart from simply granting rights of exploitation – for example, the right to perform or distribute a work – UMG licenses musical content for incorporation into other works, as in the case of samples, interpolations, remixes, film scores and soundtracks, and other audiovisual uses, and for various other categories, including fitness companies, gaming companies, health/wellness companies, various social media uses, sheet music, and lyrics databases. UMG licenses its content to social media platforms like YouTube to ensure that our artists and songwriters are compensated – and respected – on the platform. And UMG licenses its content for various technological and AI-powered purposes, including for example, for the development of fingerprint systems and music recommendation systems. While the existing market is vast, the potential market is unlimited, and AI training is the latest iteration of an emerging but very real market. Allowing training without license essentially destroys that market before it can mature.

Indeed, new valuable markets for music are constantly developing. For example, the music industry generates substantial revenue from music clips, including clips used for YouTube Shorts, TikTok videos, and similar social media. These revenue streams did not exist a decade ago. Short clips from music generating AIs will have direct commercial impact on these revenue streams.

At the output level, generative AI creates artificial music that competes directly with human-created musical content. Given the sheer breadth of the training materials and the nearly unlimited scope of potential outputs, it would be inequitable to confine the analysis to singular copyrighted works or the works of a given author. Generative AI competes directly with all classes of musical works by offering synthetic substitutes for human musical artistry. If you are a rock musician, a jazz musician, or a classical performer, AI has the capacity to supplant your

entire creative output. If the fourth fair use factor is the most important in the fair use analysis, its one-sided tilt towards copyright protection is enough standing alone to defeat fair use.

C. Neither The Intermediate Nor Mass Copying Cases Support A Fair Use Defense For Training On Copyrighted Works

UMG believes two additional fair use issues warrant brief comment, because UMG anticipates that AI developers will invoke them. They involve, first, cases concerning "intermediate copying," and, second, the rare instance where massive copying of entire bodies of works is deemed fair use. Neither support applying fair use to generative AI training on copyrighted materials.

i. Intermediate Copying

AI developers may claim that training on copyrighted materials involves only "intermediate copying" on the path to an ultimately justified purpose of enabling new, non-infringing content. As noted above, training on copyrighted music often leads to brazenly infringing outputs, but even where the outputs are not facially similar to particular training materials, intermediate copying cannot be fair use under the governing legal standards.

At the threshold, even where copying is "merely preliminary to further uses, intermediate copying can be no less an infringement of the copyright owner's exclusive reproduction right than 'final' copying." *Nimmer* § 13F.14[C][1]. However, in narrow circumstances, intermediate copying may constitute fair use. In *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1992), Plaintiff Sega manufactured the Genesis video game console on which only its own video games could be played. Accolade copied and reverse engineered the computer code for Genesis so that consumers would be able to play Accolade's games on Sega's device. In finding this intermediate copying to be fair use, the Court stressed that Accolade's purpose was to "study the

functional requirements for Genesis compatibility" to that it could make other games usable with the Sega console; Accolade's identification of those functional requirements helped increase the number games compatible with the Genesis console, thereby fostering "growth in creative expression, based on the dissemination of other creative works and the unprotected ideas contained in those works, that the Copyright Act was intended to promote'; and Sega's effort "to monopolize the market by making it impossible for others to compete runs counter to the statutory purpose of promoting creative expression." *Id.* at 1522-24. *See also Sony Computer Entertainment, Inc. v. Connectix Corp*., 203 F.3d 596 (9th Cir. 1999), *cert. denied*, 531 U.S. 871 (2000)(intermediate copying of Sony's code to so that owners of Sony's Virtual Game Station would be able to play Sony games on personal computers deemed fair use where intermediate copying necessary to study functional elements and Sony's effort to monopolize devices that play Sony games inimical to copyright law).

In these instances, intermediate copying allowed access to functional, noncopyrightable elements of computer programs so that subsequent developers could create new expressive works. Generative AI, on the other hand, copies and reproduces the essential creative expression of music – its lyrics, its harmonies and melodies, and its actual sounds as captured in sound recordings – to enable digital generation of synthetic music. Moreover, in both *Sega* and *Sony*, forbidding intermediate copying would have perpetuated unfair monopolies plaintiffs sought to protect, whereas the music industry simply asks that AI license its content the way all other technologies do. And finally, these intermediate copying cases are necessarily limited in their application by the *Oracle* decision. As the Supreme Court noted in citing both *Sega* and *Sony,* fair use plays a particularly "important role in determining the lawful scope of a computer program copyright," which the Court acknowledged was farther from the "core" of copyright

protection. *See Oracle*, 141 S. Ct. at 1198. Music, which is surely closer to the "core" of copyright protection, is not susceptible to the same breadth of fair use defenses.

ii. Mass Copying

It is only the advent of newer technologies that have permitted copying on a broad, industry-wide scale, and the result is a paucity of caselaw on the subject. Nevertheless, UMG believes that two Second Circuit cases broadly help set parameters for the application of fair use to copying at this level.

In *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015), Google made digital copies of "tens of millions of books" to create the "Google Books" searchable database that allows users to search for words and phrases appearing in published books. The search results displayed "snippets" of text, usually consisting of several lines, containing the searched-for terms. While acknowledging that the case "tests the boundaries of fair use," *id.* at 206, the Second Circuit nevertheless found the service to constitute a transformative fair use, because it not only facilitated research for specified terms but also provided appropriate textual context for researchers to assess the relevance of the search results to their topic of interest. *Id.* at 216-18. In other words, it served the "highly transformative purpose of identifying books of interest to the searcher." *Id.* at 218. The Court further held that the brevity of the snippets ensured that they did not offer a "significantly competing substitute for the copyrighted work" and thus did not displace or harm an existing market for the underlying books from which the snippets are abstracted. *Id.* at 222-224. The use, in other words, promoted a beneficial purpose of research without supplanting any licensing market.

In contrast, *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 173 (2d Cir. 2018), *cert. denied*, 139 S. Ct. 595 (2018), exceeded the limit of what the Second Circuit would tolerate

in the way of mass copying. Defendant TVEyes continuously recorded "vast quantities of television programming, compiling the recorded broadcasts into a database that is text-searchable … and allowing its clients to search for and watch (up to) ten-minute video clips that mention terms of interest to the clients." Plaintiff Fox News challenged defendant's redistribution of those clips of Fox's content to users. Although the "watch function" was modestly transformative "insofar as it enables users to isolate, from an ocean of programming, material that is responsive to their interest," *id*. at 177-78, defendant's provision of full ten-minute clips to users was "extensive" and "inclusive of all that is 'important' from the copyrighted work" to users seeking those news segments. Those clips also "undercut Fox's ability to profit from licensing searchable access to its copyrighted content to third parties." *Id.* at 179-80. On balance, making such extensive portions of Fox's content available and usurping its potential licensing of that content defeated fair use.

While neither case provides a clean analogy to AI training on mass quantities of music, the principles of these cases illustrate why training cannot satisfy the aims of the fair use defense. In *Author's Guild,* the mass copying of books facilitated research without supplanting the market for whole books. Research is, of course, a paradigmatic fair use purpose called out in the preamble to Section 107, and directly serves the interest of making copyrighted works accessible so that later researchers can contribute new and useful authorship. In *Fox News*, in contrast, the copying of newscasts lead to the dissemination of clear market substitutes – why pay for a subscription when you can essentially get all you need from a news cast through defendant's service?

Training on music serves none of the interests that justified the copying in *Author*'*s Guild*. Rather than promoting research while preserving the existing market for whole books,

training simply steals copyrighted works to enable the digital creation of market substitutes. As noted above, training does not facilitate the creation of further expressive works, but actually impedes it by supplanting human creativity with computer-generated content. Moreover, where the output in *Author's* Guild was controlled and predictable – short snippets only so long as necessary to provide context for a term of interest – the output of AI is wildly unpredictable and uncontrolled and generates everything from brazen infringements to competing new (albeit synthetic) music.[41] It is far closer to the improper substitution deemed unfair in *Fox News.* Ultimately, massive copying should always warrant the most skeptical of scrutiny, and when it operates in the ravenous and poorly controlled fashion of AI training, it cannot qualify as fair use.

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?

UMG incorporates by reference its response to Question 8 above, save for the final sentence of this Question 8.1. UMG does not believe there is any principled reason to distinguish among different stages of training, such as pre-training and fine-tuning. The question for all remains the same – does the training involve use of UMG's content and is that use excused by the fair use doctrine? As noted above, an AI model cannot effectively train on copyrighted materials without reproducing them at some stage. And as also noted above, even where those reproductions are "intermediate," they are far afield from the kinds of software code reproductions that served beneficial purposes in cases like *Sega* and *Sony*. Finally, fair use

[41] Note that in the case of music, short clips of music are themselves monetized products, as evidenced by products such as YouTube Shorts and TikTok.

should never exonerate copying that is for a clearly commercial purpose, supplants a licensing market, and generates market substitutes. The same analysis applies wherever copying takes place in the stages of training.

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

The entities that collect and distribute copyrighted material for training without an express license from the copyright owners to do so are co-equal infringers with the AI companies that use that training material. As in the case of any chain of distribution – for example, an importer who sells to a distributor, who sells to a retailer – each party to the infringing conduct is subject to either direct or secondary liability. By way of example, an entity that compiles links to copyrighted content and advertises its list as suitable for AI training may be a contributory infringer because it knowingly facilitates acquisition of infringing materials from third party links, whereas one that collects copies of copyrighted content itself might be a direct infringer. If copying copyrighted content for training without authorization from the copyright owner is infringement – and UMG contends that it certainly is – then every entity who either engages in or knowingly facilitates that copying is liable under basic principles of direct and contributory copyright liability.

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

The scenario described in this question – datasets prepared for noncommercial or research purposes subsequently adapted for commercial ventures – is far from hypothetical.

UMG has spoken to several AI developers who assert their commercial products do not infringe copyright "because the training data was openly and/or permissively licensed," when in fact those data sets include unlicensed copyrighted data. We are also aware of some who have used training data created by universities or non-profit organizations. For example, the Stable Diffusion AI relies on datasets intended for research purposes but the AI was made available for commercial use.[42] UMG is aware of other instances of "data laundering," the conversion of AIs or data compiled for research purposes to commercial uses;[43] some of these projects have even been funded by commercial entities.[44] These examples reveal that the companies often ignore the designation of an AI or dataset as being for "research or non-commercial use only."

Warhol also reminds us that use of a copyrighted work for some purposes may be infringing while another might be a fair use. *See* 143 S. Ct. at 1291. However, UMG's marketplace experience is that its musical content is routinely used for *unfair* uses, and that the nonspecific promise of "research" or "noncommercial" training is commonly broken. Given this unreliability, the only practical solution is simply to require a license to train on copyrighted music to avoid chasing after the original and subsequent provenance of training materials.

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?

---

[42] https://github.com/CompVis/stable-diffusion/blob/main/Stable_Diffusion_v1_Model_Card.md - training
and https://laion.ai/blog/laion-5b/

[43] See Is Big Tech using Data Laundering to cheat artists? Devansh- Machine Learning Made Simple, *Medium*, November 14, 2022 ("*We thank our sponsors hugging face, doodlebot and stability for providing us with computing resources to produce this dataset!"); https://laion.ai/about/; Startup Behind AI Image Generator Stable Diffusion Is In Talks To Raise At A Valuation Up To $1 Billion* Kenrick Cai, *Forbes*, September 7, 2022 *("Stable Diffusion — created in collaboration with RunwayML,… Coatue, and researchers at the Ludwig Maximilian University of Munich.")*

[44] *AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability,* Andy Baio, *Waxy*, September 30, 2022

AI text models typically use billions of data sources, including song lyrics or elements of those lyrics. Music generating AIs may use millions of data sources including sound recordings or particular elements of those recordings, like vocal stems. The sheer quantity of the taking makes the usage less likely to be deemed fair use, because it crowds out an increasingly larger licensing market. And if the second fair use factor considers the "amount and substantiality of the portion used in relation to the copyrighted work as a whole," then by logical implication, the theft of an entire catalog of music should be correspondingly less fair. But quantity ultimately should not drive the analysis. Even if AI models improve to the point where they can train on proportionally fewer copyrighted works, the harm to those works will remain acute and unfair.

Moreover, the fact that AI trains on all kinds of other materials that may not be copyrighted or that may, for one reason or another, be more susceptible to fair use is irrelevant to UMG's concerns. Training on music, which lies at the core of copyright protection and enables the creation of machine-made content, is inherently unfair.

8.5. Under the fourth factor of the fair use analysis, how should the effect on the potential market for or value of a copyrighted work used to train an AI model be measured? Should the inquiry be whether the outputs of the AI system incorporating the model compete with a particular copyrighted work, the body of works of the same author, or the market for that general class of works?

UMG incorporates by reference its response to Question 8 above with the following additions. First, it bears reemphasis that, separate and apart from outputs, training itself competes with the expansive market for licensing music for all kinds of technological and other uses. Because of the breadth of that market, competing generally with the licensing of all kinds of music is sufficient to defeat fair use. On the output side, AI is capable of outputs that directly compete with particular songs, with a particular artist, and with commercially exploited music in

general. All such competition is cognizable as defeating fair use under the fourth factor. Works in any genre of music might compete with works in different genres, and music companies all vie for the ears of listeners who favor multiple kinds of music. When a user elects to stream a rap song, she may be deciding not to stream classic rock. AI trained on unlicensed music intrudes into that competition by offering a free and instantly accessible alternative, and in that sense, fails the fourth fair use factor.

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)?

UMG is adamantly opposed to an opt out system, where copyright owners are required affirmatively to object to the use of their works for training. An opt out system is based on the erroneous premise that training on copyrighted works without permission is by default lawful unless each copyright owner objects. That philosophy does violence to basic principles of copyright law, imposes undue burdens on copyright owners, creates the wrong incentives for AI developers, and is neither practicable nor effective for protecting the rights of copyright owners or ensuring the sensible use of copyrighted works for training purposes.

Section 106 of the Copyright Act grants copyright owners the exclusive right "to do and to authorize," *inter alia*, the reproduction, distribution, and performance of their copyrighted works. To "authorize" means "to give official permission for something to happen, or to give someone official permission to do something."[45] The only way to grant "official permission" is for a copyright owner to express affirmative consent before a third party exploits the owner's copyrighted work. Copyright law works on this basis, and silence,

[45] https://dictionary.cambridge.org/us/dictionary/english/authorize

much less unawareness of an exploitation, cannot substitute for knowing consent. There is no equitable reason why AI companies merit exemption from this baseline requirement of copyright law.

Moreover, an opt out system would place an insuperable burden on copyright owners. Unlike, for example, trademark owners, copyright owners bear no obligation to police their copyrights. *See, e.g., Parisienne v. Scripps Media, Inc.*, 2021 U.S. Dist. LEXIS 154960 (S.D.N.Y. Aug. 17, 2021). An opt out system would impose that obligation in an unmanageable fashion. There is a growing universe of AI companies and no feasible way for copyright owners to know of them all, learn how each one operates, what training materials each one is using, how each one processes those materials, and what kinds of outputs those training materials enable. Copyright owners cannot reasonably be expected to canvas the marketplace, conduct research with limited information and resources, and determine which AI models warrant opting out. Moreover, as set forth in greater detail in response to Question 9.2, there is no reliable way to communicate opt-out information. Copyright owners will either be unable or unwilling to do the necessary research, and thus become unwitting licensors, or will alternatively opt out of all training because of the lack of information to make informed choices in any practical way. Neither is an appropriate outcome.

An opt out system also creates the wrong incentives at all ends of the spectrum. AI developers and companies have little reason to respect the rights of copyright owners if they can simply ingest copyrighted works for training in the first instance to attract investors and subscribers, and then respond on an ad hoc basis to individual objections. Copyright law does not operate on a "take now/get permission later" basis. And as noted above, copyright owners

may lack the wherewithal to object or even to devote the resources to determining whether objection is warranted. An opt out system breeds complacence rather than compliance.

Finally, an opt out system may be impossible to implement in a pragmatic way. Some AI companies have published ways to opt-out specific URLs,[46] *e.g.*, using robots.txt, but this approach is not effective, because content that is available on unlicensed sites (such as via the unlicensed datasets discussed elsewhere) will be crawled even if the content owner opts-out on sites they control. This effectively makes opting out pointless, because a content owner's content will just be obtained from sites they don't control. Even if this issue was somehow addressed, AI developers have historically only removed opted out content from *future* versions of their AI, not existing or past versions, which they typically leave up. This is partly due to the cost of retraining existing models and fact that there is currently no reliable way to untrain large swaths of content from an existing model (though, as described above, this is an active area of research).

[46] See this site for a collection of opt-out methods: https://neil-clarke.com/block-the-bots-that-feed-ai-models-by-scraping-your-website/

9.1. Should consent of the copyright owner be required for all uses of copyrighted works to train AI models or only commercial uses? For example, the generative AI model, Stable Diffusion, was reportedly developed in part by researchers at the Ludwig Maximilian University of Munich but is used by the for-profit company Stability AI. See Kenrick Cai, Startup Behind AI Image Generator Stable Diffusion Is In Talks To Raise At A Valuation Up To $1 Billion, Forbes (Sept. 7, 2022), https://www.forbes.com/sites/kenrickcai/2022/09/07/stability-aifunding-round-1-billion-valuation-stable-diffusion-text-to-image/?sh=31e11f5a24d6.

Consent should be required for all uses of copyrighted musical content for training. As noted above, even if "non-commerciality" were somehow considered relevant, there is no reliable means of ensuring that uses originally intended as "noncommercial" remain as such. The example provided in Question 9 proves as much. Moreover, datasets prepared for noncommercial purposes are subsequently repurposed by commercial AI models, and the provenance of data in those datasets is often unclear or even deliberately falsified in cases of data laundering. Even a noncommercial actor can infringe a copyright. Accordingly, the only equitable and practical system is a bright line rule that any use of copyrighted works for training material be licensed.

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

As noted above, UMG opposes any opt out system. In addition to the policy reasons to avoid such an approach, there are practical impediments that further counsel against any such system. To date, the proposed technical tools to facilitate objections are insufficient to support an opt out system. Some entities have proposed tools akin to those contemplated by this question. For example, Deviant Art, a social network for artists and art enthusiasts, has proposed a "noai" metatag that can be embedded into web pages to signal that the content they contain should not

be used for training. OpenAI has proposed data in a site's robots.txt file. However, copyrighted content is often available from sites not controlled by the content owner making such a system ineffectual. Furthermore, web crawlers can ignore robots.txt.[47] Watermarking is equally imperfect because watermarks can be stripped out of the content. Likewise, fingerprints could be used to identify content, but common content modifications will typically break this approach. The inherent technical limitations of an opt out system further militate against its adoption.

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?

We have already identified above the practical and technical obstacles to an opt out system, as well as the overwhelming policy reasons not to adopt one. Obtaining advance consent is not only feasible, but entirely consistent with the practices of the music industry, as developed further below in response to Questions 10-13.

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?

To the extent this question concerns remedies for failure to honor an objection in the opt out context, UMG objects to an opt out system altogether. More broadly, any copyright owner who objects to use of its copyright in training materials should be entitled to all remedies available for any infringement, including injunctive relief and the standard measures of monetary relief (actual damages, profits, and statutory damages).

[47] https://en.wikipedia.org/wiki/Robots.txt ("This relies on voluntary compliance. Not all robots comply with the standard…")

9.5. In cases where the human creator does not own the copyright—for example, because they have assigned it or because the work was made for hire—should they have a right to object to an AI model being trained on their work? If so, how would such a system work?

The owner of the copyright is generally the party with the right to object to infringement. However, there may be circumstances where the copyright owner is contractually bound to an author not to consent to certain kinds of uses of a copyrighted work. For example, even where UMG has entered into a catalog-wide license, certain artists and songwriters may retain the right to object to the use of the works under the terms of that license. UMG honors those rights, even where UMG is the actual copyright owner, thus protecting all parties' interests and honoring our contractual promises. Additionally, authors or performers may have other, non-copyright reasons for objecting to the use of a work as training material, such as rights of publicity or the Lanham Act. Other territories could have additional laws that are implicated, like moral rights laws.

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

As in the case of other answers, we respond to this question from the perspective of the music industry, and do not purport to speak for other industries.

As we have stated throughout this response, licensing is the only lawful and practical means to permit the use of musical content for purposes of generative AI training. Voluntary, negotiated licenses are necessary to protect the rights of songwriters, performing artists and businesses like UMG, and to ensure that AI is trained ethically and with the best possible data. UMG and the music industry generally have a long history of establishing appropriate license terms for virtually all available music used in countless contexts, so there is no impediment to

development of appropriate licensing schemes for generative AI. We shed further light on these issues in our responses to the questions that follow.

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

Yes. The licensing of music for training generative AI is not an insurmountable administrative task or an undue burden on the AI community. Any contrary belief is grossly mistaken. Licensing is the primary model by which the music industry does business, in connection with both traditional markets and the newest technologies. Indeed, UMG has been a leader in establishing business models for cutting edge technologies, which it confronts on a regular basis. There is no reason to believe that generative AI is not susceptible to appropriate licensing, just as there is no equitable basis for exempting AI companies from licensing musical content.

Several examples prove how effectively market-based licensing has tackled revolutionary technologies. Streaming music and social media are relatively new phenomena that took hold over the past decade. Despite this seismic takeover of music consumption by previously unknown technologies, the entire music industry managed in short order to license virtually all of its content to all of the major streaming music services, including Apple Music, Amazon Music Unlimited, Deezer, SoundCloud, Spotify, Tidal, TikTok, YouTube, and many others. In fact, Universal Music Group has music licensing arrangements with hundreds of different digital music businesses worldwide.

The sampling marketplace also stands out as a prime example of free market music licensing with parallels to generative AI.[48] In order to use a portion of a musical work and/or sound recording as an element in a new work, permission must be granted by the copyright owners of the sample being used. Today's sampling and interpolation licensing arrangements are each bespoke and built to work for creators and rightsholders alike. The industry successfully manages and transacts a multitude of these licenses that further the creation of new music while also ensuring the owners, writers, and artists whose samples are being used are rightfully respected and compensated. Just as the sampling marketplace has provided artists with lawful component materials upon which to create new music, free market licensing agreements can grant AI developers permission to use musical works and sound recordings as training material.

The advent of user-generated content on YouTube provides another prime example of how free market arrangements can address copyright concerns. In order to manage the flood of unauthorized content posted by YouTube users, the music industry collaborated with YouTube to develop the Content ID system and associated license regime, pursuant to which rightsholders license their copyrighted content to YouTube and elect whether to be compensated for user postings or to remove those postings. The system works by having rightsholders deliver their copyrighted works and identifying elements of those works to YouTube's system, which in turn uses acoustic fingerprints and other technologies to determine when a user-uploaded video

[48] For further explanation of the sampling and interpolation marketplaces, see https://www.copyright.gov/music-modernization/educational-materials/Sampling-Interpolations-Beat-Stores-and-More-An-Introduction-for-Musicians-Using-Preexisting.pdf and https://www.findlaw.com/smallbusiness/intellectual-property/obtaining-permission-before-sampling-music.html#:~:text=The%20music%20publisher%20is%20typically,public%20performances%20of%20artists%27%20music.

incorporates any of the copyrighted content and notifies the copyright owner. Again, despite the extraordinary breadth of the uploaded content – it has been reported that 3.7 million videos are uploaded to YouTube every day [49] – and the rapid development of the technology and the public's adoption of it, YouTube and the music industry were able to work out a mutually beneficial system that protects rightsholders without impeding the development of a key music platform.

The same can and will be true for generative AI. UMG is ready and willing to explore licensing content to generative AI companies and has already established our disposition to work with those companies to find ways to build fair, lawful, and ethical business models that make economic, strategic, and long-term sense for our artists and songwriters. However, two challenges currently impede progress: (1) the need for the AI community to have a reciprocal willingness to negotiate licenses with terms that respect and compensate creators, and (2) the lack of established and equally applied "generative AI rules of the road" to ensure an ethical, responsible, and trustworthy ecosystem. Pending litigation and ambiguity in the law also inhibit licensing. It took years of litigation to establish a robust digital music market via disputes with Napster, Grokster, Grooveshark, and Limewire, for example. There is no need to force the industry through that prolonged period here as the law clearly requires licensing now and imposing a bright-line licensing requirement will provide incentives for AI developers to work with rightsholders to create pragmatic and equitable solutions that benefit all.

[49] https://www.wyzowl.com/youtube-stats/ - :~:text=Around 3.7m new videos,average length of 4.4 minutes

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 collective licensing scheme is not an appropriate solution to the use of musical content to train generative AI. Given its broad experience with licensing, UMG believes that direct and individually negotiated licenses are the best path forward for the music industry. UMG's role is to protect the rights, interests and integrity of its artists and songwriters, who may have different views, preferences, and rights regarding the use of their music – and their personas – in the generative AI space. Generative AI requires the creativity, rapid response, and adaptability inherent in free market licensing. UMG can best meet these interests through direct licensing, rather than delegating those licensing duties to a collective licensing organization.

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

UMG strongly opposes the imposition of a compulsory license. Like an opt out system, a compulsory licensing scheme would prejudice the rights of copyright owners – as well as artist and songwriter contributors to the copyrighted works – without any corresponding benefit. Among other things, a compulsory license regime would limit rates rightsholder can negotiate, force copyright owners to allow the copying of their works for uses they could neither control nor anticipate, initiate costly rate setting litigation or administrative proceedings, and require the creation of a needless and complicated infrastructure to administer and police any such regime. And like an opt out system, the suggestion of a compulsory licensing scheme flows from the

faulty premise that generative AI should be exempt from the obligation, born by the rest of the market, to secure specific consent to the use of copyrighted materials.

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

Extended collective licensing is a highly *un*desirable approach. As defined by footnote 51 of the NOI, such a scheme would delegate licensing to a central licensing body with rights to given classes of copyrighted works for all rightsholders, regardless of whether those rightsholders have granted specific permission. Such a system has all the defects of an opt-out or compulsory licensing scheme, in that it essentially forces rightsholders to grant licenses without specific knowledge or permission or any control over the scope of the license granted. Extended collective licensing makes little sense when UMG (and other rights holders) already has the means and infrastructure in place needed to engage in the necessary licensing activity. In short, extended collective licensing diminishes the rights of copyright owners needlessly, particularly where the alternative of direct, individual licensing is both feasible, more efficient, and more equitable.

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

UMG's comments are confined to the music industry, where free-market licensing has enabled the industry to be readily adaptable to new technologies and has resulted in the widespread availability of music to consumers across myriad 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)?

As noted above, the music industry is adept at licensing its content in huge quantities for countless different uses, so UMG does not anticipate legal, technical, or practical barriers to granting licenses for use of its content for training purposes. We encourage AI developers who respect music and those who create it to join us in building a future that benefits us all. The negotiations simply have to begin.

All the participants in the training process must have appropriate authorization, but that authorization can be granted in different ways. For example, AI platforms can license musical content directly from the rightsholders. Alternatively, curators of training datasets can license musical content with rights to sublicense the content to AI companies for training purposes with appropriate safeguards and limitations. For example, any authorization to use musical content for training purposes might include durational limitations or limitations on the kinds of outputs than can be enabled by the training materials. Songwriters may license their lyrics for use in training subject to restrictions against using those lyrics to create infringing outputs or combining their lyrics with other, third-party lyrics. Owners of sound recordings might license their recordings for training subject to restrictions against using those materials to create soundalikes of performing artists. The key is that any usage conforms to the scope of the license. This approach is consistent with how the music industry operates, in that a license for one purpose (such as inclusion of a recording in a compilation album) would not grant rights for another purpose (such as inclusion in a movie soundtrack).

12. Is it possible or feasible to identify the degree to which a particular work contributes to a particular output from a generative AI system? Please explain.

As matters currently stand, the means to determine whether particular works used for training contributed to particular outputs are limited and often ineffectual, with some exceptions. When a generative AI system yields outputs that facially incorporate all or elements of copyrighted works, that system has trained on those works. For example, a system that can generate the lyrics to a copyrighted song must have trained on that song. Likewise, an output that creates a faithful soundalike of a celebrity singer's voice would have trained on recordings or vocal stems containing that singer's voice. There are other means of deducing what works contributed to what outputs, such as analyzing what words are used in prompts (*e.g*. artist, track names or lyrics), similarity searches examining the AI output (identifying what human created songs are closest to the output), or similarity searches in the embedding space (a kind of intermediary representation in the AI of prompts and the content they can create). Over time, we expect new means of attribution will be discovered or developed.

Importantly, this opacity is why a bright-line prohibition on the use of copyrighted works as training materials without authorization is paramount. If copyright is respected when training a generative AI, and developers are transparent about what has been used in their training sets, ensuring credit and compensation on the outputs is a matter of technology and licensing arrangements. In other words, if copyright is respected in the AI ingestion process, the remaining issues raised in this NOI are exceedingly easier to solve.

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

The cost of obtaining licenses is a necessary expense for any business seeking to use copyrighted works, and music is no different. Such economics support the Copyright Act's and the Constitution's purpose of rewarding creative and artistic expression. Imposition of that cost on those who wish to exploit copyrighted music has not slowed or impeded the growth of new technologies, and artificial intelligence is no exception. To put things in perspective, music streaming services generated $43.3 billion revenue in 2022, a 15% increase over 2021.[50] Licensing has fueled that growth by supplying those platforms with the greatest music in the world – and the expertise, metadata, and flexible arrangements that make it a value proposition for technology companies, whether they be start-ups or global behemoths.

Generative AI should follow the same path, with any expense of licensing far outweighed by the benefits of partnering with the music industry, which best understands the provenance and value of the music it controls and can supply musical training materials in the best and most reliable formats. Simply asking the AI community to take the time necessary to license music is an appropriate and necessary "cost" that far outweighs the appropriation of an entire artform without permission or compensation. Other legitimate businesses that use copyrighted works en masse bear those licensing costs and have each negotiated agreements that fit their particular needs. It would be unjust to relieve the AI industry from that same responsibility. Moreover, many of the key players in the AI industry are huge companies that should have little difficulty absorbing this necessary expense.

[50] https://www.businessofapps.com/data/music-streaming-market/

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

Copying without permission is forbidden by the Copyright Act, and in every other industry or context, those who use copyrighted materials license them. That generative AI is new, has fantastic potential for dynamic and beneficial applications, and can produce non-infringing outputs should not exempt generative AI from the basic workings of copyright law. The practical and equitable solution for a technology with so much potential married to so much potential damage is to require licensing of training materials in the first instance. As opposed to a piecemeal fight in the courts, with unpredictable and potentially inconsistent results as well as wasted resources, a system of licensing provides a bright line, equitable model born of partnership, not opposition. In UMG's view, it is the only practicable and fair solution.

TRANSPARENCY

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

Yes. As stated in response to Question 5, without a comprehensive and transparent system of recordkeeping regarding training data, rightsholders face an untenable "Catch-22." When AI training sets are opaque, rightsholders must initiate expensive litigation and pursue discovery to determine whether their content was used in data training sets, and yet without such evidence, rightsholders cannot even initiate litigation in the first place. Legislation requiring that record of materials used to train AI algorithms be made available to the public would remedy this unreasonable burden on enforcing one's copyright.

Additionally, the use of copyrighted and other protected materials in training must be licensed to comply with copyright law. Appropriate recordkeeping will require AI developers to be thoughtful and selective in the content they use for training purposes – and will encourage them to seek appropriate authorization in the first instance. Licensing is easier and more efficient if training data is transparent and detailed. Consumers benefit as well, as they may want to know if content was AI generated and if so, whether it was authorized by the people on whose content it trained or whose likeness it mimics. The answers to these questions can only be determined if training information is available.

For these reasons, UMG proposes a statutory requirement that AI developers be required to maintain accurate records of all materials and information used to train their model. Because recordkeeping is important beyond the scope of copyright considerations, a training material recordkeeping requirement could easily exist outside of Title 17.

15.1 What level of specificity should be required?

Records articulating training materials should be sufficiently detailed and understandable to a layperson: URLs detailing the source of training materials should be attached to specific and detailed descriptors of individual ingested material (but a mere list of URLs should not be deemed acceptable). When a training set is comprised of musical works and sound recordings, those works should have associated metadata that contains all relevant information about the material and its owner(s). Because it is unlawful to remove copyright management information (*e.g.*, metadata), it should not be burdensome to require that metadata for all training materials be maintained. Importantly, records of training materials should be easily searchable, particularly given the vast amount of data in many training sets.

15.2 To whom should disclosures be made?

Records regarding training materials should be available for public inspection.

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

The need to be transparent about the materials used to train an AI model does not dissipate if a developer uses a model created by someone else as the basis of their system. The reporting requirement should flow through to the entities who make the product available to end users, meaning an entity who uses a third-party AI must be able to provide access to the list of training material associated with the third-party models they use.

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?

As noted above, in all likelihood, some of the metadata associated with training data used by AI developers is already being stored, because it is needed during the training, refining and re-training processes. For a training set containing 5 billion pieces of content, assuming 1kbyte of metadata per entry, that amounts to roughly 5 terabytes (TB) of storage. Currently, a 5TB LaCie drive is available from Amazon for under $170.[51] Additionally, many AI developers are highly capitalized and have billions of dollars of investments from some of the biggest names in technology. They certainly can bear the cost of responsible recordkeeping.

[51] https://www.amazon.com/LaCie-Rugged-External-Drive-Portable/dp/B07XQXF444

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

As noted earlier, copyright law dictates that consent from rightsowners is required before copyrighted content may be used as training material. Once consent has been granted, notification has already occurred.

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

Consistent with other laws regarding the retention of records, AI developers should always maintain records regarding provenance of the materials they used for training purposes. The federal government has enacted or imposed a number of recordkeeping requirements on businesses, as have the states. Employers are required by law and regulation to retain a number of various records regarding their employees,[52] the Department of Education requires educational institutions to maintain records related to financial aid,[53] financial institutions must keep records in the interest of detecting money laundering,[54] and midwives in New Jersey,[55] chiropractors in Kentucky,[56] and talent agents in California[57] must do so as well. An internet search returned over 1,000 different recordkeeping requirements imposed by governments on businesses and other entities.[58] It is absolutely reasonable to impose recordkeeping requirements on those in the generative AI space as well.

[52] https://www.macpas.com/wp-content/uploads/2019/02/Federal-Record-Retention-Requirements.pdf
[53] https://www.ecfr.gov/current/title-34/subtitle-B/chapter-VI/part-668/subpart-B/section-668.24
[54] https://www.fincen.gov
[55] https://www.law.cornell.edu/regulations/new-jersey/N-J-A-C-10-58-1-9
[56] https://www.law.cornell.edu/regulations/kentucky/201-KAR-21-100
[57] https://www.dir.ca.gov/dlse/talent/talent_laws_relating_to_talent_agencies.pdf
[58] https://www.law.cornell.edu/search/searchResultsForm.html

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?

UMG does not believe that a human being using a generative AI system can qualify as the author of outputs of that system, because the output is never copyrightable. Authorship requires human expression, and the output of generative AI systems does not embody human expression. As noted in the NOI at 1 n.1, generative AI systems produce outputs based on learning statistical patterns in existing data. When generative AI responds to a human text prompt, it draws upon its analysis of those statistical patterns to create something that matches or "autocompletes" the human prompt, but it is the AI system that creates the output, not a human being. Whatever the nature of the output – text, music, image – its expressive content was created and arranged by a computer making statistical predictions rather than creative choices.

In this regard, it is important to distinguish between the potential copyrightability of a prompt and the output it might generate. The former might be copyrightable, but the latter is not. Many prompts may fail the test for copyrightability because they amount to no more than a non-protectible idea. Instructing a generative AI system to create a "poem about pine trees" or an image of a "lion playing an electric bass" provides only noncopyrightable concepts.

However, a prompt may cross the threshold of copyrightability to the extent it satisfies the minimal standards of originality and creativity set forth in *Feist Publications, Inc. v. Rural Telephone Service Co*., 499 U.S. 340 (1991). Imagine a two-page textual treatment laying out the plot, characters, and setting of a short story. That treatment would likely qualify for protection under traditional principles of copyright law, and if a screenwriter were authorized to

base a full-length movie on it, the resulting movie would likely qualify as a derivative work. However, if the author of the treatment used it as a prompt for a generative AI system and instructed the system to create a movie based on the treatment, everything added by the AI system would be noncopyrightable. The added "creative" aspects of the movie would be executed by a computer, so the only actual copyrightable work would be the original treatment. Of course, the copying of the AI-generated movie might nevertheless infringe the underlying, copyrightable treatment under basic principles of copyright law.

The same would hold true for a musical prompt, even if it, for example, supplied detailed instructions about a musical composition, including a textual description of its melody, harmony, and rhythm. If the AI system executed those instructions exactly, it would add nothing to any copyright inherent in the prompt, and if the AI system departed in any way from the instructions, its additions or modifications would be generated by a computer and thus not copyrightable. To return to the wording of this question, even if the user of the system were to provide an "iterative series of text commands," anything other than identical implementation of those text commands would be machine-generated and not the product of human authorship.

The actual experience of using generative AI confirms that the system makes what resembles the "creative choices" in the output, not the human prompter. For example, text prompts for OpenAI's DALL-E image generator produces multiple simultaneous outputs, each of which typically embodies the concepts of the prompt but expresses those concepts differently with different images. Likewise, serially prompting ChatGPT with the same text prompt – "a poem about an aardvark" – produces markedly different results every time. The fact that the AI system executes the same prompt in multiple different ways, with dissimilar expression, proves empirically that it is the AI system that is making the choices that control the output, not a

human being. And efforts to make the prompts more and more detailed do not change the basic analysis – anything that transcends the prompts is produced by a computer program.

The Copyright Office recently endorsed this view that AI-generated outputs cannot command copyright protection in its September 5, 2023 refusal of the Second Request for Reconsideration for Refusal to Register Théâtre D'opéra Spatial (SR # 1-11743923581; Correspondence ID: 1-5T5320R). *See* https://copyright.gov/rulings-filings/review-board/docs/Theatre-Dopera-Spatial.pdf. The applicant sought to register his edited version of a "two-dimensional artwork" generated by Midjourney – a popular image-generator – without disclaiming those elements that were created by AI. In arguing that he was an author of the Midjourney image, the applicant protested that he "input numerous revisions and text prompts at least 624 times to arrive at the initial version of the image," and that his "creative input" into the Midjourney image included "enter[ing] a series of prompts, adjust[ing] the scene, select[ing] portions to focus on, and dictat[ing] the tone of the image." His series of prompts began with a "big picture description" that "focuse[d] on the overall subject of the piece," followed by prompts allegedly dictating "the overall image's genre and category," "certain professional artistic terms which direct the tone of the piece," "how lifelike [he] wanted the piece to appear," a description of "how colors [should be] used," a description "to further define the composition," "terms about what style/era the artwork should depict," and "a writing technique that Mr. Allen has established from extensive testing" that would make the image "pop." *Id.* at 6-7. In other words, he provided the kind of "iterative series of text commands" that Question 18 specifically identifies for analysis.

The Copyright Office correctly rejected this claim of authorship in the underlying image:

> In the Board's view, Mr. Allen's actions as described do not make him the author of the Midjourney Image because his sole contribution to the Midjourney Image was inputting the text prompt that produced it. Although Mr. Allen describes "input[ing] numerous revisions and text prompts at least 624 times" before producing the Midjourney Image …the steps in that process were ultimately dependent on how the Midjourney system processed Mr. Allen's prompts. According to Midjourney's documentation, prompts "influence" what the system generates and are "interpret[ed]" by Midjourney and "compared to its training data." … As the Office described in its March guidance, "when an AI technology receives solely a prompt from a human and produces complex written, visual, or musical works in response, the 'traditional elements of authorship' are determined and executed by the technology—not the human user."

Id. at 6-7. As the Office then explained, whether the prompts themselves are copyrightable was not the issue – some prompts could certainly qualify – because the output was ultimately determined by software. *Id.* at 7. This analysis crystallizes UMG's view on this issue. No matter how detailed the prompts, or even whether they independently qualify for copyright protection, the "expressive" elements of outputs are ultimately determined by a machine, not a human being and accordingly cannot embody human authorship and are not copyrightable.

UMG also believes that selecting training data cannot elevate the party selecting that data to an author of any resulting output. First, in practice, training material for a machine-learning based system is selected by categories of content rather than individual items of content. For example, developers do not hand select each of the billions of images used for an image generator, but rather capture categories of content. Individual selection would defeat the purpose of generative AI to train on vast swaths of information to create the best predictive generative model. Moreover, as in the case of a prompt, while training material may influence an output, that output remains the product of machine generation, not human authorship.

Ultimately, the question of human authorship for the outputs of generative AI is controlled by the concepts first announced by the Register of Copyrights (and quoted in the NOI) back in 1965:

> The crucial question appears to be whether the "work" is basically one of human authorship, with the computer merely being an assisting instrument, or whether the traditional elements of authorship in the work (literary, artistic, or musical expression or elements of selection, arrangement, etc.) were actually conceived and executed not by man but by a machine.

NOI at 2 (quoting U.S. Copyright Office, *Sixty-Eighth Annual Report of the Register of Copyrights for the Fiscal Year Ending June 30, 1965*, at 5 (1966). As noted above, certain kinds of artificial intelligence can serve as an "assisting instrument" to a musician, such as the studio technologies that help refine sound recordings. However, as currently constituted, the generative AI that is the subject of this NOI "conceives and executes" the traditional elements of authorship, regardless of the training materials or prompts humans put into the system. Accordingly, outputs are uncopyrightable and thus not susceptible to human authorship.

Finally, UMG recognizes that humans can add authorship to previously generated AI outputs. Thus, if an AI system renders an image or a set of lyrics or a musical composition, a human being could alter or modify that output by adding expressive, copyrightable content. For example, a musician could create an arrangement of an AI-generated musical composition, just as the same musician could create an arrangement of a public domain composition. In both cases, the traditional rules of authorship would govern and might accord copyright protection to the new materials added to the underlying, unprotectible work.

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?

UMG believes that the requirement of human authorship for copyright protection is well-established. UMG agrees with the exhaustive analysis set forth by the Copyright Office in the *Thaler* matter, as subsequently affirmed by a federal district court. *See Thaler v. Perlmutter*, 2023 U.S. Dist. LEXIS 145823 (D.D.C. August 18, 2023), *affirming*, U.S. Copyright Office Review Board, *Decision Affirming Refusal of Registration of A Recent Entrance to Paradise* at 2 (Feb. 14, 2022). Accordingly, no revision to the Copyright Act is necessary.

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?

There is no policy advantage to affording copyright protection to AI-generated content. Copyright law rewards human authorship and provides incentives for human creativity; computer programs do not respond to incentives. Moreover, granting copyright protection to AI-generated content is inimical to the policies of copyright law and the underpinnings of the music industry because it would threaten to displace human authorship. As noted above, AI-generated musical outputs compete with genuine music created by real people. AI can generate generic music, canned lyrics, and even imitations of a human performer's voice or style, all to the detriment of real musicians who make a living presenting their genuine musical talents.

It has also been observed that generative AI outputs are inherently creatively limited compared to human expression. Generative AI relies on existing copyrighted works to create statistical associations that enable the generation of things that look or sound like those existing works. It essentially tries to resemble human expression based on past expression. Human

beings, on the other hand, have the ability to transcend past expressive works and create new forms of expression, new artistic genres, and new works of music and art never previously imagined. Copyright law is designed to promote that kind of artistic innovation, rather than to reward a computer program that is by its very nature limited to what already exists artistically.

There are ample existing means to encourage development of generative AI. Developers can claim copyright protection in the code that operates generative AI. They can obtain patents in their AI technologies (as UMG itself has) and can rely on trade secrets protection to ensure exclusivity in their technologies. AI companies have multiple additional means of building value without legal protection for AI outputs. They can charge users for accessing AI tools (Stability AI, Adobe, and many others do today) or for accessing AI-generated content (as in the case of, for example, Endel). The value is not in the individual outputs but access to the generative AI system. The breathtaking investment in AI proves that no additional incentives are needed.

20.1. If you believe protection is desirable, should it be a form of copyright or a separate sui generis right? If the latter, in what respects should protection for AI-generated material differ from copyright?

As stated above, no protection is desirable.

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?

UMG's answer to this question is closely related to its answer to Question 20 above. Copyright protection for AI-generated content runs contrary to the policies and purposes of the Copyright Clause. Justice Ginsberg described the purposes of the Copyright Clause as follows:

> As we have explained, "the economic philosophy behind the [Copyright] Clause . . . is the conviction that encouragement of individual effort by personal gain is the best way to advance public welfare through the talents of authors and inventors." *Mazer v. Stein*, 347

> U.S. 201, 219, 98 L. Ed. 630, 74 S. Ct. 460, 1954 Dec. Comm'r Pat. 308 (1954). Accordingly, "copyright law celebrates the profit motive, recognizing that the incentive to profit from the exploitation of copyrights will redound to the public benefit by resulting in the proliferation of knowledge . . . . The profit motive is the engine that ensures the progress of science." *American Geophysical Union v. Texaco Inc*., 802 F. Supp. 1, 27 (SDNY 1992), *aff'd*, 60 F.3d 913 (CA2 1994).

Eldred v. Ashcroft, 537 U.S. 186, 212 n.18 (2003). The Copyright Clause, in other words, is designed to encourage "individual effort by personal gain," so that the public gains access to the fruits of the "talents of authors." By financially rewarding authors, copyright law promotes further creativity that ultimately redounds to the public's benefit.

Granting protection to AI-generated content does the exact opposite. Rather than rewarding authors and encouraging them to continue creating, such protection would encourage the proliferation of machine-generated content that can supplant human authorship. The more the market is flooded with AI-generated content, the less human authors have the economic incentives to continue composing and recording music to benefit the public. Rather than "promoting" creativity and the public benefit that it confers, granting copyright protection to AI-generated content would stifle creativity and reduce the store of creative works available to the public. The Copyright Clause, accordingly, forbids rather than allows granting copyright protection to such works.

UMG notes that AI developers themselves benefit from copyright's incentives. Their creative works are the AI programs and models they develop. They earn copyright protection for the code that embodies those creations and thus enjoy the same incentives that composers and recording artists do. Granting protection to AI outputs, however, strays far from this appropriate allocation of incentives.

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?

There is no question that AI-generated outputs can and often do violate copyright owners' exclusive rights, including without limitation the reproduction and derivative work rights. For example, many of the AI systems generate identical or nearly identical replicas of copyrighted lyrics. The same systems also create derivative works based on copyrighted lyrics. They will combine existing copyrighted lyrics with AI-generated lyrics or combine the lyrics of two different songwriters into a single song. The only reason they are able to create these infringements is because they have trained on the underlying copyrighted works.

More broadly, it is widely known that generative AI models can reproduce all kinds of copyrighted works based on the materials on which they were trained, including copyrighted text, artwork, photography, musical compositions, and sound recordings (which are subject to a different standard for infringement, as discussed in response to Questions 5 and 23). This is known as "memorization" – the AI model "memorizes" the training materials and then reproduces them identically or in some recognizable form as a derivative work. Researchers have written about this phenomenon and the associated copyright jeopardy faced by these models.[59]

It is not possible to anticipate all the prompts that will cause an AI system to violate the reproduction or derivative work rights. For example, UMG has observed that its copyrighted lyrics will be reproduced in response to various types of prompts, ranging from those that specifically ask for the lyrics to a given song to prompts simply asking for lyrics about a

[59] *See, e.g.*, *AI Spits Out Exact Copies of Training Images, Real People, Logos, Researchers Find*, Chloe Xiang, Vice, February 1, 2023

particular topic. There are likely countless other types of prompts that will result in an infringing output. The reality remains throughout that these systems are able to reproduce and infringe copyrighted works because have trained on those 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?

Answering this question requires a three-part analysis. First, as set forth above, training on copyrighted works, including sound recordings and musical compositions, necessarily infringes the right of reproduction. Assessing infringement at the training stage does not require resort to the substantial similarity test, because the AI systems make an actual copy of the underlying copyrighted work itself. The nature of the output – and whether it is infringing or not – does nothing to exonerate the initial infringement occasioned by training. Thus, training materials must always be licensed.

With respect to "outputs from a generative AI system," the substantial similarity test should continue to provide the appropriate test for infringement. Whether in the form of text, images, or music, AI-generated content is designed to imitate human forms of expression. Whether it infringes may thus be assessed by the same standards courts use for human expression – does AI-generated output appropriate substantial protectible expression from the copyrighted work? A set of lyrics or a melody that is substantially similar to a copyrighted work will infringe, regardless of whether those lyrics or that melody was created by a human being or generated by an AI system that trained on the copyrighted work. We have ample jurisprudence on substantial similarity to manage the task of evaluating AI-generated content for infringement.

Finally, as discussed above in response to Question 5, in the case of sound recordings, a finding of infringement should extend to AI-generated sound recordings that either incorporate

the actual sounds of a copyrighted sound recording or are substantially similar to a copyrighted sound recording. In contrast to all other copyrightable subject matter, for which substantial similarity is the governing test, section 114(b) of the Copyright Act limits the reproduction and derivative work rights of sound recordings to copies that capture the "actual sounds" embodied in the copyrighted sound recordings. While there may have been historical policy reasons or perhaps a legislative compromise that imposed this limitation on liability for human efforts to imitate sound recordings, that limitation no longer makes sense with respect to AI-generated content, which can create sound recordings with uncanny speed and accuracy. Indeed, AI-generated soundalikes and deep fakes have attracted particular outrage in the artistic community because they intrude so unfairly on genuine human artistry. Accordingly, as noted above, Section 114(b) should be amended to expand the scope of sound recording copyrights to proscribe AI-generated recordings that either capture the actual sounds of or are substantially similar to copyrighted sound recordings on which the AI was trained.

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 matters now stand, without access to records of training materials, copyright owners can only establish copying anecdotally, where outputs are sufficiently similar to copyrighted works that initial reproduction of those works in the training stage may be inferred. Locating outputs that are facially similar, moreover, means a trial-and-error process of inputting prompts until one fortuitously produces a facially infringing output. Attempting enforcement through that kind of investigation, where AI systems are so variable in how they respond to prompts, is inefficient, unfair, and a burden copyright owners should not have to face. And, where no

facially infringing output reveals itself at all, the copyright owner is left to guess whether the copyright was used or not. As explained above, copyright owners are left in an insoluble dilemma – they may not be in a position to file suit and yet need the tool of civil discovery to determine whether their works were copied in the first place. And even when they file suit, discovery is a blunt, expensive, and time-consuming instrument. The expense of discovery alone will almost certainly discourage some copyright owners from pursuing meritorious claims of infringement.

This state of affairs is neither rational nor equitable and lays bare the acute need for transparency and recordkeeping, as discussed elsewhere in this response. In what other industry can a business so easily hide its unauthorized exploitation of *millions and even billions of copyrighted works* with practical impunity? And in what other context is simply detecting infringement so disproportionately difficult? Keeping the public in the dark about training sets serves only to mask infringement, with no corresponding public benefit. Accordingly, AI companies must maintain and make readily available identifying information about the works on which their systems were trained.

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?

Traditional principles of direct and secondary liability should govern who is liable for infringing outputs. Because an infringing output will result from training on the copyrights that have been infringed, any party who either directly reproduced that copyrighted work in training, or who materially and knowingly contributed to that reproduction or induced it, would be liable in the first instance. Such parties could include the owner of the dataset who provided training

data including the copyrighted work to the AI developers, the developer of the AI model, the developer of the system incorporating that model, and any independent contractors who assisted any of the foregoing.

Likewise, an end user who prompts an AI system to generate an infringing output is potentially liable. For example, if a user prompts an AI system to produce the lyrics or melody of a copyrighted song – "give me the melody and lyrics to Elton John's 'Goodbye Yellow Brick Road'"– that user has knowingly contributed to or induced the system to create an unlawful reproduction. And if the user then further reproduces or distributes that output, the user will be a direct infringer. For example, if a user asks an AI system to create an infringing image and then imprints those images on t-shirts for sale, the user has reproduced and distributed infringing copies and is directly liable. Given the variety of possible scenarios, it is not possible to anticipate all bases for liability, but the law of infringement is sufficiently developed that it will supply the governing principles.

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

The fact that an AI model is open source does not fundamentally change the infringement analysis. "Open source" typically means that the model's code is publicly available, so that third parties may use that code and potentially create infringing outputs. Whether that free availability of code supports a claim of secondary liability will depend upon the particular facts – *e.g.*, does the model's developer know that others are using the code to generate infringing content? But otherwise, the basic law governing infringement should not change.

26. If a generative AI system is trained on copyrighted works containing copyright management information, how does 17 U.S.C. 1202(b) apply to the treatment of that information in outputs of the system?

As in the case of copyright infringement, outputs should be subject to the same standards as any other works with respect to the removal, alteration, or falsification of copyright management information. No special treatment for AI-generated content is required.

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

UMG reiterates here what it states throughout. The complexities and uncertainties about copyright liability for infringement of musical content in the AI context can be substantially reduced by imposing a bright line rule that training materials must be licensed. Those licenses will authorize the initial reproduction of copyrighted works for training, and the scope of those licenses will be germane to assessment of whether given outputs infringe. Dealing with copyright liability only at the output stage will yield an unpredictable patchwork of case-by-case determinations that will not provide the necessary guidance or predictability.

LABELING

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?

Content purely generated by AI should be labeled as such, particularly when there is a chance of harm. Unfair, deceptive, or fraudulent materials have long been restricted in the interest of protecting consumers, and the use of generative AI does not negate the fundamental intent of consumer protection law. The higher the potential for consumer deception, the greater the need for content to be labeled.

Beyond the need to protect against deception, a label identifying content generated by AI could be a helpful tool for consumers. For example, music fans who desire authenticity and human connection may want a label in order to help them make informed choices to support those creators who do not use AI. However, if an artist elected to use elements of generative AI in the creative process, their fans may also want to know it was used with the artist's consent and/or at their direction.

In addition, given that elements of content generated by AI are not eligible for a copyright, it is essential that proper labeling and metadata be used throughout the ecosystem to ensure content created by AI does not unjustly siphon compensation from human creators.

Of course, any government labeling requirement should impose responsibility across the entire ecosystem. And labeling needs and the format of any label could vary across content type and the associated and inherent risk of a particular generative work. The burden imposed by any labeling system should also be weighed against its potential need and benefits.

Finally, it is important to consider whether there is any harm if AI information is *not* disclosed. Potential use cases should be evaluated to understand (1) the purpose of the label; (2) what label information is required and how it should be communicated, and (3) who benefits from the information provided. Consider these three different scenarios:

- An FM radio station uses generative AI to create a station identification to be broadcast over the air such as "You are listening to KAI 106.7." If this audio is generated using an AI algorithm that was only trained on legally obtained content – and the AI voice is not trained on an identifiable human voice used without their consent – arguably, it should not need to be "labeled" since it is not readily apparent what the purpose of such a "label" would be. This content would not be eligible for copyright or reside in an economic value chain. It would not necessarily be deceptive or harmful to consumers. A station identification is not something a consumer "chooses."

- A consumer is looking to buy or stream music and wants to support only human-created art. That consumer choice is only possible if music generated by AI (and/or human created music) is labeled as such. And a consumer could be deceived or misled if no label exists.

- A recording artist authorizes their voice to be used to train generative AI for the purpose of making backing vocals on a new song. Here, the consumer may be interested to know the backing vocals were generated by AI, but also that the artist authorized the use of their voice to train AI for that purpose and wants fans to hear the song.

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

In many (if not most) cases, the only entity who will have direct knowledge that a piece of content has been generated by AI is the AI developer whose tool created it and/or the user who employed the generative AI algorithm.

As the Office is aware, proper data in the music industry is the fuel that makes the music consumption engine run: The *earlier* proper metadata is attached to creative content, the better.

Regardless of who identifies a work as AI-generated, it will be critical that there is an obligation for entities farther down the distribution chain to maintain the integrity of that data and ultimately display it to end users.

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

Technological solutions are nascent, but market demand should accelerate scalable technological solutions that enable accurate labeling. However, while standards for labeling and, ultimately, enriched metadata will certainly help, one simple solution available today would be to require inclusion of some additional information (such as "AI") in the content's description.

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 party who fails to label as required – or removes a label – should be fined. The most logical enforcement body is the Federal Trade Commission. In particularly egregious instances, a private right of action should be available to an aggrieved party.

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?

This is an active area of research and development and numerous companies and researchers working on tools to identify AI-generated content. A more reliable approach may be identifying human created content.

Also relevant are standards to carry provenance data. UMG is a member of the Content Authenticity Initiative (CAI),[60] a group of "media and tech companies, NGOs, academics, and others working to promote adoption of an open industry standard for content authenticity and provenance." The Coalition for Content Provenance and Authenticity (C2PA),[61] comprised of a group of leading technology companies, is a related group working on the development of technical standards certifying authenticity of content. However, it may be faster to add AI-related data to existing and widely adopted industry-specific standards, such as DDEX in the case of the music industry.

[60] https://contentauthenticity.org/how-it-works
[61] https://c2pa.org/

Additional Questions About Issues Related to Copyright

RIGHT OF PUBLICITY

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?

Generally, the unauthorized use of an individual's name, image, likeness, voice, and other personally identifiable characteristics is governed by state law rights of publicity. Currently, about half the states have right of publicity statutes, while many states recognize those rights at common law. Most states recognize rights of publicity for living individuals, whereas only some states recognize such rights for deceased individuals, with varying limitations on the duration of postmortem rights. Different states also apply different choice of law rules. The state law rights of publicity, whether statutory or common law, are a patchwork, which is one of the reasons UMG advocates for a federal right of publicity in the context of generative AI that establishes a meaningful right available to all.

While state right of publicity laws thus differ in material ways, each broadly allows an individual to control the commercial use of his or her "persona," which may include not only the individual's name and likeness, but also look-alikes and even individual's signature phrases. *See, e.g.*, *Onassis v. Christian Dior-N.Y., Inc.*, 472 N.Y.S.2d 254 (Sup. Ct. N.Y. County 1984) (pictorial advertisement with Jacqueline Kennedy Onassis look-alike constituted a "portrait" or "picture" of plaintiff under New York right of publicity statute); *Hilton v. Hallmark Cards*, 599 F.3d 894 (9th Cir. 2010) (Paris Hilton had cognizable right of publicity claim against Hallmark for use of her catchphrase "that's hot").

State common law rights have also been applied to "soundalikes" – imitations of the voice of recognized performers. For example, in *Midler v. Ford Motor Co*., 849 F.2d 460 (9th Cir. 1988), after failing to entice Bette Midler to sing for a TV commercial, Ford Motor hired a singer to imitate Midler's voice singing a song closely associated with the star and used the recording in the commercial. The Court reversed a grant of summary judgment in defendant's favor and held that Midler could pursue a claim under the common law right against "appropriation of the attributes of one's identity." Noting that "a voice is as distinctive and personal as a face," the Court explained that "when a distinctive voice of a professional singer is widely known and is deliberately imitated in order to sell a product, the sellers have appropriated what is not theirs and have committed a tort in California." *Id.* at 463. The Ninth Circuit applied *Midler* to affirm singer Tom Waits' successful challenge to a Frito-Lay commercial that similarly used a soundalike of his voice singing a song closely associated with him. *See Waits v. Frito-Lay, Inc*., 978 F.2d 1093 (9th Cir. 1992). *Waits* also affirmed Waits' victory on his Lanham Act false endorsement claim, holding that such claims "premised on the unauthorized imitation of an entertainer's distinctive voice, are cognizable under" 15 U.S.C. §1125(a). *Id.* at 1107.

These claims for unauthorized use of an individual's name, likeness, or voice, as well as imitations of those identifying features, are equally applicable to AI-generated content. Thus, whether a person or entity uses a recording of an actual human being or an AI-generated soundalike of a recognized performing artist in a TV commercial or in some other commercial or trade context, that person or entity will likely have violated the performer's right of publicity under state law and may also be liable for false endorsement under the Lanham Act. Similarly, a t-shirt or print advertisement with an image that looks like a celebrity will be equally unlawful regardless of whether the image was created by a human being or an AI system. Nevertheless, as

discussed elsewhere, a federal right of publicity that specifically addresses AI-generated imitations of the images and voices of performing artists will shore up this protection and ensure that it applies uniformly in all U.S. jurisdictions.

31. Should Congress establish a new federal right, similar to state law rights of publicity, that would apply to AI-generated material? If so, should it preempt state laws or set a ceiling or floor for state law protections? What should be the contours of such a right?

Yes, Congress should enact a federal right of publicity statute that would apply to content generated by AI. Rather than preempt state right of publicity laws, a federal statute should establish a floor for the protection of an individual's name, image, likeness, and voice.

When AI has been used to steal or appropriate an artist's persona, we have asked distribution platforms hosting, displaying, and transmitting that content to remove the material on the grounds that it violates state right of publicity laws. To date, the response to such requests has been uneven, and some services have refused to remove the material citing claims of Section 230 immunity.

This is likely the result of the ruling in *Perfect 10 v. CCBill, Inc.*, 488 F.3d 1102, 1107–08 (9th Cir. 2007) (holding that Section 230's intellectual property exemption does not apply to state law claims). Other Circuits have held that state "IP" statutes are exempted from section 230, *e.g.*, *Hepp v. Facebook*, 14 F.4th 204, 209–12 (3d Cir. 2021) (disagreeing with Ninth Circuit and holding that both state and federal IP claims are carved out of Section 230). The *Hepp* case provides a useful overview of the various cases that have addressed this issue as well as a cogent explanation for why the Ninth Circuit is wrong. A federal statute should establish that the right of publicity is an intellectual property right in order to provide clarity, as there is no question that federal IP rights are carved out of Section 230.

In addition, a federal statute should prohibit unauthorized generative AI voice clone "models," algorithms designed with the intent of allowing users to create new recordings in a mimicked voice (or voices) used to train the model without consent. A meaningful statute will also include a First Amendment balancing test to harmonize the potentially conflicting interests of the individual and free speech. We firmly believe in freedom of expression, but broad categorical exemptions in some existing state statutes can enable the exception to swallow the rule. A court is in the best position to weigh the facts and balance the interests; defendants should not be able to trample on an individual's right of publicity by simply labeling their AI creation a form of content included in an all-encompassing list of categorical exemptions.

Beyond those contours, it is important that a federal statute (1) institute fines and enable damage recoveries sufficient to establish a deterrent effect, (2) authorize both federal enforcement and a private right of action, and (3) as with all forms of intellectual property, the right should be eligible for assignment or licensing either in whole or in part, so that enforcement may be delegated.

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?

As a general matter, copyright law does not protect the style of an artist or musician. *See, e.g., Judith Ripka Designs v. Preville*, 935 F. Supp. 237, 248 (S.D.N.Y. 1996) ("The copyright laws do not protect styles, but only particular original designs"); *McMahon v. Prentice-Hall, Inc.*, 486 F. Supp. 1296, 1302 (E.D. Mo. 1980) ("A writer may not claim a monopoly on a particular writing style by virtue of a copyright"). Thus, musicians, composers, and artists

influenced by greats from prior generations may freely draw upon stylistic characteristics of those predecessors.

This lineage of artistic influence and evolution must be sharply distinguished from AI-generated soundalikes, which result from training on copyrighted sound recordings and produce faithful imitations of the identifying features of an individual's voice. In this sense AI-generated soundalikes invade precisely the property rights identified in *Midler* and *Waits*. Whereas style is a diffuse similarity in artistic approach, a soundalike is theft of features of an artist's identity, which must be protected against unauthorized commercial use. *Waits* recognized this distinction. In rejecting defendant's argument that the right of publicity claim was preempted by the Copyright Act, the Ninth Circuit held that the "trial's focus was on the elements of voice misappropriation, as formulated in *Midler*: whether the defendants had deliberately imitated Waits' voice rather than simply his style and whether Waits' voice was sufficiently distinctive and widely known to give him a protectable right in its use." *See Waits* 978 F.2d at 1100. Accordingly, that "style" *per se* is not protectible in no way diminishes the legal protections against copying a performer's voice.

33. With respect to sound recordings, how does section 114(b) of the Copyright Act relate to state law, such as state right of publicity laws? Does this issue require legislative attention in the context of generative AI?

As noted above, UMG proposes that section 114(b) of the Copyright Act may need to be expanded to protect copyrighted sound recordings against AI-generated recordings that either capture the actual sounds of or are substantially similar to those copyrighted sound recordings. As we have discussed above, generative AI's ability to create instant and faithful imitations of sound recordings requires robust protection. There is no longer any good reason to insulate a

technology with this nearly unlimited potential for exact imitation from the basic standards of infringement that apply to all other copyrighted works.

Question 33, however, asks about how 114(b) as currently constituted "relates to" state laws and in particular state rights of publicity. UMG interprets this question to ask whether the current limitations of 114(b) preempt such laws to the extent those laws prohibit soundalikes. In other words, if 114(b) expressly permits "imitations" of sound recordings, can state laws prohibit them? With respect to the matters of great concern in the AI context – the scrupulous imitation of performers' voices – UMG believes that there should be no preemption. In both *Midler* and *Waits*, the Ninth Circuit held that a voice is not copyrightable subject matter, so efforts to protect it are not preempted by Section 301 of the Copyright Act. *See Waits*, 978 F.2d at 1100 ("a state cause of action escapes Copyright Act preemption if its subject matter 'does not come within the subject matter of copyright . . . including works or authorship not fixed in any tangible medium of expression.' 17 U.S.C. § 301(b)(1) … Waits' claim, like Bette Midler's, is for infringement of voice, not for infringement of a copyrightable subject such as sound recording or musical composition"); *Midler*, 849 F.2d at 462 ("A voice is not copyrightable. The sounds are not 'fixed.' What is put forward as protectible here is more personal than any work of authorship"). Moreover, not only is a voice not within the subject matter of copyright, but the right to protect one's identity from commercial exploitation is not "equivalent" to rights reserved to copyright owners in their copyrighted works.

UMG acknowledges that there is some dispute on the subject of preemption and rights of publicity. For example, Nimmer is critical of Ninth Circuit jurisprudence on this issue, although he ultimately agrees that the holdings are correct insofar as the soundalikes were used for "purposes of trade" and thus should not be preempted. *See Nimmer* §1:17 [B][2] and [3][A].

Thus, while UMG maintains that rights of publicity as applied to AI-generated soundalikes are not preempted, the potential for disagreement further counsels in favor of a federal right of publicity that will eliminate debate on this issue.

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

The discussions above are focused on AI systems that use machine learning. However, those are not the only kind of generative AIs. AIs may generate content, especially music, using a rules engine. Many existing and past music generation systems use this approach. These systems typically have separate collections of rules (templates) hand-coded to generate different genres or styles of music. The rules allow for variation in key, timing, length, and other attributes. They typically generate compositions based on those rules and then apply samples, loops, and effects in order produce audio. To generate music, users typically select from a list of genres or musical styles, then optionally specify variables such as key, length, and beats-per-minute rather than typing in a prompt.

Because these systems do not typically train on audio recordings, they are unlikely to raise all of the concerns discussed above. However, they can still infringe copyrighted content if they use unlicensed audio from copyrighted sources as samples or loops, and they could infringe musical works if the developers used copyrighted compositions when building the rules. Rules-based AI developers should seek licenses for any copyrighted content they use.

* * *

On a final note, we anticipate future court decisions, changes in the law here and abroad, advances in technology and a rapidly evolving marketplace will impact our views of and

approach to these issues. We recognize this is an ever-changing landscape, but one principle will always remain the same: Our copyright system is essential, and creators and copyright owners must be protected in the era of AI. We look forward to working with the Office and other policymakers in the months ahead as circumstances evolve.