BEFORE THE

U.S. COPYRIGHT OFFICE

Artificial Intelligence and Copyright	**Docket No. 2023–6**

COMMENTS OF THE COPYRIGHT ALLIANCE

The Copyright Alliance appreciates the opportunity to submit the following comments in response to the notice of inquiry and request for comments ("NOI") published by the U.S. Copyright Office in the Federal Register on August 30, 2023 (and supplemented by the extension of the comment period on September 21, 2023), regarding the Office's study of the copyright law and policy issues raised by artificial intelligence ("AI") systems.

The Copyright Alliance is a non-profit, non-partisan public interest and educational organization that is dedicated to advocating policies that promote and preserve the value of copyright, and to protecting the rights of creators and innovators. We represent the copyright interests of over 2 million individual creators, including established authors and artists, performers and photographers, and software coders and songwriters, as well as a new generation of creators. Some of these creators are career professionals, while others are hobbyists. Some have years of experience, while others are just embarking on their burgeoning careers. Some are critically acclaimed, while others toil in relative obscurity or have limited audiences. Perhaps most

importantly for the purposes of this study, some of these creators are long-time users of AI, while others are just beginning to use these tools.

We also represent the copyright interests of over 15,000 organizations in the United States, across the spectrum of copyright disciplines. These include motion picture and television studios, record labels, music publishers, book and journal publishers, newspaper and magazine publishers, video game companies, software and technology companies, visual media companies, sports leagues, radio and television broadcasters, database companies, standard development organizations and many more. Importantly, these also include companies that have developed their own AI tools,[1] companies that have been using AI in some form for many years, and companies that have just begun exploring how to use generative AI. Each of these organizations comes to the Copyright Alliance with somewhat different experiences, views, and interests. Regardless of how their approaches to AI may differ, they all fall under the Copyright Alliance umbrella for a reason—their strong support for the value and importance of copyright and protecting the rights of human creators and copyright owners.

All Copyright Alliance members—whether they are an individual creator or an organization, whether they are big or small, or whether they are more traditional creators/copyright owners or a new generation of creators/copyright owners—share two things in common: (1) they rely on copyright law to protect their creativity, efforts, and investments in the creation, reproduction, distribution and adaptation of copyrighted works for the public to enjoy, and (2) they are interested in and concerned about copyright-related issues raised by generative AI. During its 16-year history, other than online piracy, no copyright issue has drawn more interest from the Copyright Alliance membership than generative AI.

As discussed more in the answers to the questions below, Copyright Alliance members' interests and concerns may vary depending on factors such as the size of the creator/copyright owner, the type of copyrighted work, business and licensing experience and models, different AI models that affect their industry, provisions in the copyright law that may be applicable, their existing

[1] For example, some of our members (or members of members) who are both creators/copyright owners and also developers of generative AI foundational models, include Adobe, Oracle, and Getty.

and/or intended future use of generative AI, and many other factors. Despite these variables, as discussed further below, there are core generative AI principles that every Copyright Alliance member supports. These core principles can be summarized in one sentence: *We support responsible, respectful, and ethical development and use of AI.*

Introductory Comments

Before responding to the questions in the NOI, a few prefatory remarks are necessary. First, unless stated otherwise, our responses to the questions focus exclusively on generative AI models, which are a distinct form of machine learning algorithm that is programmed for a specific purpose—to manufacture output (when prompted) based on preexisting works that are ingested by the system.[2] Generative AI is different than traditional (sometimes called analytical) AI in that it ingests creative works in order to manufacture new material, whereas traditional AI is used to do more rote and mechanical tasks and calculations like detect patterns, hone analytics, or classify data.[3] The output of generative models could be text, images, audio, or audiovisual material, which the AI model typically manufactures only after ingesting similar copyright works. We focus our comments on generative AI systems because they are of most concern and pose the greatest risks to the copyright community.

We also want to make clear that when we use the terms "ingestion" or "ingestion stage" throughout our responses, we are using them to describe any and all activities that occur during the development of datasets for training generative AI models *and* that surround the feeding of the dataset into the model. These "ingestion-stage" activities include the collection (which includes scraping) and curation of copyrighted works for training purposes, regardless of whether the entity engaged in such acts is the same as the entity that owns or operates the generative AI system that ingests the works. Additionally, when we use the term "industry" in our responses, we are using it as shorthand to mean individuals, businesses, and others that own

[2] Kim Martineau, *What Is Generative AI?*, IBM (Apr. 20, 2023), https://research.ibm.com/blog/what-is-generative-AI.

[3] Elysse Bell, *Generative AI: How It Works, History, and Pros and Cons,* INVESTOPEDIA (May 26, 2023), https://www.investopedia.com/generative-ai-7497939#.

copyrighted works, as well as creators, rights administrators, and the associations that represent these groups.

Second, it bears reminding that while, in the future, AI *may* be a significant contributor to the economy and perhaps jobs, the contributions of U.S. creative industries—made possible through copyright law—have been one of the most significant contributors to the U.S. economy and to job creation for decades. A report on the economic impact of copyright by the International Intellectual Property Alliance (IIPA) notes that, in 2021, the total copyright industries contributed more than $2.9 trillion to the U.S. gross domestic product (GDP) (accounting for 12.52% of the U.S. economy) and employed nearly 16.1 million workers (or 8.14% of the workforce).[4]

In addition to growing at a rate more than three times that of the rest of the economy, the report notes that the core copyright industries:

- make up an increasingly large percentage of value added to GDP;
- create more and better paying jobs than other sectors of the U.S. economy;
- grow faster than the rest of the U.S. economy;
- contribute substantially to U.S. foreign sales and exports, outpacing many industry sectors; and
- make significantly large contributions to what the U.S. Bureau of Economic Analysis defines as the digital economy, which does not even encompass the full scope of the copyright industries' digital activities.[5]

The U.S. continues to be the world leader in intellectual property[6]—an attribute that contributes significantly to this country's vast cultural influence and its standing as the world's leading

[4] *See* Robert Stoner & Jéssica Dutra, *Copyright Industries in the U.S. Economy: The 2022 Report*, INT'L INTELL. PROP. ALL. 8 (Dec. 2022), https://www.iipa.org/files/uploads/2022/12/IIPA-Report-2022_Interactive_12-12-2022-1.pdf.

[5] *Id*. at 7.

[6] U.S. CHAMBER OF COM. GLOB. INNOVATION POL'Y CTR., INTERNATIONAL IP INDEX 6–7 (11th ed. 2023), https://www.uschamber.com/assets/documents/GIPC_IPIndex2023_FullReport_final.pdf.

economy. Any AI-related copyright policies that are considered must take into account the effect such policies may have on copyright's importance to the economy and job creation.

Third, it should be recognized how important copyright is to empowering marginalized and underrepresented communities. By incentivizing and rewarding the creation and dissemination of copyrighted works, copyright encourages participation in the creative industries by a diverse range of creators and copyright owners. For example, by enabling creators to earn a living from the works they create, copyright law helps to ensure that meaningful contribution to the arts and entertainment is not a privilege reserved for those with financial means. Likewise, the exclusive rights afforded by copyright provide creators the autonomy to create works that are reflective of their experiences, viewpoints, and communities, which in turn aids in increased (and more authentic) representation of marginalized and underrepresented groups in the media and entertainment. As evidenced by the data on the contributions of the core copyright industries to job creation, copyright also creates employment opportunities for creators and creative professionals from underserved and marginalized backgrounds. Any AI-related copyright policies that are considered must recognize the importance of copyright to empowering marginalized and underrepresented communities.

Fourth, as the Office has already recognized in the NOI, the terms used when discussing these copyright-AI issues are very important. In particular, we want to highlight an important distinction between "data" and copyrighted works. In discussing the massive amount and array of material ingested by generative AI systems, some people have begun to incorrectly lump copyrighted works under the umbrella term "data." We want to make it clear that copyrighted works are not data. This is not just a matter of semantics. Instead, it is an issue that gets to the core of our concerns relating to generative AI and copyright. The term data refers to "factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation."[7] By contrast, copyrighted works—books, music, movies, photographs, paintings, sculptures, video games, etc.—are works of creative expression. In fact, to be protected under copyright law, a work cannot be mere data—i.e., facts or information. To garner copyright

[7] *Data*, MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/data (last visited Oct. 30, 2023).

protection, a work must be an authored work of expression; as the Supreme Court has made clear, "facts are not copyrightable." *Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340, 344 (1991); see also *Harper & Row, Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539, 547 (1985) ("[N]o author may copyright facts or ideas. [17 U.S.C.] §102."). To mislabel a copyrighted work as mere "data" is to strip it of the critical essence by which it avails itself of copyright protection: its expressive value and human creativity. It should also be understood that while *Feist* and *Harper* held that facts or ideas are not copyrightable, both decisions recognized that expressions of information or compilations of facts can be copyrightable because they reflect sufficient originality.[8]

While there are important discussions to be had about the ingestion of unprotectable data by generative AI systems, those discussions differ in substance, and must remain separate, from discussions about the ingestion of copyrighted works. It is therefore vital that the term "data" be reserved for clearly unprotectable facts and information and not be used to refer to copyrighted works.

Relatedly, there is a troubling trend surrounding the terminology used when discussing issues raised by generative AI and copyright. That trend involves the aforementioned de-humanization of human creativity (by using words like "data" to refer to copyrighted works) and the humanization of generative AI operations through the use of terms like "hallucinate," "learn," "unlearn," and "create" when referring to AI-generated output. We caution the Copyright Office to not fall into the trap of anthropomorphizing the actions of generative AI systems. Doing so devalues human creativity and wrongfully prioritizes the interests of AI companies over the rights and interests of creators and copyright owners.

Lastly, we commend the Copyright Office for asking such thoughtful and comprehensive questions. We understand and appreciate how challenging this study is and that formulating the number and type of questions to ask, as well as how to organize and phrase the questions, is no

[8] See *Feist Publications, Inc.,* 499 U.S. at 348 ("Thus, even a directory that contains absolutely no protectible written expression, only facts, meets the constitutional minimum for copyright protection if it features an original selection or arrangement."); *see Harper Row*, 471 U.S. at 547 (1985) ("Creation of a nonfiction work, even a compilation of pure fact, entails originality.").

small task. We also want to thank the Office for granting an extension to respond to the NOI. Our goal is to provide the Office with the most comprehensive and helpful responses possible. The extra two weeks has helped us immensely in achieving this goal, while at the same time, ensuring that the results of this study can be published quickly—hopefully early next year.

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?

The continuing development of a broad range of AI systems represents a great achievement of the digital age that brings with it tremendous opportunities. In fact, many in the creative industries are already using or plan to use AI-based technologies to assist in the creation of a wide range of works that will benefit society. Some—like the motion picture, video game, and music industries—have been using AI-based assistive tools for many years. Others—like many independent illustrators and authors—have just begun exploring how to incorporate AI tools into their work process.

There is little doubt that AI tools give rise to opportunities for new forms of creativity and expand existing forms of creativity for all types of creators and all types of creative works. For example:

- Artists who have physical, mental, or other challenges may have struggled with or been unable to perform certain aspects of the creative process, but through the assistance of AI they may now be able to do what previously they could not. Relatedly, marginalized groups whose creative efforts may have been limited in the past due to a lack of resources or access to creative tools may find that the democratization of generative AI "levels the playing field" and results in new creative endeavors.

- Fiction authors who struggle with "writers' block" use generative AI to assist with their ideation, as well as to do things like develop characters, location names, plot lines etc.

- AI permits artists to take on more ambitious projects which formerly required extensive labor and outsourcing. For example, artists have reported that they use AI image generators to create large background elements that they then combine with original artwork they create off-platform.

- Artists who do not use generative AI tools for creating client work because of concern and confusion about the copyrightability of such work, may still use generative AI to assist with their ideation. This may be done at the outset of a client project, or for personal growth, such as training models on their own works to explore new directions they pursue off-platform.

- Use of responsibly trained AI tools can create new artistic opportunities and markets for creators to commercialize their work. For example, in partnership with Nvidia, Getty Images recently released its GenAI tool which allows users to generate high-quality AI-generated visuals from a text prompt. Adobe Firefly is another good example of how AI technology can successfully augment human artistic expression when trained on proprietary or licensed copyrighted works.

- The voluntary licensing of creative works to AI companies for training has created new revenue streams for some copyright owners. As discussed in more detail in our answer to question six, there are numerous examples of AI companies that are directly entering into licensing agreements with (and thereby compensating) copyright owners for the ingestion of their works. With every day that passes, more new license agreements are being reached. That is welcome progress.

As with most advances in technology, new opportunities are often accompanied by new challenges. AI is no different. The breakneck speed with which generative AI technology is being

developed, in conjunction with the decision of AI developers to develop their systems by engaging in large-scale copying of copyright-protected works without the consent or input of the works' owners and creators presents legitimate causes for concern for not just the creative community, but the whole of society.

As generative AI technology continues to evolve and questions arise about how copyright laws apply to the ingestion of copyrighted works, development and dissemination of generative AI systems, and the output of generative AI systems, *it is critical that the underlying goals and purposes of our copyright system are upheld and that the rights of creators and copyright owners are respected*.

Listed below are some of the known risks that are directly or indirectly related to copyright that concern the Copyright Alliance and our members. Obviously, there are numerous risks that fall outside the scope of copyright and therefore we will either not address or merely mention in passing in the process of answering this question. The risks our members would like to highlight to the Copyright Office for the purposes of this study include:

- Generative AI models are often developed based on copyrighted works without any advance permission from or remuneration for the original creators. In addition, the outputs from these AI systems may frequently act as a substitute or supplant the market for the copyright-protected works they are trained on, threatening the very livelihoods and careers of the human creators whose works enable the generative AI systems to work. Copyright owners and creators invest large amounts of time, money, ingenuity, creativity, and resources to create these works. If they are not appropriately compensated when their works are ingested, they will not be able to continue creating them.

- With AI art generators, the base models for the AI platforms (such as Stable Diffusion and Midjourney) come pre-trained on some artists' works and styles, typically famous artists. This becomes a copyright issue when users then fine-tune the dataset using as few as 20-30 images, permitting artists' works to potentially appear in whole or in

part in AI-generated images. We understand that some artists have reported that AI image generators may have been weaponized by using their artwork for fine-tuning models in retaliation for them speaking out against generative AI.[9]

- Vocal models are also being fine tuned with copyrighted works to reconstruct the voice of famous recording artists without authorization from the copyright holders nor the artists, and then those models are used to modify the sound of the voice on copyrighted sound recordings to sound like the artist for which the vocal model was developed. This harms the copyright holders of the works being ingested and the works being modified, as well as harming the artists whose voice is exploited or modified without authorization.

- As noted above in our discussion of the benefits of voluntary licensing, while there are examples of AI companies that are directly entering into licensing agreements with (and thereby compensating) copyright owners for the ingestion of their works, unfortunately, so far, many AI companies have been slow or unwilling to license works from creators and the vast majority of works ingested from copyright owners of all types have not yet been licensed. The voluntary licensing of works for AI use can bring creators much needed additional compensation, and it can insulate AI companies from the risk of expensive and time-consuming litigation.

- For certain industries, like news media, there are societal risks posed by impacting existing business models. High-quality publisher content supports a healthy democracy and vibrant communities with publishers investing considerable time and resources to produce journalism and creative content that combats misinformation,

[9] *Webinar: Protect Your Artistic Style from AI Mimicry*, GRAPHIC ARTISTS GUILD (Sept. 20, 2023), https://graphicartistsguild.org/product/protect-your-style-from-mimicry/ ("People then use models to fine-tune on individual artists, and that is where most of the damage is done. Basically, AI models can now be weaponized to target people. It's really disturbing . . . I hear stories from Japan where artists who speak up against AI are basically targeted with AI models as a weapon. So, if you speak up against AI, some AI bro will literally go find all your art and intentionally train a model on you and then use that model, use your style to draw images that are disturbing or offensive or against everything you stand for. And that is their way of getting back and retaliating against some of these artists who are not in favor of AI."); @JonLamArt, X (*formerly Twitter*) (Jan. 9, 2023, 9:01 AM), https://twitter.com/JonLamArt/status/1612494765203009536?s=20.

encourages democratic engagement, strengthens community ties, safeguards consumers, keeps decision makers accountable, and supports the free flow of ideas and information. By undermining publishers' ability to benefit from their investments in high-quality content, AI systems using unlicensed content risk the very foundation of our society, including through the closure of local newspapers, magazines, and online-only outlets, the spread of mis- and disinformation, and reduced access to information that can fundamentally only be created by humans.

- AI output is no replacement for human creation. For example, research has shown that AI platforms that are continuously trained on AI output will generate very low-quality, incorrect, or biased outputs—highlighting the value and continual need of preserving and protecting human created expressions.[10] There may be other unintended effects of prioritizing AI technologies at the expense of its backbone—human creativity—that could result in a regression in the progress of the culture and arts of our country.

- If developers of AI models are not accountable, there might be no incentive for them to act responsibly, ethically, and respectfully with regard to copyright, which in turn could stifle innovation and creativity as well as harm people's trust in technology.

- Without appropriate and effective transparency rules, AI developers can exploit copyright owners and creators, especially independent creators, without their knowledge.

[10] This may reduce the diversity of novel content which may ultimately harm the public. *See, e.g.*, Anil R. Doshi & Oliver Hauser, *Generative Artificial Intelligence Enhances Creativity but Reduces the Diversity of Novel Content* 7 (Aug. 8, 2023) (unpublished paper), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4535536 ("[O]ur findings suggest that the produced stories would become less unique in aggregate and more similar to each other . . . Initial evidence suggests that GenAI models trained with GenAI content become unstable."); Vishakh Padmakumar, *Does Writing with Language Models Reduce Content Diversity?*. CORNELL UNIV. arXiv (Sept. 11, 2023), https://arxiv.org/pdf/2309.05196.pdf ("We find that the set of essays written with InstructGPT does not only have lower lexical diversity, but also exhibits lower diversity in terms of the key points they present . . . Reduced content diversity is not only detrimental to personal expression and creativity . . .").

- Flooding the market with AI-generated works creates competition against human-created works and makes it difficult for consumers to find the higher-quality, human-created works they prefer in the sea of low-quality outputs.

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

As noted in our introductory comments, the Copyright Alliance does not represent a particular sector or industry—we represent *all* sectors and industries that rely on copyright for their livelihoods, careers, and businesses, as well as some that are also generative AI foundational model developers. Therefore, we are not in a position to represent the views of one particular industry and compare and contrast it to another—we will leave that to our members. What we are in a position to do is present the amalgamated, overall view of the copyright community—one that balances and takes into account the views of all the creative sectors, as well as the interests of AI developers.

As the only organization that represents the interests of the entire copyright community, we bring a unique perspective to the AI-copyright discussion. We understand and appreciate the interests and concerns of each particular copyright industry and group of creators and how those interests and concerns compare to the interests and concerns of others within and outside the copyright community. In our view, that puts the Copyright Alliance in exclusive company.

With that important background in mind, we will now respond to what we believe is the most important question of the NOI because our answer to this question informs and contextualizes our answers to all the questions that follow.

When thinking about the issues raised by generative AI and copyright, it is essential to understand that AI impacts each of the different copyright industries differently. This is because each industry has very different business models. There are different ways that they create their works, different ways they make their works available to the public, different industry policies and standards that they adhere to, different approaches to licensing the works of and to third

parties, different ways they are remunerated for their works, different ways they remunerate others for works they use, and different internal business structures (e.g., some have collective bargaining agreements).

It is also because certain provisions in copyright law or other laws apply to each of them differently. For example, there are licensing provisions in the Copyright Act that apply to musical works and sound recordings that do not apply to other types of copyrighted works.[11]

Characteristics of a creator or copyright owner—such as their size, their name recognition, the number of works they have created—will also influence their views of AI. We are already seeing evidence of that difference in the AI market right now. As license agreements between the creative industries and AI companies grow in number, that growth so far seems to be limited to large companies with large corpuses of copyrighted works that can negotiate more easily with large AI companies. We have yet to see much in the way of AI companies licensing the works of independent creators whether directly or through existing copyright management companies.

All these factors play a role in determining the views that the copyright community has about AI and inform why those views might differ from industry to industry and from creator to creator. But perhaps most importantly, their views might differ because AI impacts each copyright industry differently due to the fact that the AI models that impact them are different. For example, an LLM is different from an audio model—most LLM models appear to have been developed on vast quantities of online text, while music models appear to have been developed on orders of magnitudes of less content.[12]

---

[11] This should not be construed as an endorsement of those music licensing provisions. The Copyright Office is aware of the music communities' concerns with these licensing provisions, but those provisions are not within the scope of this NOI and therefore we will not discuss them here.

[12] One example of a music AI that does not ingest copyrighted works is Boomy. *See* Listening Session on Music and Sound Recordings, held by U.S. Copyright Off. (May 31, 2023), https://www.copyright.gov/ai/listening-sessions.html#sound-recordings. There are reports that suggest a model trained with higher quality music, but lower amounts of licensed music (e.g., MusicGen) works better than music AI models trained with more, but likely lower quality music. As discussed in our answer to question 9.3, here is no reason to think that it cannot be extrapolated to apply to others AI models, like large visual models (LVM) and LLMs. *See* Matt Mullen, *AI Music Wars: Meta Takes on Google and Releases Its Own AI Music Generator – But Whose Is Better?*, MUSICRADAR (June 16, 2023), https://www.musicradar.com/news/meta-google-ai-music-wars-musicgen (concluding that Meta's product, which is

As a result of these considerations, during the course of this study the Office has heard and will continue to hear different responses from different types of copyright owners and creators. That does not mean they disagree with one another—it just means that the factors informing their responses differ and that they are responding for their specific industry. The fact that industries differ in their approaches, does not mean the Office should conclude that there is no consensus. We urge the Office to drill deeper and to examine the many areas where there is intra-industry consensus.

What this means is that—as the Copyright Office, Congress, the Administration or any other policymakers consider copyright-related AI issues—it's important that the question being asked by policymakers is *not* how AI impacts the copyright community as a whole, but rather how AI impacts copyright owners and creators in the book publishing industry, how it impacts copyright owners and creators in the music industry, how it impacts copyright owners and creators in the motion picture and television production industries, how it impacts copyright owners and creators in the visual arts community and so on—*because the impact will be different and therefore the responses and solutions may need to be different*.

Similarly, we urge the Office to examine the areas where there is multi-copyright industry consensus because, despite their different approaches to certain AI-related copyright topics, there are basic tenets/principles that all our members can agree on. Much of this will be apparent as you review our answer to the NOI questions, but for simplicity we list several of these principles below.

---

trained on significantly less music than Google's product, creates better music); *see also* Ali Shutler, *New Meta AI Music Tool Is Trained on 10,000 Hours of 'Licensed Music'*, EMERGE (June 13, 2023), https://decrypt.co/144425/new-meta-ai-music-tool-musicgen-trained-hours-licensed-music; *see also* Jade Copet, et al., *Simple and Controllable Music Generatio*n, CORNELL UNIV. ARXIV 2, 7 (June 8, 2023), https://arxiv.org/pdf/2306.05284.pdf ("[H]uman evaluation suggests that MusicGen yields high quality samples which are better melodically aligned with a given harmonic structure, while adhering to a textual description . . . Results suggest that MusicGen performs better than the evaluated baselines as evaluated by human listeners, both in terms of audio quality and adherence to the provided text description.").

There are seven fundamental principles that must form the basis of a common understanding amongst stakeholders, courts, policymakers, and the public when it comes to the relationship between copyright and generative AI.

1. *When formulating new AI laws and policies, it is essential that the rights of creators and copyright owners be respected*. When making determinations about AI policies, it is vital for policymakers and stakeholders to understand that any new laws and policies relating to AI must be based on a foundation that preserves the integrity of the rights of copyright owners and their licensing markets. The interests of developers who use copyrighted materials for ingestion by AI systems must not be prioritized over the rights and interests of creators and copyright owners.

2. *Longstanding copyright laws and policies must not be cast aside in favor of new laws or policies obligating creators to essentially subsidize the development of AI technologies*. Established copyright laws must not be weakened based on a mistaken belief that doing so is necessary to incentivize the development of AI technologies. This is especially true when there is no evidence of market failure or problems warranting changes to the law. AI-specific statutory exceptions to copyright law that would effectively strip rightsholders of their ability to control and be compensated for the use of their copyrighted works for ingestion purposes are unnecessary and harmful and should be rejected.

3. *The ingestion of copyrighted material by AI systems implicates the right to reproduce copyrighted works*. Section 106(1) of the Copyright Act vests copyright owners with the right to prevent the reproduction of their copyrighted works. When an unauthorized copy is made of a work protected by copyright, there is a violation of the copyright owner's right to reproduce the work, absent a valid defense. It is important to understand that copyright infringement at the input stage is distinguishable from infringement at the output stage because the reproduction right is a "stand-alone" right—it is violated by copying a work (without authority or defense) regardless of whether a specific output of an AI system is infringing.

4. *The ingestion of copyrighted material by AI systems is not categorically fair use*. Determining whether a particular use qualifies for the fair use defense to infringement requires a fact-specific inquiry that is considered on a case-by-case basis. There are no uses that always, categorically qualify as fair use. That is no less true when copyrighted work are used for AI ingestion. In fact, the typical commercial system's ingestion of copyrighted works is particularly unlikely to qualify as fair use when the AI system generates competing works. Courts will need to evaluate fair use defenses involving AI systems the same way they evaluate fair use in all contexts: by applying the four factors set forth in section 107 of the Copyright Act to the specific uses at issue. Under the first factor, ingestion is unlikely to be a transformative use since the output generated by these AI systems will often serve the same exact purpose as the works ingested, especially in the case of music and art. However, even if the use is held to be transformative, as the Supreme Court recently made clear in *Andy Warhol Foundation v. Goldsmith*, whether a use is transformative is not dispositive of the question of fair use and is merely one of the considerations under the first fair use factor. In addition, under the fourth factor, when courts consider the extent of the "effect of the use upon the potential market for or value of" the works ingested by that system, they may conclude that such ingestion will have a significant adverse impact on the value and market for the copyrighted work. This is especially true when copyright owners have made licenses available in the market for AI training. Finally, as we discuss more in response to question eight, the second factor may often weigh against a finding of fair use, and the third factor will either weigh against fair use or be neutral.

5. *AI companies should license works they ingest*. No AI-copyright policy should be adopted in response to generative AI that interferes with the free market or the freedom to license. It is essential that the licenses be respected by any copyright or AI legal regime. Obtaining a license to use copyrighted works is the best way for developers to ensure they avoid infringement liability. Further, if licensing markets exists or are being developed, it can weigh against a finding that copying without the permission of the copyright owner is excused by the fair use defense. The marketplace should continue to

properly value and incentivize creativity, and AI policy should not interfere with the right of copyright owners to choose whether to license, or not to license, their works for AI purposes. Copyrighted works provide immense value to AI developers, and they can and should pay for that value—as many today are already doing. In other words, when properly applied, copyright law sets the conditions for the market to prevail.

6. *AI systems must implement safeguards to prevent infringing AI-generated outputs*. Overfitting and allowing prompts that call for copyright protected-material and "in the style of" are more likely to result in AI-generated outputs that infringe one or more copyrighted works. While merely imitating the style of an existing artist does not constitute infringement, it is essential that AI companies implement effective safeguards to prevent the likelihood of output-related infringements. This is yet another reason why AI companies should voluntarily license ingested works because when they do so, the parties can negotiate these safeguards.

7. *Transparency regarding ingestion of copyrighted works by businesses that offer generative AI systems to the public will help ensure that the rights of copyright owners are respected, and that AI development is being implemented in a way that is responsible and ethical*. Adequate and appropriate transparency and record-keeping benefit both copyright owners and AI developers in resolving questions regarding infringement, fair use, and compliance with licensing terms. Transparency has many other benefits unrelated to copyright such as promoting safe, ethical, and unbiased AI systems. Consequently, transparency by businesses that offer generative AI systems to the public is a crucial component of any AI policy. Best practices should include maintaining records of what copyrighted works are being ingested and how those works are being used, except where the AI developer is also the copyright owner of the works being ingested by the AI system. Those records should be publicly accessible and searchable as appropriate and subject to reasonable confidentiality provisions the parties to a license might negotiate as well as the aforementioned exception.

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.

In addition to the several papers and studies we reference throughout the answers to these questions, there are other papers and studies that are relevant to this NOI that might be of interest to the Copyright Office as it conducts its study; some of these include[13]:

- AI Art and its Impact on Artists: This article explores the general legal, economic, and cultural effects of image-generative AI technologies on visual artists and creators.[14]

- Investigating Data Replication in Diffusion Models: This article explores the extent to which various AI models generate outputs that reproduce an ingested work used to train the model. The authors conclude that many AI models are capable of generating output that is a reproduction of an underlying work and that in actuality the likelihood of replicated material in AI output is much higher.[15]

- Machine Unlearning: Solutions and Challenges: This article provides an overview of the various "unlearning" methods that AI developers employ to selectively remove various

[13] We note these for the interest of the Copyright Office and not to imply endorsement of the views therein.

[14] *See* Harry Jiang et al., *AI Art and Its Impact on Artists*, *in* PROC. OF THE 2023 AAAI/ACM CONF. ON AI, ETHICS, & SOC'Y (AIES '23) (Aug. 2023), https://dl.acm.org/doi/pdf/10.1145/3600211.3604681 ("The proliferation of image generators poses a number of harms to artists, chief among them being economic loss due to corporations aiming to automate them away.").

[15] *See* Gowthami Somepalli et al., *Diffusion Art of Digital Forgery? Investigating Data Replication in Diffusion Models*, CORNELL UNIV ARXIV (Dec. 12, 2022), https://arxiv.org/pdf/2212.03860.pdf ("[I]f we look only at very close matches . . . these match images are replicated on average 34.1 times – far more often than a typical image. It seems that replicated content tends to be drawn from training images that are duplicated more than a typical image.").

training data points of certain ingested works from the AI algorithm. Importantly, it shows that complete "unlearning" is challenging.[16]

- The Curse of Recursion: Training On Generated Data Makes Models Forget: This article explores the implications of when generative AI models are trained on its outputs. These findings show that in order to continue generating quality outputs, AI machines require human-created expressive works as inputs and highlights the need to support and protect the creators behind such works.[17]

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?

U.S. rightsholders are not isolated or unaffected by international developments, and so it is vital that international approaches to AI and copyright are harmonized in that they respect and uphold the copyrights of human creators and copyright owners. The United States should continue collaborating with their international counterparts on actively developing and supporting standards, guidelines, and policies which promote copyright and protect against attempts to undermine these rights, especially when it comes to creating novel copyright exceptions. The Hiroshima AI Process international commitment is an example of an opportunity for the U.S. to prioritize and champion copyright, as the G7 countries are discussing the need to address generative AI topics including the "safeguard[ing] of intellectual property rights including copyright . . ." in developing the G7's AI guiding principles.

[16] *See* Jie Xu et al., *Machine Unlearning: Solutions and Challenges*, CORNELL UNIV. ARXIV (Aug. 14, 2023), https://arxiv.org/pdf/2308.07061.pdf ("Machine unlearning faces challenges from inherent properties of ML models as well as practical implementation issues.").

[17] *See* Ilia Shumailov et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget*, ARXIV.ORG (May 31, 2023), https://arxiv.org/pdf/2305.17493v2.pdf ("[T]o avoid model collapse, access to genuine human-generated content is essential.").

In terms of current international regulatory or statutory approaches, other than the EU countries, there are only a handful of countries considering AI regulations with respect to exceptions in copyright laws and even fewer that have enacted such laws. Among those countries that have considered or are considering the adoption of copyright exceptions for text-and-data mining ("TDM"), Brazil,[18] Hong Kong,[19] South Korea,[20] Australia,[21] and Canada[22] significantly, have so far declined to do so.

In varying degrees, only the European Union, Japan, Singapore, and the United Kingdom have AI exceptions within their copyright laws. But none of these approaches should be considered in the United States, as this would not only require a change to the Copyright Act but could also potentially result in U.S. noncompliance under the Berne Convention for the Protection of Literary and Artistic Works. That is because such exceptions may not be compliant under the Berne three-step test, especially as copyright owners' AI licensing markets have been developing.[23] A brief summary of these four problematic approaches follows.

---

[18] Projeto de Lei nº 21/2020, de 2 de Fevereiro 2022, Diário Oficial da União [D.O.U.] de 2.4.2022 (Braz.).

[19] *See* COMM. & ECON. DEV. BUREAU, UPDATING HONG KONG'S COPYRIGHT LAWS: PUBLIC CONSULTATION PAPER, 31–32, (Nov. 24, 2022), https://www.cedb.gov.hk/archive/assets/resources/citb/consultations-and-punblications/(Eng)%20Consultation%20Paper%20on%20Copyright.pdf.

[20] *See* Jeojaggwonbeob Jeonbugaejeongbeoblyul-an [Total Amendment to the Copyright Act], Bill No. 2107440, Jan. 15., 2021 (S. Kor.), https://likms.assembly.go.kr/bill/billDetail.do?billId=PRC_Q2T1M0X1D0M4W1T4M3O0R3Y4C7O3D2. South Korea announced that the government will be releasing AI guidelines related to copyright and AI. *South Korea to Set New Standards for Copyrights of AI-Generated Content*, DIGIT. WATCH (May 3, 2023), https://dig.watch/updates/south-korea-to-set-new-standards-for-copyrights-of-ai-generated-content#:~:text=South%20Korea%20to%20set%20new%20standards%20and%20guidelines%20for%20copyrights,advancements%20and%20encourage%20citizens'%20participation.

[21] *See* AUSTRALIAN L. REFORM COMM'N (ALRC), COPYRIGHT AND THE DIGITAL ECONOMY (DP 79) , 8.41–.63, (May 2013), https://www.alrc.gov.au/wp-content/uploads/2019/08/dp79_whole_pdf_.pdf.

[22] *See* INNOVATION, SCI, & ECON. DEV. CANADA (ISED), A CONSULTATION ON A MODERN COPYRIGHT FRAMEWORK FOR ARTIFICIAL INTELLIGENCE AND THE INTERNET OF THINGS 7–10, (2021), https://ised-isde.canada.ca/site/strategic-policy-sector/sites/default/files/attachments/2022/ConsultationPaperAIEN.pdf. The Canadian government recently launched consultations on AI and copyright issues. *Government of Canada Launches Consultation on the Implications of Generative Artificial Intelligence for Copyright*, GOV'T OF CAN. (Oct. 12, 2023), https://www.canada.ca/en/innovation-science-economic-development/news/2023/10/government-of-canada-launches-consultation-on-the-implications-of-generative-artificial-intelligence-for-copyright.html.

[23] *See* Agreement on Trade-Related Aspects of Intellectual Property Rights, art. 9, Apr. 15, 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299, 33 I.L.M. 1197 (1994) [hereinafter TRIPS Agreement]. Specifically, the TRIPS Agreement states that: "Members shall confine limitations

- *European Union*: The EU has a limited exception that excuses TDM of copyrighted works for the purposes of scientific research, with no ability for rightsholders to opt out. TDM of copyrighted works for commercial purposes is also allowed, but subject to a machine-readable opt-out request from the rightsholder. Lawful access to the copyrighted work is required.

- *Japan*: Japan has an ambiguous and overbroad exception that excuses TDM of copyrighted works which (1) is "not a person's purpose to personally enjoy or cause another person to enjoy the thoughts or sentiments expressed in that work" and (2) "does not unreasonably prejudice the interests of the copyright owner in light of the nature or purpose of the work or the circumstances of its exploitation."[24] There are no prohibitions as to commercial use and no requirements that the work be lawfully accessed or published.

- *Singapore*: Singapore has an overbroad exception that excuses TDM of copyrighted works with no ability for rightsholders to opt out.[25] There are no prohibitions as to commercial use and there are no requirements that the work be lawfully published—permitting TDM of even pirated and illegal works.

- *United Kingdom*: The UK has a limited exception that excuses TDM of copyrighted works for the sole purpose of research for a non-commercial purpose. Lawful access to the copyrighted work is required.

---

or exceptions to exclusive rights to certain special cases which do not conflict with a normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the right holder." *Id.* art. 13.

[24] Chosakukenhō [Copyright Act], Law No. 48 of 1970, art. 30-4 (Japan). The Japanese government has attempted to provide clarity on the law.

[25] Copyright Act of 2021, No. 22, Part 5 Division 8 (Oct. 8, 2021) (Sing.), https://sso.agc.gov.sg/Acts-Supp/22-2021/Published/20211007?DocDate=20211007&WholeDoc=1 - pr243-.

Broad and vaguely worded exceptions, like the ones enacted in Japan and Singapore, fail to meet international treaty obligations because they prejudice the copyright interests of rightsholders and lack clear or appropriate safeguards that protect rightsholders. Such broad exceptions should be rejected. Even the viability and application of the other, more limited exceptions have been questioned and criticized by some rightsholders for failing to pass muster under the Berne Convention and will likely be tested in courts.[26] This is even more apparent in light of existing and quickly developing licensing markets for use of copyrighted works by AI developers and the inevitable market harm such exceptions would cause to copyright owners.[27] None of these currently existing exceptions should be considered or adopted in the United States.

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.

Subject to the explanatory notes and caveats set forth below, we do not believe that any amendment to the Copyright Act that is specifically targeted at artificial intelligence and would apply broadly to *all* copyright owners is needed at this time. To avoid any confusion as to our position, however, we provide the following, further clarifications and caveats.

We specifically reference the "Copyright Act" in our response, because (i) as we explain in our introductory remarks, the Copyright Alliance's mission does not extend beyond copyright law and therefore we take no position on the need for copyright-adjacent legislation, such as antitrust legislation or *sui generis* legislation that would address protections for style, image, likeness, voice, etc.; and (ii) as we discuss in detail in our responses to questions 15-17, legislation related to transparency is necessary, but it is unlikely that that such legislation would be specific to copyright and therefore require an amendment to the Copyright Act itself.

---

[26] Letter from Creators' Coalition to European Union (EU) (July 7, 2023), https://nwu.org/wp-content/uploads/2023/07/creators-coalition-AI-exceptions.pdf. ("But allowing these exceptions to be applied to copying for ingestion and reuse by generative AI systems constitutes a significant violation of the obligations of EU member states as parties to the Berne Convention2 and the WIPO Copyright Treaty.").

[27] *Id*.

We use the phrase "specifically targeted at artificial intelligence" to make clear that this statement is not intended to have any bearing on legislative initiatives we have supported in the past or will support in the future that are intended to have broad applicability and may also have implications for the relationship between copyright and AI. Some examples include our support for legislation related to no-fault injunctions and improvements to section 512 and the copyright management provision in section 1202.[28]

Perhaps most significantly, we use the phrase "would apply broadly to all copyright stakeholders" as a reference back to our response to question two. Each industry and each category of creative authorship is impacted by AI differently and thus a blanket change to the Copyright Act that impacts all copyright owners and creators without being narrowly tailored to a particular industry or type of copyrighted work is, in our view, unnecessary and inappropriate at this time.

At the same time, as we note in our response to question two, each type of copyright industry, creator, owner, and work is impacted by AI differently, and because the impact is different, the responses and solutions may need to be different. And thus, narrowly targeted legislation to amend the Copyright Act might be appropriate for a particular industry, group of creators, or type of work.[29] We leave that decision up to each industry. The Copyright Alliance itself is amenable to considering such legislation so long as: (i) it is narrowly targeted to a specific industry and type of works and would not directly or indirectly affect (through inadvertent consequences, or otherwise) those not intended to be covered by the legislation; (ii) there is a general consensus within that particular industry that legislation is necessary or appropriate; and (iii) the legislation

[28] For instance, we have long supported amendments to section 1202 to, among other things, provide that a copyright owner should only be required to prove that the information was removed or altered knowingly or recklessly, not that the copyright management information (CMI) was removed or altered with the knowledge that it would induce, enable, facilitate, or conceal infringement. In the AI context, this becomes even more important because it is crucial to maintain metadata which can be used to determine whether a work has been ingested by an AI system, and possibly to indicate the provenance of derivative works containing both AI and human-authored elements.

[29] For example, while songwriters are "authors," an approach supported by certain authors groups may not be appropriate for songwriters and, if so, should be drafted narrowly so as to not include them.

does not create a new copyright exception for AI training or use of copyrighted materials, a compulsory license, an opt-out approach for training or use, or allow copyright protection in outputs that are solely AI-generated.

Finally, when we use the phrase "at this time," it is in recognition of the numerous AI-related copyright infringement cases that are pending in the courts.[30] We certainly hope the courts will engage in a proper and comprehensive analysis of the legal issues in the case and reach the correct conclusions. But if they do not, it may be necessary or appropriate to revisit the legislative question and to enact narrowly focused legislation to correct a misinterpretation or misapplication of copyright law. This phrase should not be construed to mean that, if a particular industry or its creators are supporting industry specific-AI legislation that meets the criteria set out in the preceding paragraph, such legislation should not be considered.

It is worth noting that on February 24, 2023, the Congressional Research Service (CRS) published a report titled *Generative Artificial Intelligence and Copyright Law*,[31] exploring legal questions that courts and the U.S. Copyright Office are confronting with generative AI, including authorship and ingestion issues. The report suggests that though Congress may wish to consider whether copyright questions raised by generative AI warrants any amendments to the Copyright Act or other legislation, "given how little opportunity the courts and Copyright Office have had to address these issues," ultimately Congress may wish to adopt a wait-and-see approach as courts and the Copyright Office gain experience handling generative AI issues and cases. Since the initial publication of the report, CRS has updated the report twice (in May and September) to include developments that occurred since the initial release of the report.[32] The updates includes discussion about "Heart on My Sleeve" (the AI-generated song using the voices of Drake and The Weeknd), new class-action lawsuits brought by authors and visual artists, resolution of the Copyright Office registration dispute with visual artists Kristina Kashtanova and Jason Allen, the

[30] *See* Appendix A for a summary of the AI-related copyright cases presently pending in the courts.

[31] CONG. RSCH. SERV., LSB10922, GENERATIVE ARTIFICIAL INTELLIGENCE AND COPYRIGHT LAW (2023), https://crsreports.congress.gov/product/pdf/LSB/LSB10922.

[32] CONG. RSCH. SERV., LSB10922, GENERATIVE ARTIFICIAL INTELLIGENCE AND COPYRIGHT LAW: VERSIONS (Sept. 29, 2023), https://crsreports.congress.gov/product/details?prodcode=LSB10922.

District Court of D.C.'s affirmance of the Copyright Office's rejection of Dr. Stephen Thaler's registration application, and the Copyright Office's registration guidance.

TRAINING

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

Any kind of copyright-protected work that is available in digital copies can be ingested by AI models for training purposes. The primary ways that copyright-protected training materials are collected and ingested are: (1) scraping them off websites (where the works may have been posted either by the copyright holder, a third party—with the copyright owner's authority—or an infringer) and then including the scraped-copies in pre-processed datasets; (2) licensing them from rightsholders; and (3) using proprietary works owned by the AI developer.

Several major AI companies, including StabilityAI, OpenAI, and Meta, have trained their AI models by scraping and ingesting copyright-protected works from all over the internet, including copying pirated works that appear on rogue websites and circumventing firewalls in order to access copyrighted material on subscription-based websites.

Other AI developers of specialized models and companies who are developing in-house models have often taken a more responsible, ethical, and respectful approach by only ingesting works that they have licensed from copyright owners, works that are in the public domain, or their own proprietary materials. These sourcing methods are not mutually exclusive. For example, Adobe sources their training materials in a variety of ways including using copyrighted works from Adobe Stock images, openly licensed content, and public domain content.[33] We discuss licensing partnerships in more detail in future answers.

---

[33] *Adobe Unveils Firefly, a Family of New Creative Generative AI*, ADOBE: NEWS (Mar. 21, 2023), https://news.adobe.com/news/news-details/2023/Adobe-Unveils-Firefly-a-Family-of-new-Creative-Generative-AI/default.aspx, ("Adobe's first model, trained on Adobe Stock images, openly licensed content and public domain content where copyright has expired . . .").

(1) Scraped From the Internet

Some developers of large, general-purpose foundational AI models use datasets containing *billions* of copyright-protected works scraped from the internet. They also use datasets of specific types of copyrighted works—such as books in the case of large language models. The copyrighted works contained in the datasets may be harvested from the internet by the AI developers themselves, but in many other cases developers may use datasets created by third parties, as discussed in our response to 6.1. In the latter scenario, the scraping would be done by third parties using bots and web crawlers that scrape and ingest works that exist on the internet, including on pirate websites; or copyrighted works would be downloaded *en masse* from pirate sources. Nevertheless, in almost every case the developers would be "cleaning up," copying, and further processing the datasets during training.

Investigative journalists and outside researchers have discovered that some of the most popular training datasets contain large corpuses of illegal copies of copyright-protected works. Unfortunately, these illicit sourcing practices are not uncommon amongst AI developers, as revealed in investigations of popular LLMs such as OpenAI's ChatGPT, Meta's LLaMA, Google's Bard, and others.[34]

The Washington Post discovered that Google's C4 dataset—a version of the Common Crawl dataset—contains copyrighted works that are located behind a firewall on subscription-based websites like scribd.com and major news outlets including *The New York Times* and *Los Angeles Times*.[35] Moreover, this dataset also included pirated books scraped from b-ook.cc, a notorious

[34] In light of these revelations, AI companies have become increasingly secretive and less transparent about their training data, which underscores the need for increased transparency requirements for AI companies as discussed later in our comments. For example, Meta refused to disclose the details of how the second version of its LLaMA tool was trained. Sharon Goldman, *Generative AI Datasets Could Face a Reckoning*, VENTUREBEAT: THE AI BEAT (Aug. 21, 2023), https://venturebeat.com/ai/generative-ai-datasets-could-face-a-reckoning-the-ai-beat/.

[35] Kevin Schaul, Szu Yu Chen & Nitasha Tiku, *Inside the Secret List of Websites That Make AI Like ChatGPT Sound Smart*, THE WASH. POST: TECH (Apr. 19, 2023, 9:00 AM), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

pirate website connected to the Z-library network, which is under federal investigation and indictment for criminal copyright infringement.[36]

Unfortunately, this is not a one-time occurrence. Many other training datasets have the same issues, including a training set known as "The Pile,"[37] which included the sub-set "Books3"—a dataset of the full text of almost 200,000 books, scraped from the pirate tracker, Bibliotik.[38] This dataset is popular among AI companies and developers, and has been used to train models like Meta's LLaMA.[39] Current class action lawsuits against OpenAI and Meta include additional allegations of training datasets sourcing from pirate repositories and websites.[40]

In addition to using copyrighted works obtained from pirate sources, AI companies and third-party developers scrape websites belonging to individual creators, news outlets, stock image and footage companies, and tens of thousands of other sites operated by other copyright owners and ingest whatever copyrighted works are posted on the site without regard to the status of copyright protections in those works or to website terms and conditions.[41] While in theory it is possible for rightsholders to signal that they do not want their works scraped and used for AI training purposes through tools like robots.txt, or tagging works with industry-developed "do not

---

[36] *Id*.

[37] Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, ELEUTHERAI, https://pile.eleuther.ai/paper.pdf.

[38] The Books3 dataset was recently taken offline as per a takedown request by rightsholders. Alex Reisner, *Revealed: The Authors Whose Pirated Books Are Powering Generative AI*, THE ATL.: TECHNOLOGY (Aug. 19, 2023), https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/; *see also* Leah Asmelash, *These Books Are Being Used to Train AI. No One Told the Authors*, CNN (Oct. 8, 2023, 8:00 AM), https://www.cnn.com/2023/10/08/style/ai-books3-authors-nora-roberts-cec/index.html.

[39] Reisner, *supra* note 38. Books3 is not the only books dataset compiled from pirated books. It is just the only one on the open internet.

[40] *See, e.g.*, Kadrey v. Meta Platforms, Inc., 2023cv03417 (N.D. Ca. July 7, 2023); Tremblay v. OpenAI, Inc., 2023cv03223 (N.D. Ca. June 28, 2023); Silverman v. OpenAI, Inc., 2023cv03416 (N.D. Ca. July 7, 2023).

[41] *See e.g.*, Andy Baio, *AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability*, WAXY (Sept. 30, 2022), https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/.

train" credentials,[42] as we discuss in detail in our response to question 9.2, these tools are not very effective. Further, "do not train" tags are forward looking only and do nothing to address scraping and ingesting that's already occurred. We hope and anticipate that new tools will be created in the future to more effectively enable rightsholders to signal that they do not want their online works to be scraped and used for AI training purposes, but those tools will also be forward-looking and therefore will suffer from similar problems.[43]

Despite these flaws, various rightsholders have no other choice than to use these tools in a desperate attempt to protect their valuable copyrighted works and prevent AI developers from profiting off these works.[44] The reality is that the use of these tools has not stopped the widespread ingestion of their copyrighted works by AI companies. Of course, it goes without saying that these measures are completely ineffective to prevent scraping illegal copies from illicit, pirate websites. In that scenario, the author or copyright owner has no ability or authority to use such tools.

(2) Licensed from Copyright Owners

Voluntary licensing of copyrighted works for training material is not novel. Even before the explosion of generative AI, copyright owners offered machine learning licenses, particularly in

[42] Andy Parsons, *Reaching Major Milestones with 1,000 Members, Content Credentials in Adobe Firefly and Much More*, CONTENT AUTHENTICITY INITIATIVE (Apr. 3, 2023), https://contentauthenticity.org/blog/meeting-the-moment-with-c2pa-and-firefly.

[43] Ultimately, all these mechanisms represent imperfect ways for copyright owners to "opt out" of having their works scraped and ingested by generative AI developers. As we make clear in our responses to questions nine and its subparts, we oppose any opt out approach to generative AI training.

[44] *E.g.*, Ariel Bogle, *New York Times, CNN and Australia's ABC Block OpenAI's GPTBot Web Crawler from Accessing Content*, THE GUARDIAN (Aug. 24, 2023), https://www.theguardian.com/technology/2023/aug/25/new-york-times-cnn-and-abc-block-openais-gptbot-web-crawler-from-scraping-content; Dan Milmo, *The Guardian Blocks ChatGPT Owner OpenAI from Trawling its Content*, THE GUARDIAN (Sept. 1, 2023 12:54 PM), https://www.theguardian.com/technology/2023/sep/01/the-guardian-blocks-chatgpt-owner-openai-from-trawling-its-content; *see also* Nitasha Tiku, *Newspapers Want Payment for Articles Used to Power ChatGPT*, THE WASH. POST (Oct. 20, 2023, 5:51 AM), https://www.washingtonpost.com/technology/2023/10/20/artificial-intelligence-battle-online-data/.

the fields of scientific and academic journal publishing.[45] An increasing number of generative AI companies are entering into licensing deals with some rightsholders for the use of copyrighted works to train AI models. Some examples include:

- a partnership between OpenAI and The Associated Press ("AP");[46]
- AI-startups like Bria licensing works from rightsholders like Getty Images[47] and individual photographers and artists;[48]
- a partnership between Nvidia and Getty Images to build new generative AI technologies that ingest only fully licensed content;[49]
- a collaboration between IBM and Adobe to assist customers in implementing generative AI models based on Adobe's Firefly technology.[50]

However, as discussed in our responses to questions 2, 9.2, 10.4 and throughout our other responses, licensing and partnership opportunities are not available to all creators or to all

---

[45] *See generally* COPYRIGHT CLEARANCE CENTER, https://www.copyright.com/solutions-rightfind-xml/ (last visited Oct. 17, 2023) (offering access to licensed scientific articles for the purposes of training AI machines); CAMBRIDGE CORE, https://www.cambridge.org/core/open-research/text-and-data-mining (last visited Oct. 17, 2023); ELSEVIER, https://www.elsevier.com/about/policies/text-and-data-mining/elsevier-tdm-license (last visited Oct. 17, 2023) ("Text mining access for subscription content is provided to subscribers for non-commercial research purposes."); WILEY, https://onlinelibrary.wiley.com/library-info/resources/text-and-datamining (last visited Oct. 17, 2023) ("Academic subscribers can perform TDM under license . . .").

[46] Matt O'Brien, *ChatGPT-Maker OpenAI Signs Deal with AP to License News Stories*, AP (July 13, 2023, 11:41 AM) https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

[47] BRIA , https://bria.ai/ (last visited Oct. 27, 2023).

[48] Kyle Wiggers, *This Startup Wants to Train Art-Generating AI Strictly on Licensed Images*, TECHCRUNCH (Apr. 13, 2023, 8:30 AM) https://techcrunch.com/2023/04/13/this-startup-wants-to-train-art-generating-ai-strictly-on-licensed-images/.

[49] Rick Merritt, *Moving Pictures: NVIDIA, Getty Images Collaborate on Generative AI*, NVIDIA (Mar. 21, 2023). https://blogs.nvidia.com/blog/2023/03/21/generative-ai-getty-images/.

[50] *IBM Expands Partnership with Adobe To Deliver Content Supply Chain Solution Using Generative AI*, IMB NEWSROOM (June 19, 2023) https://newsroom.ibm.com/2023-06-19-IBM-Expands-Partnership-with-Adobe-To-Deliver-Content-Supply-Chain-Solution-Using-Generative-AI.

creative industries. This is especially true for independent creators who often lack the resources and negotiating power to deal with AI developers. Finally, when examining the varied licensing practices of copyrighted works, it is vital that copyright owners' choice whether to license works for AI ingestion be respected. We discuss this core principle more in our responses to questions 9, 10, and their respective subparts, and in our general principles outlined earlier.

(3) Proprietary Works

Copyright owners not only provide valuable training inputs for AI but are also actively developing AI technologies of their own. In doing so, these developers utilize proprietary works, sometimes in combination with other sources, to develop training datasets for their AI models. Getty Images recently announced a partnership with major tech-manufacturer Nvidia to develop two generative AI models exclusively trained on Getty-owned images.[51] As previously mentioned, Adobe's image generative AI model, Firefly, utilizes Adobe Stock images as part of its training datasets.[52] Shutterstock's recent six-year agreement with OpenAI provides for using Shutterstock images, video and music libraries, and associated metadata to develop OpenAI products and Shutterstock's own AI tools.[53]

> ***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 acquire materials or datasets in the ways we detail above in our answer to question six—but ultimately, they can and do acquire them anyway from anywhere. There are widely popular, prepared datasets, like The Pile, Books3, LAION, and WebVid-10M datasets, which are

---

[51] Merritt, supra note 49; GETTY IMAGES, https://www.gettyimages.com/ai/generation/about (last visited Oct. 17, 2023).

[52] *Adobe Unveils Firefly, a Family of New Creative Generative AI*, ADOBE: NEWS (Mar. 21, 2023), https://news.adobe.com/news/news-details/2023/Adobe-Unveils-Firefly-a-Family-of-new-Creative-Generative-AI/default.aspx.

[53] *Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data*, SHUTTERSTOCK (July 11, 2023), https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year.

uploaded to and shared on various websites and repositories.[54] As discussed above, the problem is that these datasets have been proven to contain illegal, pirated copies of entire copyrighted works (or links to them). When these stolen works are indiscriminately scraped by AI developers, the harm to copyright owners is exacerbated. We discuss this compounded threat to copyright owners more in response to question eight.

In response to the second part of this question, the supply chain in the AI training process does not look the same across the board for every generative AI tool. An AI company could follow links in a dataset to perform scraping and internally develop training datasets on their own, outsource such activities to a third party, and further alter, add to, or subtract from prepared datasets, subject to any terms and conditions set on the datasets by the dataset developers.

A practice commonly referred to as "data laundering" is worth mentioning here as it blurs the distinction between noncommercial, research uses and commercial uses of copyrighted works. Data laundering entails private, commercial AI companies funding research or nonprofit institutions to develop training datasets and sometimes even the AI tools themselves, which often use copyright-protected works, under the guise of supporting noncommercial research activities. Once these training sets or models are developed, the funding AI company then uses them to develop proprietary commercial AI platforms.[55] AI developers may engage in this kind of activity in an effort to avoid infringement liability that would otherwise attach to clearly commercial, unauthorized use of copyrighted materials. By funding these endeavors, commercial AI developers aim to avail themselves of broader exceptions in copyright law (such as fair

[54] Recently, several websites including *The Eye* and *Academic Torrents* took down the Books3 dataset upon Digital Millennium Copyright Act (DMCA) notices sent by a Danish anti-piracy group, The Rights Alliance, upon the organization's discovery that the dataset included pirated copies of at least 150 works of authors they represented. Kate Knibbs, *The Battle Over Books3 Could Change AI Forever*, WIRED (Sept. 4, 2023, 6:00 AM) https://www.wired.com/story/battle-over-books3/.

[55] Viki Auslender, *Why Meta's Open Source Is Not Really Open*, CTECH (Aug. 3, 2023, 8:23 AM), https://www.calcalistech.com/ctechnews/article/atv6xnkya.

use)—sometimes due to the fact that they operate in countries with broader copyright law exceptions or limitations.[56]

Another way that some AI developers blur the line between research and clear commercial activity is by initially developing AI models for research use that are later incorporated into commercial products that are licensed to others. For example, the first version of Meta's LLaMA AI model was released for research use[57] but has subsequently shifted into an "open-source licensing" model. Under current open-source licensing terms for LLaMA2, Meta requires "special licensing" for applications or services with more than 700 million users—illustrating how commercial AI companies blur the lines between research and commercial purposes to do an end-run around copyright infringement liability for their own commercial gain.[58]

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?

As discussed in our answer to question six, there are examples of large rightsholders licensing their copyrighted works for commercial AI uses, and the AI licensing market for copyrighted works continues to grow. The types of AI licenses, including the terms and conditions, being offered differ greatly from industry to industry and are evolving rapidly, and thus our members are better able to speak to their particular licensing experiences and business models.

---

[56] Andy Baio, *AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability*, WAXY (Sept. 30, 2022), https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/. For example, StabilityAI—a for-profit corporation from its inception—was a major donor of the work done by researchers at the Ludwig Maximilian University of Munich, who essentially developed the Stable Diffusion product. 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, 1:38 PM), https://www.forbes.com/sites/kenrickcai/2022/09/07/stability-ai-funding-round-1-billion-valuation-stable-diffusion-text-to-image/?sh=7256acd624d6. Moreover, StabilityAI funded the German nonprofit organization, Large-scale Artificial Intelligence Open Network (LAION), to create LAION 5B, a training dataset containing image-text pairs of 5.6 billion images from the entire internet, which enabled StabilityAI to scrape those billions of images to train its AI models. GITHUB, https://github.com/CompVis/stable-diffusion (last visited Oct. 17, 2023).

[57] Shirin Ghaffary, *Why Meta is Giving Away its Extremely Powerful AI Model*, VOX (July 28, 2923, 6:00 AM), https://www.vox.com/technology/2023/7/28/23809028/ai-artificial-intelligence-open-closed-meta-mark-zuckerberg-sam-altman-open-ai.

[58] *Llama 2 Community License Agreement*, META AI (July 18, 2023), https://ai.meta.com/llama/license/ ("If . . . the monthly active users of the products or services made available by or for Licensee [] is greater than 700 million monthly active users . . . you must request a license from Meta . . . .").

Although the AI licensing market is showing welcome signs of growth, the overwhelming majority of copyrighted works that are ingested for training have not yet been licensed and not all creators and rightsholders have had the same successes in reaching licensing deals with AI companies. In particular, many AI companies have been slow or unwilling to license works from independent creators. As history has shown us, creators and copyright owners are usually willing to license their works on reasonable terms for reasonable fees; that is, of course, how creators typically earn a living. Copyrighted works provide immense value to AI developers, and they can and should pay for that value. When properly applied, copyright law sets the conditions for the market to prevail. The marketplace should continue to properly value and incentivize creativity, and AI policy should not interfere with the right or ability of copyright owners to license, or choose not to license, their works for AI purposes.

6.3. To what extent is noncopyrighted 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 in our response to question six, we know that AI companies ingest public domain works to train AI models. However, we do not know the extent to which public domain materials are being used for AI training purposes other than to conclude that: (i) the level of dependence on public domain works for AI training will likely differ among the various AI models because each AI company desires different qualities and characteristics for their AI outputs; and (ii) because public domain works may contain outdated information, outdated language and styles, biased information, and other traits that could negatively influence the quality of the output, most AI developers do not solely rely on public domain works to train their models and instead ingest copyright-protected works because they lead to higher-quality outputs.

Some developers of special-purpose models create or commission their training materials. As discussed in our answers above, there are AI developers that use proprietary works as training materials to develop their own AI models, commercial AI companies that fund non-commercial or research institutions to develop training materials on their behalf (which is data laundering), and AI companies that take prepared datasets and further develop them or develop subsequent

training materials that contain copyright-protected works. Some of our members, including stock images licensors and publishers, have produced custom datasets for AI developers to license.

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.

Retention practices among AI companies vary. As we understand it, some AI companies delete copyrighted works used in training datasets after copying and ingesting them, while others choose to store and retain them. Importantly, whether an AI company deletes or retains the copyrighted works in a dataset after copying them is wholly irrelevant to any copyright infringement analysis. During the AI ingestion process, when a copyrighted work is ingested without a license from the copyright owner, a copy of the entire work is made, thereby infringing the copyright owner's right to control the reproduction of the work under section 106(1) (absent a valid defense). Significantly, there is no requirement that a copy of the work be retained or stored in order for the use to be deemed infringing. (This is discussed in more detail in our response to question eight.)

Although storage of copyrighted works is not relevant to an infringement analysis as a general matter, significant security concerns are raised when copyright-protected training materials are retained and stored by AI companies. We live in an age where cyberattacks and mass online piracy is the norm. Therefore, it is critical for AI companies and/or entities that curate datasets to employ stringent security measures and safeguards to prevent cyberattacks that lead to the retained works being leaked and to prevent misuse of the retained copies.

This is not a new issue. In its fair use analysis in the Google Books case, the Court of Appeals for the Second Circuit took note of how Google took significant safeguards to secure the copies of books it used in its database, such as only showing "snippets" of works to highlight a search term and implementing anti-hacking measures.[59] Due to these safeguards, the court concluded that there was little risk that Google's actions could serve as a substitute for the copied works.

[59] Authors Guild v. Google, Inc., 804 F.3d 202, 226–29 (2d Cir. 2015).

Implementation of these safeguards were one of several essential elements that led to the court's ultimate finding of fair use in the case.

In the generative AI context, it would appear that such safeguards are rarely implemented.[60] If works retained by AI companies are not secure, the AI dataset can be hacked, and the copyrighted works can be leaked. The harm to copyright owners if that were to occur would be catastrophic. Therefore, it is crucial that AI companies and dataset curators employ safeguards to prevent the mass piracy of copyrighted works that have been stored by the AI companies.

When copyrighted works are licensed, whether and how the works are retained and the security measures an AI company must take to protect the works from being leaked can all be set forth in the license agreement. Additionally, licenses can convey worldwide rights, which is yet another benefit of AI companies licensing copyrighted works.

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.

The training process for generative AI typically involves wholesale copying of ingested copyrighted works. In many cases, leading general-purpose AI companies work with vendors that employ vast numbers of people around the world to sort, tag, "annotate," and otherwise process massive amounts of material, which include copyright-protected works, that are being

[60] For example, by allowing for prompts that are "in the style of" an author or artist or by allowing prompts including copyrighted characters.

ingested into the AI system.[61] The fact that a copy of a work is perceptible to human employees who review and sort them is clear evidence that a copy (as defined by the Copyright Act) is being made at some point in the ingestion process, and that (absent a valid defense) a copyright owner's right of reproduction (among others) is being violated. Even when humans are not directly involved in the ingestion process, a computer making copies of the ingested works is sufficient to satisfy the definition of copying in the Copyright Act because the copies last for "more than transitory duration."[62]

Training sets themselves may contain links to copyright-protected works (as discussed in our answers to questions six and its subparts, this includes works posted on rightsholders' websites or pirated works found on rogue websites and elsewhere) or lists of cloud-based files containing copyright-protected works (again, as discussed in our responses to questions six and its subparts, pirated works have been found in such files). But even when the training sets only include links, at some point in the AI training process, the works that are being linked to must be copied in order to be processed for ingestion. That act of making a copy of the protected work is the *sine qua non* of the reproduction right and satisfies the definition of copying in the Copyright Act. The Copyright Act makes clear that a copy is made whenever a work is fixed and "can be perceived, reproduced, or otherwise communicated."[63] The definition goes on to say that such perception can occur "either directly or with the aid of a machine or device."[64] This is a very low standard that is easily met.

---

[61] Josh Dzieza, *AI Is a Lot of Work*, THE VERGE (June 20, 2023), https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots.

[62] Transcript from Online Webinar, *International Copyright Issues and Artificial Intelligence*, U.S. COPYRIGHT OFF. (July 26, 2023), https://www.copyright.gov/events/international-ai-copyright-webinar/International-Copyright-Issues-and-Artificial-Intelligence.pdf; *id.* at 11 ("Jane Ginsburg: On transient copying, I don't think that the AI training data would meet the criteria of the Article 5(1) of the EU Infosoc Directive . . . it's not clear under U.S. law whether a transient copying approach would apply. It's not an exception because if the copying is too transient, it doesn't count as copying."); *id.* at 12 ("Matthew Sag: I think that both in the EU and the U.S. there's no way this falls under transient copying. You know, if you actually look at the mechanics of how you—like how machine-learning training works, like people aren't storing files or parts of files for anything you would measure in seconds or fractions of seconds. They're storing them for months.").

[63] 17 U.S.C. § 101.

[64] *Id*.

A common argument that some AI developers make is that even if copying occurs, the only things being copied are unprotectable facts. This position is based on the assertion that when datasets are created for AI training purposes, expressive works of authorship are reduced into mere "data" about the "relationships" between elements of a work that is then processed by an algorithm. However, even if works are eventually converted into binary code, that would not excuse copies of the works being made—as they exist in their original form. Moreover, simply because a work is converted into a format that can more easily be ingested by an AI system does not mean that the work suddenly ceases to include copyrightable expression or loses copyright protection. As discussed elsewhere in our comments, what some AI developers consider to be unprotectable "data" is actually protected copyrightable expression.[65] Those AI developers also argue that they are merely copying the "relationships" of different elements of a work. Inaccurately classifying copyrightable expression as "data" would eviscerate well-established tenets of copyright law that have existed for well over the past two centuries.

Finally, some claim that AI systems merely "observe" copyrighted works, rather than copy them. But that is not how computers work and is an attempt to humanize the functions of a machine and avoid liability.[66] AI can be a powerful tool for creativity—and many Copyright Alliance members are already using different AI tools in service of their own artistry. But, simply put, AI is not human. At the very least, where a copy of an ingested work is made in the random-access memory (RAM) of the computer system, courts have made clear that that copy qualifies as a reproduction under section 106(1).[67]

---

[65] Lee, Katherine et al., *Talkin' 'Bout AI Generation: Copyright and The Generative-AI Supply Chain*, arxiv.org, Sept. 14, 2023, available at https://arxiv.org/ftp/arxiv/papers/2309/2309.08133.pdf. ("[T]he works that have been transformed into data have copyrights. In turn, for generative-AI systems that generate potentially copyright-infringing material, the training data itself will often include copyrightable expression.").

[66] See Kailey Jacomet, *Legal Issues with Using AI-Generated Content in Your Business*, CONTRACTISTA (June 25, 2023) (quoting ChatGPT that works are "observed during training"), https://www.contractista.com/blogs/the-contractista-blog/legal-issues-with-using-ai-generated-content-in-your-business; Rob Enderle, *The Problem With Suing Gen AI Companies for Copyright Infringement,* TECH NEWS WORLD (July 17, 2023) ("AIs observe digitized data at a massive scale that renders individual contributors unidentifiable. This observation process leads to the formation of an amalgamated knowledge that constitutes the AI's brain."), https://www.technewsworld.com/story/the-problem-with-suing-gen-ai-companies-for-copyright-infringement-178470.html.

[67] MAI Sys. Corp. v. Peak Comput., Inc., 991 F.2d 511, 518 (9th Cir. 1993) (finding that "MAI has adequately shown that the representation created in the RAM is 'sufficiently permanent or stable to permit it to be perceived,

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

In order to make "inferences," an AI model must extract copyrightable expression from copyrighted works, as discussed in our response to question 7.1. While AI developers claim they are merely extracting unprotectable data or facts from the training data, that is incorrect. What they are extracting are valuable, expressive elements of a work that merit copyright protection. For example, in the case of a literary work, the words an author chooses to express herself, the relationship of those words into a sentence, the relationship of that sentence to other sentences within a paragraph, and so on, represent that author's copyrightable expression. It is that type of expression that makes literary works protectable under copyright. The same is true for music, images, audiovisual works, and other copyrighted works. Simply because these copyrighted works can be processed into an AI model does not mean that the work should lose its protection or that a copyright owner should lose their right to control or be compensated for use of that work. There are several examples of AI models being prompted to reproduce almost verbatim text from ingested books, song lyrics or reproducing ingested pictures, further supporting the notion that these works are embedded in the model itself, to varying degrees.[68]

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?

There is a continuing debate about whether an AI model that has been trained on a work can be retrained to "unlearn" inferences that it gained from training on that work. Some indicate that, at

---

reproduced, or otherwise communicated for a period of more than transitory duration'"). While there is an exception in section 117 of the Copyright Act that excuses the making of RAM copies that are made as part of computer maintenance or repair, that exception is very narrow and does not apply to the unauthorized ingestion of copyrighted works by AI systems.

[68] *See, e.g.*, Universal Music–Z Songs v. Anthropic, PBC, 23cv10192 (M.D. Tenn. Oct. 18, 2023); Getty Images, Inc. v. Stability AI, Inc., 23cv00135 (D. Del. Feb. 3, 2023); Kadrey v. Meta Platforms, Inc., 2023cv03417 (N.D. Ca. July 7, 2023); Tremblay v. OpenAI, Inc., 2023cv03223 (N.D. Ca. June 28, 2023); Silverman v. OpenAI, Inc., 2023cv03416 (N.D. Ca. July 7, 2023).

present, the model's unlearning is challenging,[69] though technologies could develop in the future to change this.[70] Others claim that a degree of unlearning is possible. They point to methods, like algorithmic disgorgement, that may be used to decrease the impact a particular copyrighted work has on the AI algorithm and generated outputs.

Still others claim that AI models can be fully retrained, but that doing so would be expensive. In the event an AI model can be retrained in whole or in part, the fact that it may be expensive or otherwise burdensome to do so is evidence that licensing from the outset is the better and most cost-effective option. Further, it is not known whether or how "unlearning" methods can be validated to confirm whether the effects of a particular copyright-protected work are truly scrubbed out of an AI's algorithms.[71]

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?

It is a well-documented phenomena that generative AI models can be prompted to generate output that replicates particular copyrighted works that were used to train the AI.[72] This is a universal problem affecting the spectrum of copyright-protected works including images, songs, literary works, and computer code. When a AI model generates output that reproduces copyrighted content, it is a clear sign that models are trained on more than mere uncopyrightable

---

[69] *See* Jie Xu et al., *Machine Unlearning: Solutions and Challenges*, CORNELL UNIV. ARXIV (Aug. 14, 2023), https://arxiv.org/pdf/2308.07061.pdf ("Machine unlearning faces challenges from inherent properties of ML models as well as practical implementation issues.").

[70] *See* Nicholas Carlini et al., *The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks*, CORNELL UNIV. ARXIV (July 16, 2019), https://arxiv.org/pdf/1802.08232v3.pdf; Xulong Zhang et al., *Machine Unlearning Methodology Base on Stochastic Teacher Network*, CORNELL UNIV. ARXIV (Aug. 28, 2023), https://arxiv.org/pdf/2308.14322.pdf (proposing "using a stochastic network as a teacher to expedite the mitigation of the influence caused by forgotten data on the model"); Ronen Eldan & Mark Russinovich, *Who's Harry Potter? Approximate Unlearning in LLMs*, CORNELL UNIV. ARXIV (Oct. 4, 2023), https://arxiv.org/pdf/2310.02238.pdf (acknowledging that large language models (LLMs) "are trained on massive internet corpora that often contains copyright infringing content" and they propose a "novel technique for unlearning a subset of the training data from an LLM, without having to retrain it from scratch").

[71] Matthew Duffin, *Machine Unlearning: The Critical Art of Teaching AI to Forget*, VENTUREBEAT (Aug. 12, 2023), https://venturebeat.com/ai/machine-unlearning-the-critical-art-of-teaching-ai-to-forget/.

[72] Gowthami Somepalli et al., *Diffusion Art of Digital Forgery? Investigating Data Replication in Diffusion Models*, CORNELL UNIV. ARXIV (Dec. 12, 2022), https://arxiv.org/pdf/2212.03860.pdf.

"data" and that the expressive elements of particular works are being ingested as training materials.[73] It is also important to recognize that even if a generative AI model does not reproduce copyrighted content, it does not mean that no copyrightable content was ingested—it may only be an indication that the developer employed a measure to prevent it.

Many Copyright Alliance members have tested different generative AI models and have confirmed this phenomenon, as they have been able to prompt generative AI models to produce identical copies of copyright-protected works. For example, one of our members reported that earlier this year, they were able to prompt ChatGPT with "What are the lyrics to [song] by [artist]" to generate verbatim song lyrics of multiple songs they own. Furthermore, all of the lawsuits brought against AI companies rely in part on the allegations that the models were able to reproduce works or produce detailed derivative works (for e.g., summaries and outlines for sequels of ingested books).[74] Some AI companies have started to implement safeguards to prevent prompts that will lead to outputs that are identical or substantially similar to copyrighted works they are trained on, but most have not.[75] Even where models prevent such prompts, that only addresses the issue of outputs, it does not alter the landscape with respect to the unauthorized reproduction of copyrighted works as part of the ingestion process.

There are tools, such as "Have I Been Trained,"[76] that can be used to help rightsholders discover whether their works have been used to train AI models. But there is still significant progress to be made in this area. Researchers have also developed methods of introducing watermarked

---

[73] *See* Carlini et al., *Extracting Training Data from Diffusion Models*, CORNELL UNIV. ARXIV (Jan. 30, 2023), https://browse.arxiv.org/pdf/2301.13188.pdf ("For example, memorizing and regenerating copyrighted text and source code has been pointed to as indicators of potential copyright infringement.").

[74] *See, e.g.*, Kadrey v. Meta Platforms, Inc., 2023cv03417 (N.D. Ca. July 7, 2023); Tremblay v. OpenAI, Inc., 2023cv03223 (N.D. Ca. June 28, 2023); Silverman v. OpenAI, Inc., 2023cv03416 (N.D. Ca. July 7, 2023); Authors Guild v. OpenAI, Inc., 23cv08292 (S.D.N.Y. Sept. 19, 2023).

[75] Brad Smith & Hossein Nowbar, *Microsoft Announces New Copilot Copyright Commitment for Customers*, MICROSOFT (Sept. 7, 2023), https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/.

[76] *Have I Been Trained?*, SPAWNING, https://haveibeentrained.com/ (last visited Oct. 25, 2023).

works into training datasets to track AI models that use works without authorization.[77] But there are some drawbacks to these tools including how they can interfere with the operability of the AI algorithm.[78]

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.

When considering fair use and generative AI systems, it's important to understand that the ingestion of copyrighted material by generative AI systems is *not* categorically a fair use. Determining whether a particular use qualifies for the fair use defense to infringement has always required a fact-specific inquiry that is considered on a case-by-case basis.

While the preamble of section 107 of the Copyright Act provides examples of uses that are more likely to be a fair use,[79] even those examples do not categorically qualify as fair use. Courts will need to evaluate fair use defenses involving AI systems the same way they evaluate fair use in all contexts: by applying the four factors set forth in section 107 of the Copyright Act to the specific uses at issue. Blanket assertions that the ingestion of copyrighted works by an AI system should always qualify as fair use are legally inaccurate,[80] and a categorical exception for the broad

---

[77] *See, e.g.*, Ruixiang Tang et al., *Did You Train on My Dataset? Towards Public Dataset Protection with Clean-Label Backdoor Watermarking*, CORNELL UNIV. ARXIV (Apr. 10, 2023), https://browse.arxiv.org/pdf/2303.11470.pdf.

[78] *Id*.

[79] The six examples in section 107 include "criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, [and] research." 17 U.S.C. § 107.

[80] *See e.g.*, OpenAI, LP, Comments on USPTO's Request for Comments on Intell. Prop. Prot. for A.I. Innovation 4, 8 (Oct. 30, 2019), https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf ("[P]roper application of fair use factors requires a finding of fair use, especially considering the highly transformative nature of training AI systems. . . . Prior cases have generally supported a finding of fair use for uses of large digital corpora that were less transformative than the training of AI systems. A fortiori, training AI systems should be considered fair use."); *Artificial Intelligence and Intellectual Property: Part I—Interoperability of AI and Copyright Law*: *Hearing Before the Subcomm. on Cts., Intell. Prop., & the Internet of the H. Comm. on the Judiciary*, 118th Cong. (2023) (written testimony of Chris Callison-Burch, Assoc. Professor of Comput. & Info. Sci., Univ. of Penn.),

notion of "training" AI would betray the flexible, fact-specific fair use analyses that our copyright system has long relied on. With those words as an instructive framework, when considering fair use in the generative AI context, the typical commercial system's ingestion of copyrighted works is particularly unlikely to qualify as fair use when the AI system generates competing works.

Before engaging in any fair use analysis, it is crucial to distinguish infringement and fair use issues related to ingestion of copyrighted works by AI from those related to AI-generated output. Copyright owners are granted certain exclusive rights in their works under U.S. copyright law. These rights include the right to reproduce (i.e., copy) the work, to prepare derivative works, to distribute the work, to perform the work publicly, and to display the work publicly.[81] When a third party engages in one or more of those actions, that party is liable for infringement unless either (i) there is an applicable exception in the copyright law that excuses the specific infringement(s) at issue, such as fair use; or (ii) that person is acting with the authorization of the copyright owner, such as when they have a license (or permissible sublicense) that permits them to engage in the otherwise-infringing act(s).

In the AI environment there are at least two potential infringements: (1) infringement that occurs during the ingestion process;[82] and (2) infringement that occurs during the output stage—when a work is generated by the AI system. During the ingestion process, when a copyrighted work is used without a license from the copyright owner, a copy of some or all of the work is made,

---

https://judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/callison-burch-testimony-sm.pdf ("In considering whether pre-training AI systems on copyright is fair use, it is important to highlight that the copying of works at this stage is 'non-expressive' in the same way that is for making a copy of a work in other digital media. Pre-training also has a transformative nature . . . I find there is a compelling argument that training AI systems on copyrighted works is fair use under US copyright law."); *Artificial Intelligence and Intellectual Property—Part II: Copyright, Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. 8 (2023) (written testimony of Ben Brooks, Head of Pub. Pol'y Policy, Stability AI), https://www.judiciary.senate.gov/imo/media/doc/2023-07-12_pm_-_testimony_-_brooks.pdf ("Training these models is an acceptable, transformative, and socially-beneficial use of existing content that is protected by the fair use doctrine and furthers the objectives of copyright law . . .").

[81] 17 U.S.C. § 106.

[82] Ingestion for machine learning refers to the process of collecting and preparing data for use in machine learning models.

thereby infringing the copyright owner's right to control the reproduction of the work under section 106(1) (absent a valid defense, such as fair use). In practice, derivative copies of a work may also be made during the ingestion process, in order to increase the total amount of training material (e.g., images may be flipped or cropped). At the output stage, the AI system might generate an output that is substantially similar[83] to an ingested copyrighted work, potentially infringing not only the copyright owner's reproduction right, but also the adaptation right in section 106(2) and the distribution right under section 106(3) (again, absent a valid defense).

It is important to distinguish between infringements that occur during the ingestion stage and the output stage because the legal analysis necessary to determine whether an infringement has taken place is different. And because the use and the user are different, the fair use analysis will also be different. Thus, when determining whether an infringement occurs during the ingestion stage, there is no need to compare the ingested work to the AI-generated output to determine if the two are substantially similar because an identical copy of the work is being made during ingestion, and thus the substantial similarity test is not relevant to the legal analysis for ingestion.

Some have argued that because the output of generative AI is generally not substantially similar to any particular copyrighted work that is ingested, there is no infringement taking place.[84] But that argument ignores the fact that the right of reproduction enumerated in section 106 of the Copyright Act is a stand-alone right—meaning that it can be violated through unauthorized copying regardless of whether there is a separate act of infringement on the output side.[85] So, while the question of whether any given generative AI *output* is infringing will depend on specific circumstances and might involve a determination of substantial similarity, making unauthorized copies *during the ingestion process* would constitute a distinct violation of a

[83] When the copyrighted work and the alleged infringing work are not identical to one another, the test to determine whether an infringement has occurred is whether the two works are "substantially similar."

[84] *See Hearing on Artificial Intelligence and Intellectual Property: Part I – Interoperability of AI and Copyright Law 2023 Before the H. Subcomm. on Cts., Intell. Prop., & the Internet*, 118th Cong. 8 (2023) (statement of Sy Damle), https://judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/damle-testimony.pdf ("But absent some aberration in the training data or model design (as discussed below), the output will not be a "copy" of (i.e., substantially similar to) any individual work on which the model has been trained.").

[85] 17 U.S.C. § 106(1).

copyright owner's exclusive right of reproduction (absent an applicable defense or a license to use the work), and the substantial similarity of the output is irrelevant at that stage.

It is also important to recognize that for a copyright owner's right of reproduction to be violated, there is no requirement that a copy of the work be downloaded, retained, or stored. The Copyright Act makes clear that a copy is made whenever a work is fixed and "can be perceived, reproduced, or otherwise communicated."[86] The definition goes on to say that such perception can occur "either directly or with the aid of a machine or device."[87] The only requirement is that the copy exists for "more than transitory duration."[88] This is a very low standard that is easily met.[89]

As noted above in response to question 7.1, leading AI companies work with vendors that employ vast numbers of people around the world to sort, tag, and "annotate" massive amounts of material that are being ingested into the AI system.[90] The fact that a copy of the work is perceptible to human employees who review and sort the works is clear evidence that a copy is being made at some point in the ingestion process, and that (absent a defense) a copyright owner's right of reproduction is being violated. Even when humans are not involved in the process, computers making copies of the ingested works is sufficient to satisfy the definition of copying in the Copyright Act. We now turn to an analysis of the four factors.

---

[86] 17 U.S.C. § 101.

[87] *Id*. § 101.

[88] *Id*.

[89] *See* Transcript from Online Webinar, *International Copyright Issues and Artificial Intelligence*, U.S. COPYRIGHT OFF. (July 26, 2023), https://www.copyright.gov/events/international-ai-copyright-webinar/International-Copyright-Issues-and-Artificial-Intelligence.pdf; *id.* at 11 ("Jane Ginsburg: On transient copying, I don't think that the AI training data would meet the criteria of the Article 5(1) of the EU Infosoc Directive . . . it's not clear under U.S. law whether a transient copying approach would apply. It's not an exception because if the copying is too transient, it doesn't count as copying."); *id.* at 12 ("Matthew Sag: I think that both in the EU and the U.S. there's no way this falls under transient copying. You know, if you actually look at the mechanics of how you—like how machine-learning training works, like people aren't storing files or parts of files for anything you would measure in seconds or fractions of seconds. They're storing them for months.").

[90] Josh Dzieza, *AI Is a Lot of Work*, THE VERGE (June 20, 2023), https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots.

Factor One: Purpose and Character of the Use

The first factor considers how the party claiming fair use is using the copyrighted work. Some AI developers have taken the position that the unauthorized ingestion of copyrighted works for purposes of training AI systems constitutes a transformative use under the first fair use factor, and that this supposed transformative use standing alone is enough to categorically qualify AI ingestion as fair use under section 107.[91] In the case of some AI tools that generate music, images, or written material, the copyrighted works being ingested are sound recordings, works of visual art, and literary works, and the output generated by these AI systems will typically *serve the same purpose* as the works ingested. For example, consider a music model and consider what the purpose of the ingested music is to those who created the music. It's for the end user to listen to and enjoy for all of the myriad reasons that humans seek out recorded music. Now consider what the purpose of ingesting the copyrighted works is to the AI developer. The AI developer's purpose is to train the AI to generate music that the end user can listen to and enjoy. The purposes of the ingested works and the AI-generated outputs are the same.[92] And where the

[91] *See e.g.*, OpenAI, LP, Comments on USPTO's Request for Comments on Intell. Prop. Prot. for A.I. Innovation 4, 8 (Oct. 30, 2019), https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf ("[P]roper application of fair use factors requires a finding of fair use, especially considering the highly transformative nature of training AI systems. . . . Prior cases have generally supported a finding of fair use for uses of large digital corpora that were less transformative than the training of AI systems. A fortiori, training AI systems should be considered fair use."); *Artificial Intelligence and Intellectual Property: Part I—Interoperability of AI and Copyright Law*: *Hearing Before the Subcomm. on Cts., Intell. Prop., & the Internet of the H. Comm. on the Judiciary*, 118th Cong. (2023) (written testimony of Chris Callison-Burch, Assoc. Professor of Comput. & Info. Sci., Univ. of Penn.), https://judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/callison-burch-testimony-sm.pdf ("In considering whether pre-training AI systems on copyright is fair use, it is important to highlight that the copying of works at this stage is 'non-expressive' in the same way that is for making a copy of a work in other digital media. Pre-training also has a transformative nature . . . I find there is a compelling argument that training AI systems on copyrighted works is fair use under US copyright law."); *Artificial Intelligence and Intellectual Property—Part II: Copyright, Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. 8 (2023) (written testimony of Ben Brooks, Head of Pub. Pol'y Policy, Stability AI), https://www.judiciary.senate.gov/imo/media/doc/2023-07-12_pm_-_testimony_-_brooks.pdf ("Training these models is an acceptable, transformative, and socially-beneficial use of existing content that is protected by the fair use doctrine and furthers the objectives of copyright law . . .").

[92] As noted above, the fair use analysis for the ingestion of copyrighted works to train AI models and the output of generative AI are different. The purpose of the ingestion is determined by the AI developer. In the example, that purpose is to train the AI system to generate music. On the other hand, the purpose of the output is determined by the prompter. In the example, the prompter's purpose may be to sell music, to educate themselves on how to make music, or any number of things.

purpose of the defendant's use is the same as the plaintiff's, a court is unlikely to conclude that the use is transformative.[93]

Some AI companies might argue that the purpose of their ingestion is to simply train the AI, and that training represents a transformative use. But "training" standing alone is unlikely to constitute a sufficiently transformative use or purpose any more than "licensing" standing alone was a transformative use or purpose in *Warhol* or *"*sampling" standing alone was a transformative use or purpose in *Campbell*. As we discuss more in response to question 8.1, whether a use is transformative is closely related to the "justification" for the use, which requires an analysis of the purpose of the use.

To determine the purpose of the use, one must consider why the action is being done, and whether there is a justification for that action. For example, in *Warhol*, the use was not simply "commercial licensing"[94] but rather licensing for a story about Prince (the celebrity in the image), which was the same use that the copyright owner, Goldsmith, was using her images for.[95] The court in *Warhol* made clear that if the licensed use was for use in an article about Warhol's art, then perhaps the uses and the fair use analysis would be different and there might have been more of a justification.[96]

[93] In fact, this is why AI developers often discuss substantial similarity here–because their arguments on transformation are weak.

[94] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 143 S. Ct. 1258, 1278 n.11 (2023) ("The Court does not define the purpose as simply 'commercial' or 'commercial licensing.'").

[95] *Id.* at 1281 n.15 ("Both Goldsmith and AWF sold images of Prince (AWF's copying Goldsmith's) to magazines to illustrate stories about the celebrity, which is the typical use made of Goldsmith's photographs.")

[96] *Id.* at 1281 n.15 ("Both Goldsmith and AWF sold images of Prince (AWF's copying Goldsmith's) to magazines to illustrate stories about the celebrity, which is the typical use made of Goldsmith's photographs."); *Id.* at 1291 (Gorsuch, J., concurring) (stating that "if the Foundation had sought to display Mr. Warhol's image of Prince in a nonprofit museum or a for-profit book commenting on 20th-century art, the purpose and character of that use might well point to fair use.").

Similarly, in *Campbell*, the Court explained that the sampled music was for parodical purposes, but had it been for satire, that would have been insufficient.[97] Thus, one cannot stop short and just conclude the purpose was for "training." Rather, when determining the purpose, one must consider why the AI needs the work for training and whether that purpose justifies the use. Generating AI outputs that mimic or are otherwise based on the originals is not a purpose that justifies copying. As the Court made clear in *Warhol*, "[c]opying might [be] helpful to convey a new meaning or message. It often is. But that does not suffice under the first factor."

In the example above involving the ingestion of music, the why is "to create AI-generated music." Unlike a parody where the use is justified by the purpose of criticizing or commenting on that specific work, "creating AI-generated music" is not a purpose that justifies the unauthorized use of massive amounts of sound recordings that serve the same purpose as the generative AI output. The same applies to other models where the general purpose is to create other works of visual art or literary works that compete with the original work without providing a compelling justification for copying. The outcome is the same in all instances—the uses are not justified or legally transformative.

As noted above, it is not impossible that a particular AI developer could develop an AI model for the purpose of generating outputs for completely different purposes than the purposes an artist creates copyrighted works, and if so, that could result in the use being a transformative use. But even if a court were somehow to find that AI ingestion qualifies as a transformative use under the first fair use factor, that does not mean that AI ingestion is a fair use. In *Andy Warhol Foundation v. Goldsmith,* the Supreme Court made unequivocally clear that whether a use is transformative does not control a fair use analysis.[98] Rather, transformative purpose is merely one subfactor under the first fair use factor. Therefore, claims by some AI developers that the allegedly

[97] *See* Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 580–581 (1994) ("Parody needs to mimic an original to make its point, and so has some claim to use the creation of its victim's (or collective victims') imagination, whereas satire can stand on its own two feet and so requires justification for the very act of borrowing.").

[98] *Warhol*, 143 S. Ct. at 1264 ("Otherwise, 'transformative use' would swallow the copyright owner's exclusive right to prepare derivative works, as many derivative works . . . add new expression of some kind.").

transformative nature of generative AI weighs heavily in favor of fair use are clearly not supported by the law.[99]

Many of the generative AI platforms that have recently been launched are clearly commercial ventures, designed to attract as many users as possible and solidify a position in the market for the company. This commercial purpose would tend to weigh against fair use under the first factor, even if the ingestion of copyrighted works was found to have a transformative purpose. After *Warhol*, any factor-one analysis must involve a weighing of other considerations, such as the commercial nature of and justification for the use, both of which will factor prominently. We discuss *Warhol* and its factor-one implications, including the justification requirement, in more detail in response to question 8.1 below.

Factor Two: Nature of the Copyrighted Work

The second factor analyzes the nature of the work that was used without authorization. Works that are more creative or imaginative (such as a novel, movie, or song) are less likely to support a claim of a fair use than highly factual works or functional works like computer code.[100]

---

[99] *See e.g.*, OpenAI, LP, Comments on USPTO's Request for Comments on Intell. Prop. Prot. for A.I. Innovation 4, 8 (Oct. 30, 2019), https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf ("[P]roper application of fair use factors requires a finding of fair use, especially considering the highly transformative nature of training AI systems. . . . Prior cases have generally supported a finding of fair use for uses of large digital corpora that were less transformative than the training of AI systems. A fortiori, training AI systems should be considered fair use."); *Artificial Intelligence and Intellectual Property: Part I—Interoperability of AI and Copyright Law*: *Hearing Before the Subcomm. on Cts., Intell. Prop., & the Internet of the H. Comm. on the Judiciary*, 118th Cong. (2023) (written testimony of Chris Callison-Burch, Assoc. Professor of Comput. & Info. Sci., Univ. of Penn.), https://judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/callison-burch-testimony-sm.pdf ("In considering whether pre-training AI systems on copyright is fair use, it is important to highlight that the copying of works at this stage is 'non-expressive' in the same way that is for making a copy of a work in other digital media. Pre-training also has a transformative nature . . . I find there is a compelling argument that training AI systems on copyrighted works is fair use under US copyright law."); *Artificial Intelligence and Intellectual Property—Part II: Copyright, Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. 8 (2023) (written testimony of Ben Brooks, Head of Pub. Pol'y lPolicy, Stability AI), https://www.judiciary.senate.gov/imo/media/doc/2023-07-12_pm_-_testimony_-_brooks.pdf ("Training these models is an acceptable, transformative, and socially-beneficial use of existing content that is protected by the fair use doctrine and furthers the objectives of copyright law . . .").

[100] 4 *Nimmer on Copyright* § 13F.06 (2023) ("Under factor two, the more creative a work, the more protection it merits from copying; correlatively, the more informational or functional is plaintiff's work, the broader should be the scope of the fair use defense.").

Generative AI systems often ingest highly creative copyrighted works because the works provide immense value to AI developers. While this factor, like others, will depend on the specific facts of the AI model and what copyrighted works are being ingested, when generative AI tools ingest creative, copyright-protected works without authorization, this factor would weigh against fair use.

Another consideration under factor two is whether a work is published or unpublished, with the use of an unpublished work being less likely to qualify as a fair use.[101] This is particularly relevant to AI developers' scraping of massive amounts of works from the internet. Countless creators post their songs, written material, videos, photographs, and other works of visual art to the internet with no intention to sell or transfer ownership, and without "purposes of further distribution, public performance, or public display."[102] While the publication status of works posted to the internet will vary, it's unlikely that merely posting material online without any intention of further distribution would constitute a publication. As the Copyright Office Compendium explains, "the Office does not consider a work to be published if it is merely displayed or performed online, unless the author or copyright owner clearly authorized the reproduction or distribution of that work, or clearly offered to distribute the work to a group of intermediaries for purposes of further distribution, public performance, or public display."[103] Thus, a great deal of material that exists online and is inevitably scraped by AI developers may constitute unpublished works. On the one hand, the unauthorized use of these unpublished works, combined with the fact that the works are often creative (and not factual), would weigh the second factor even more against a finding of fair use. On the other hand, section 107 makes clear that "[t]he fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the [fair use] factors." So, while the publication status of a work certainly weighs heavily in the second factor analysis, that status, standing alone, is not dispositive of fair use.

---

[101] *Harper & Row v. Nation Enterprises*, 471 U.S. 539, 551 (1985) ("The unpublished nature of a work is '[a] key, though not necessarily determinative, factor' tending to negate a defense of fair use.").

[102] *Definitions*, U.S. COPYRIGHT OFFICE, https://www.copyright.gov/help/faq-definitions.html (last visited Oct. 17, 2023) (defining what "publication" means in copyright law).

[103] U.S. COPYRIGHT OFF., COMPENDIUM OF U.S. COPYRIGHT OFFICE PRACTICES § 1008.3(B) (3d ed. 2017), https://www.copyright.gov/comp3/chap1000/ch1000-websites.pdf.

Factor Three: Amount and Substantiality of the Portion Used in Relation to the Copyrighted Work as a Whole

The third factor looks at both the quantity and quality of the copyrighted material that was used. When complete and identical copies of copyrighted works are being scraped and fed into AI systems, this weighs against fair use. However, under this factor, courts do consider whether copying of entire works is necessary to achieve the particular use—so long as such use is justified (see discussion below). If that's the case, then the two considerations would be weighed together. As a result, this factor will likely either weigh against fair use or, at best, be neutral.

Fourth Factor: Effect of the Use Upon the Potential Market for or Value of the Copyrighted Work

Factor four, which considers whether, and to what extent, the unlicensed use harms the existing or future market for the copyright owner's original work, is often the most important factor to consider in a fair use analysis.[104] The existence of a licensing market weighs against a finding that copying without the permission of the copyright owner is excused by the fair use defense.[105] The fact that licenses are available for the use of copyrighted material for AI ingestion would tend to weigh against fair use. However, even if a copyright owner has not yet entered a particular market, that does not mean that this factor would weigh in favor of fair use. The Copyright Act is clear that this factor requires consideration of *potential* future markets as well as existing ones, meaning that a copyright owner need not currently be exploiting a certain market for there to be a harmful effect. Most copyright owners have recognized the value of generative AI licensing and have developed, or are in the process of developing, licensing

[104] *Harper & Row v. Nation Enterprises*, 471 U.S. 539, 566 (finding that factor four "is undoubtedly the single most important element of fair use."). While some say that the Supreme Court qualified this statement in *Campbell v. Acuff-Rose* by stressing the importance of the first factor, the Court was merely reacting to the circuit court's analysis of market substitution related to a parody, and not altering its view of the importance of the fourth fair use factor.

[105] *See* Am. Geophysical Union v. Texaco Inc., 60 F.3d 913, 929 (2d Cir. 1994) ("It is indisputable that, as a general matter, a copyright holder is entitled to demand a royalty for licensing others to use its copyrighted work, *see* 17 U.S.C. § 106 (copyright owner has exclusive right "to authorize" certain uses), and that the impact on potential licensing revenues is a proper subject for consideration in assessing the fourth [fair use] factor . . . .").

models. The marketplace should continue to properly value and incentivize creativity, and AI policy should not interfere with the right of copyright owners to license, or choose not to license, their works for AI purposes. Accordingly, these considerations should weigh heavily in the fourth fair use factor analysis.

Perhaps most significantly, the output of generative AI systems might act as a substitute in the market for the ingested copyrighted works, which would harm the market for the original works and weigh against fair use under the fourth factor. Consider the case of Greg Rutkowski, a digital artist who creates fantasy landscapes that have been licensed for use by numerous media companies. It's been reported that Rutkowski is one of the most commonly used AI image generator prompts, having been used over 93,000 times in Stable Diffusion alone and allowing anyone to generate competing outputs that incorporate Rutkowski's work.[106] Rutkowski and other visual artists are understandably concerned by the unauthorized use of their works to train AI models that in turn flood the market with works that directly compete with the ingested works. It's likely that companies outside the AI context that would have licensed these artists' works in the past will now simply use a substitute that is generated by an AI tool instead. It is one thing to use an AI-generated work as a substitute for an artist's work, but an entirely different thing when that substitute is only possible because it was generated from the artists' works themselves. The undeniable harm to the market for original works like Rutkowski's should be recognized as a grave concern to the future of human creativity and the incentives that our copyright system is based on.

Finally, when conducting a factor-four analysis, courts often consider the harm that would occur if the use were to "become widespread."[107] In the generative AI context, the widespread unauthorized ingestion of copyrighted works would certainly appear to cause immeasurable harm to creators and copyright owners—both by destroying existing, nascent, and to-be-developed licensing markets and by flooding the market with low-quality substitutional material.

---

[106] Melissa Heikkilä, *This Artist is Dominating AI-Generated Art. And He's Not Happy About It.*, MIT TECH. REV. (Sept. 16, 2022), https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it.

[107] *See, e.g.,* Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 451 (1984).

Significantly, it should also be recognized that courts can take into account the fact that some AI companies are scraping the internet indiscriminately to harvest material that is then ingested as training material, and that inevitably involves taking copyrighted works that are illegally offered on pirate websites or services. At the same time, it should be recognized that not all AI developers use datasets comprised of works that were indiscriminately scraped from the internet. As noted in other responses, some companies take a more responsible approach to collecting materials for ingestion purposes, and these developers would be far less likely to ingest pirated material. That said, some of the most popular AI models, including ChatGPT, were trained on datasets created by Common Crawl, a service that crawls the entire internet and archives material in publicly accessible repositories. This type of indiscriminate collection of works would almost certainly include pirated material.[108]

These problems have been highlighted by a series of recent lawsuits[109] brought by authors against various generative AI companies for the unauthorized ingestion of literary works to train their AI models.[110] The lawsuits allege that OpenAI and others have used datasets to train their AI models that contain hundreds of thousands of literary works, and that the only "Internet-based books corpora" that have ever offered that much material are notorious pirate eBook repositories like Library Genesis (aka LibGen), Z-Library (aka Bok), Sci-Hub, and Bibliotik.[111] According to the complaints, these illegal sites have long been of interest to the generative-AI-training community.

---

[108] Laura Herijgers, *ChatGPT Based on Illegal Sites, Private Data and Piracy*, TECHZINE (June 8, 2023), https://www.techzine.eu/blogs/privacy-compliance/107181/chatgpt-based-on-illegal-sites-private-data-and-piracy/.

[109] *See, e.g.*, Kadrey v. Meta Platforms, Inc., 2023cv03417 (N.D. Ca. July 7, 2023); Tremblay v. OpenAI, Inc., 2023cv03223 (N.D. Ca. June 28, 2023); Silverman v. OpenAI, Inc., 2023cv03416 (N.D. Ca. July 7, 2023); Authors Guild v. OpenAI, Inc., 23cv08292 (S.D.N.Y. Sept. 19, 2023).

[110] *See e.g.*, Nick Breen & Josh Love, *Attack of the Clones: AI Soundalike Tools Spin Complex Web of Legal Questions for Music*, BILLBOARD (May 19, 2023), https://www.billboard.com/pro/ai-music-tools-copy-artists-voices-legal-questions/.

[111] *See, e.g.*, Kadrey v. Meta Platforms, Inc., 2023cv03417 (N.D. Ca. July 7, 2023); Tremblay v. OpenAI, Inc., 2023cv03223 (N.D. Ca. June 28, 2023); Silverman v. OpenAI, Inc., 2023cv03416 (N.D. Ca. July 7, 2023); Authors Guild v. OpenAI, Inc., 23cv08292 (S.D.N.Y. Sept. 19, 2023).

If AI developers know or should have known they are ingesting works that have been made available illegally, a fair use defense would be much less likely to succeed. This concept—that to invoke fair use, an individual must possess an authorized copy of a work—was addressed by the Supreme Court in *Harper & Row Publishers Inc. v. Nation Enterprises*, which found that one consideration weighing against Nation availing itself of the fair use defense was the fact that it "knowing exploited a purloined manuscript."[112] Significantly, the Federal Circuit expanded on the concept in *Atari Games Corp. v. Nintendo of America, Inc.*, finding that because Atari gained access to an *unauthorized copy* of the Nintendo's source code by submitting false information to the U.S. Copyright Office, "any copying or derivative copying…does not qualify as a fair use."[113]

AI developers cannot claim in good faith that the indiscriminate scraping of massive amounts of material from the internet doesn't inevitably include stolen works. Applying a constructive "known or should have known" standard to AI developers that engage in mass scraping for ingestion purposes, they are aware or at the very least should be aware of the massive problem of online piracy, and that by indiscriminately scraping the entire internet they are also ingesting pirated works. Many AI developers are also some of the world's largest online service providers (OSPs). These OSPs are fully aware of the amount of online infringement occurring on their platforms, as evidenced by their own transparency reports and accounts of the notice and takedown system.[114] As a result, under existing precedent, any AI developer that scrapes pirate websites or uses material scraped from pirate websites for training may be precluded from successfully raising a fair use defense.

[112] Harper & Row, Publrs. v. Nation Enters., 471 U.S. 539, 562 (1985).

[113] Atari Games Corp. v. Nintendo of Am., Inc., 975 F.2d 832, 843 (Fed. Cir. 1992).

[114] *See e.g.*, *Transparency Report*, GOOGLE, https://transparencyreport.google.com/copyright/overview?hl=en (last visited Oct. 17, 2023); JENNIFER M. URBAN, JOE KARAGANIS & BRIANNA SCHOFIELD, NOTICE AND TAKEDOWN IN EVERYDAY PRACTICE, U.C. BERKELEY PUB. L. RSCH. PAPER NO. 2755628 (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2755628.

Case Law Does Not Support Claims that AI Ingestion Qualifies as Fair Use

Some have argued that certain cases support the position that AI ingestion of copyrighted works (for training purposes) categorically qualifies as fair use—particularly *Sega v. Accolade,*[115] *Sony v. Connectix,*[116] and *Google v. Authors Guild* (the Google Books case).[117] However, none of these cases support the categorical positions espoused by some AI developers and their supporters. As noted above, fair use is a very fact specific analysis, and thus, while prior court decisions are instructive, there are often different facts that render them distinguishable. For example, the fair use analysis for generative AI is different than for search-functionality uses that convey information about a work or merely "point" to where a work can be found (as was the case in the Google Books case), rather than generate a new work that might be substantially similar or a substitute for the original.[118]

Below we discuss some cases that are often listed as supporting a fair use defense, and we explain why they do not.

- *Sega v. Accolade*: In this Ninth Circuit case, Accolade, one of Sega's competitors, developed its own computer games to be played on the Sega consoles. To make its game software compatible with Sega's game consoles, Accolade copied small portions of object code from Sega's games, converted it to source code—a form of reverse engineering—and used what it learned to write its own computer code to make its games compatible with Sega consoles. The court held that Accolade's reverse engineering of the computer program for compatibility purposes (combined with a lack of other means to access the

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

[116] Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596 (9th Cir. 2000).

[117] Authors Guild v. Google, 804 F.3d 202 (2d Cir. 2015).

[118] Perfect 10, Inc. v. Amazon, 508 F.3d 1146 (9th Cir. 2007) (finding that defendant's creation of thumbnail images as part of a search function that pointed to where a consumer could find the full images was transformative in that it provided information about where the original works could be found).

elements not protected by copyright) constituted fair use.[119] In its decision, the court was clear that its analysis was specific to the functional computer code at issue. In contrast, generative AI systems are making unauthorized use of clearly expressive, non-functional works of authorship. In fact, the Ninth Circuit was clear that its analysis would be different if the works at issue were more expressive (and less functional). The court explains that because "Sega's video game programs contain unprotected aspects that cannot be examined without copying, we afford them a lower degree of protection than more traditional literary works."[120]

- *Sony v. Connectix*: In another Ninth Circuit reverse engineering case, Sony sued Connectix for copying the software program that operated its PlayStation video game console for the purpose of emulating the console on a regular computer. The court concluded that Connectix's intermediate copying for reverse engineering qualified as fair use, necessary to permit Connectix to make its non-infringing game station function with PlayStation games. The *Sony* case is similar to *Sega,* in that it involved the copying of software code for reverse engineering purposes to develop non-infringing competitive products. Following *Sega's* precedent, the court found that Sony's copyrighted software code included functional elements that resulted in a lower degree of protection. This factor two analysis was central to the ultimate fair use finding, and it would be inapplicable to the highly expressive works ingested by generative AI systems. It should also be noted that both *Sega* and *Sony* are based on the understanding that interoperability exceptions to copyright law are justifiable when they support legitimate forms of competition. This focus on copying functional elements of a work for interoperability purposes (as a means to developing legitimate competitive products) is inapplicable to generative AI's use of non-functional, highly expressive works.

- *Authors Guild v. Google*: Google Books is the case most AI developers rely on when they claim AI ingestion of copyrighted materials qualifies as a fair use, but it is highly

[119] Sega Enters. Ltd., 977 F.2d at 1514.

[120] *Id.* at 1526.

distinguishable as the use of the copyrighted materials involved a completely different purpose. This Second Circuit case involved Google's mass digitization for its Google Books project, which made the digital copies available for library collections and for the public to search electronically using a search engine. The Authors Guild and individual copyright owners complained that Google scanned more than twenty-million books without permission or payment of license fees. The Second Circuit agreed with the district court's ruling that Google's digitization and subsequent use of the copyrighted works was fair use. Concluding that Google's use was transformative, the circuit court found that "Google's making of a digital copy to provide a search function . . . augments public knowledge by making available information about [p]laintiffs' books without providing the public with a substantial substitute for matter protected by the [p]laintiffs' copyright interests in the original works or derivatives of them."[121] The decision also made clear that the case "tests the boundaries of fair use" and may have come out differently if the purpose of Google's scanning of literary works was to create substitutes for the original works.[122] In other words, it was critical that Google used the books to provide information *about* the works and by serving as a "pointer" for readers to consume the original works. In contrast, AI copies particular types of copyrighted works to manufacture the same type of work and thus is much more likely to usurp the market for the underlying work.

Unlike the activity at issue in Google Books, the purpose of generative AI currently has nothing to do with providing factual information about the copyrighted works to users. Instead, generative AI systems typically reproduce and use the expressive elements from ingested copyrighted works as part of a process that results in the manufacture of AI-generated works that compete in the same market as the original copyrighted works.

Importantly, the court noted that Google also implemented significant safeguards to secure the copies of books it used in its database, such as only showing "snippets" of

[121] Authors Guild, 804 F.3d at 207.

[122] *Id*. at 206, 225.

works to highlight a search term and implementing anti-hacking measures.[123] Due to these safeguards, the court concluded that there was little risk that Google's actions could serve as a substitute for the copied works. In the generative AI context, safeguards are often not implemented,[124] works scraped from the internet are taken from pirate websites or services, and the harm to copyright owners from substitution could be catastrophic if the massive amounts of copyrighted materials that AI developers take without permission were compromised.

One other factor that the court cited (with regard to the fourth factor) was that there was no actual or potential market for the licensing of copyrighted works to search engines.[125] This is significantly different than for AI, where (as discussed in detail in our responses to question six and its subparts) there very clearly is a burgeoning market for the licensing of copyrighted works for ingestion.

In sum, while these cases might be instructive in different contexts, they deal in distinct fact patterns that are clearly distinguishable from AI ingestion. By no means do any of these cases stand for the proposition that AI ingestion is categorically fair use.

It's also important to note that many of the arguments made in defense of unauthorized use of copyrighted works for ingestion purposes focus entirely on output-related infringement. Specifically, OpenAI's motions to dismiss in multiple lawsuits filed against it by groups of creators claim that the plaintiffs did not sufficiently show direct infringement based on substantial similarity between outputs and ingested works, while not addressing direct

[123] *Id*. at 228.

[124] For example, by allowing for prompts that are "in the style of" an author or artist or by allowing prompts including copyrighted characters.

[125] Authors Guild, 804 F.3d at 226–27.

infringement of plaintiff's right of reproduction at the ingestion stage or raising a fair use defense for such unauthorized use.[126]

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?

Warhol v. Goldsmith

On May 18, 2023, the Supreme Court's decision in *Andy Warhol Foundation v. Goldsmith* made unequivocally clear that whether a use is transformative does not by itself control a fair use analysis, and that it is inappropriate for courts to attribute so much weight to what is simply a subfactor of the first fair use factor. The decision reaffirmed a critical tenet of the fair use doctrine—that transformative purpose is merely one subfactor that not only is not dispositive of a fair use analysis, but doesn't even control a factor-one determination.[127] Therefore, even if the use of copyrighted works for ingestion by AI developers were considered to be a transformative use in a particular case, which we do not believe that it is, that would not mean that the use categorically qualifies as fair use.

Central to the Supreme Court's decision in *Warhol* was its confirmation that factor one requires "justification" for copying to qualify as a fair use. This standard, which was first explained by the Supreme Court's discussion of parodic uses in *Campbell v. Acuff-Rose*, was unmistakably reaffirmed in *Warhol*, which confirmed that when an original work and secondary use share the

[126] *See e.g.*, Defendant's Motion to Dismiss at 3, Tremblay v. OpenAI, Inc., 23-cv-03223 (N.D.Ca. 2023) ("Plaintiff's claims for vicarious infringement are based on the erroneous legal conclusion that every single ChatGPT output is necessarily an infringing 'derivative work' . . . regardless of whether there are any similarities between the output and the training works.").

[127] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 143 S. Ct. 1258, 1276 (2023) ("First, the fact that a use is commercial as opposed to nonprofit is an additional 'element of the first factor.' The commercial nature of the use is not dispositive."); Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 584 (1994) ("The language of the statute makes clear that the commercial or nonprofit educational purpose of a work is only one element of the first factor enquiry into its purpose and character . . . the commercial or nonprofit educational character of a work is 'not conclusive,' but rather a fact to be 'weighed along with other[s] in fair use decisions.'").

same or highly similar purposes, and the secondary use is commercial, "a particularly compelling justification is needed." Applying this standard to commercial generative AI developers that use copyrighted works for the same or highly similar purpose as the ingested works, there is insufficient justification to support their claims that they must scrape the entire internet (including pirate websites and copyrighted works behind firewalls) or ignore existing licenses offered by copyright owners.

Some have argued that licensing copyright protected material for AI ingestion is not practical because, in order to be successful, AI systems *must* use *every* piece of available content.[128] It might be desirable to train an AI tool on everything and anything that can be scraped from the internet, but that is not always the case (sometimes less is more) and—as evidenced by the successful AI developers that do not scrape the internet for ingestion purposes—that does not make it necessary and is not a "compelling justification" that would weigh in favor of fair use. Simply because AI developers *want* to use everything to train their systems doesn't mean that it is necessary.

Evidence suggests that generative AI and licensing can coexist successfully. In fact, Adobe's Firefly suite of generative AI tools—which are trained on proprietary stock images, licensed images, and public domain images whose copyrights have expired—has seen broad consumer adoption.[129] Additionally, leading AI developer Nvidia has partnered with Getty Images to build

---

[128] *See e.g.*, *Artificial Intelligence and Intellectual Property: Part I—Interoperability of AI and Copyright Law, Hearing Before the Subcomm. on Cts., Intell. Prop., & the Internet of the H. Comm. on the Judiciary*, 118th Cong. 3 (2023) (written testimony of Sy Damle, Partner, Latham & Watkins LLP), https://judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/damle-testimony.pdf ("Successfully training an AI model requires using many *billions* of pieces of content, so the scope of any statutory or collective licensing scheme would be many orders of magnitude larger than any similar scheme in the history of American law."); *id.* at 23 (written testimony of Chris Callison-Burch, Assoc. Professor of Comput. & Info. Sci., University of Pennsylvania), https://judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/callison-burch-testimony-sm.pdf ("If it were to be ruled that . . . that every work in the training data set needed an explicit license from the copyright holder, then progress on developing capable AI systems would be jeopardized."); Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 748 (2021), https://texaslawreview.org/fair-learning/ ("[T]raining sets are likely to contain millions of different works with thousands of different owners, there is no plausible option simply to license all of the underlying photographs, videos, audio files, or texts for the new use.").

[129] Rashi Shrivastava, *Adobe Brings Its Generative AI Tool Firefly to Businesses*, FORBES (June 8, 2023), https://www.forbes.com/sites/rashishrivastava/2023/06/08/adobe-brings-its-generative-ai-tool-firefly-to-businesses/?sh=53f4dd94582b.

new generative AI technologies that ingest only fully licensed content, and IBM recently announced a collaboration with Adobe to assist customers in implementing generative AI models based on Adobe's Firefly technology.[130] Ultimately, simply because licensing is not a financially desirable avenue for AI developers does not mean unauthorized use is justified in way that would favor fair use.

The *Warhol* decision also makes clear that the analysis of transformativeness under the first fair use factor varies depending on the particular use; use of a work in one instance may qualify as a fair use, but in another instance a different use of the same work may not. For example, Justice Gorsuch's concurring opinion explains that if the Andy Warhol Foundation "had sought to display Mr. Warhol's image of Prince in a nonprofit museum or a for-profit book commenting on 20th-century art, the purpose and character of that use might well point to fair use."[131] Thus, under *Warhol*, different uses of a particular work should be considered separately, and it is possible that one use is considered to be transformative while the other is not. In the generative AI context, it is critical to recognize that there may be different stages of AI development (carried out by different entities), some that may claim to be noncommercial and others that are clearly commercial, and that any analysis of transformativeness under factor one of the fair use defense must be tied to the particular use and considered independently.

Another important part of the *Warhol* decision that will have implications for generative AI fair use analyses is its clarification that subjective opinions as to the transformativeness of a use are largely irrelevant under factor one. Responding to Warhol's claims of transformative purpose based on new meaning or message, the Court made clear that "[a] court should not attempt to evaluate the artistic significance of a particular work….Nor does the subjective intent of the user (or the subjective interpretation of a court) determine the purpose of the use."[132] Further,

---

[130] Rick Merritt, *Moving Pictures: NVIDIA, Getty Images Collaborate on Generative AI*, NVIDIA (Mar. 21, 2023), https://blogs.nvidia.com/blog/2023/03/21/generative-ai-getty-images/; *IBM Expands Partnership with Adobe to Deliver Content Supply Chain Solution Using Generative AI*, IMB NEWSROOM (June 19, 2023), https://newsroom.ibm.com/2023-06-19-IBM-Expands-Partnership-with-Adobe-To-Deliver-Content-Supply-Chain-Solution-Using-Generative-AI.

[131] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 143 S. Ct. 1258, 1291 (2023).

[132] *Id*. at 1284.

discussing the intersection of transformativeness and commerciality, the Court went on to explain that when perception of new meaning or message is a matter of subjective opinion, the commercial nature of the use "looms larger." Thus, when considering clearly commercial generative AI endeavors, the subjective opinion of an AI developer or any particular commentator or technologist should have no effect on a factor one analysis.

In the wake of *Warhol*, when the four factors are considered in relation to the unauthorized ingestion of copyrighted works for the purpose of generating substitutional output, in many cases such use will fall outside of the bounds of fair use.

Google v. Oracle

Many have recognized the limited application of Google v. Oracle.[133] This is especially true when attempting to apply the holding in the case to generative AI. *Google v. Oracle* has very limited applicability because the decision was expressly limited to the specific type of computer

[133] *See e.g.*, Brief for the United States as Amicus Curiae Supporting Defendants, Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 143 S. Ct. 1258 (2023) (No. 21-869), at 25 ("[Andy Warhol Foundation's] reliance on Google is likewise misplaced. Emphasizing the difficulty of "apply[ing] traditional copyright concepts in th[e] technological world" of "functional" computer programs, the Google Court principally focused on the second statutory factor, emphasizing the copied code's distance 'from the core of copyright.')"; Jonathan Bailey, *How the Warhol Ruling Could Change Fair Use*, PLAGIARISM TODAY (May 18, 2023), https://www.plagiarismtoday.com/2023/05/18/how-the-warhol-ruling-could-change-fair-use/ ("However, given how narrow [the *Google v. Oracle* case] was, applying solely to software code, the Warhol case is the first broad SCOTUS decision on fair use in nearly 30 years."); PRYOR CASHMAN, *Circuit to Warhol Estate: Google v. Oracle Does Not Dictate A Different Result* (Aug. 26, 2021), https://www.pryorcashman.com/publications/circuit-to-warhol-estate-google-v-oracle-does-not-dictate-a-different-result ("The Circuit explained how the Supreme Court "took pains to emphasize" that the unique context of the [*Google v. Oracle*]case (software code) may make its holding less applicable in other contexts, especially in the context where the copyrighted material "serves an artistic rather than a utilitarian function."); PROSKAUER ROSE LLP, *Supreme Court Affirms Andy Warhol's Prince Series Not Transformative Fair Use* (June 14, 2023), https://www.proskauer.com/blog/supreme-court-affirms-andy-warhols-prince-series-not-transformative-fair-use ("Recently, the court in Google v. Oracle interpreted fair use broadly but limited its decision to the context of software codes."); Tyler Ochoa, *U.S. Supreme Court Vindicates Photographer But Destabilizes Fair Use — Andy Warhol Foundation v. Goldsmith (Guest Blog Post)*, TECHNOLOGY & MARKETING LAW BLOG (June 20, 2023), https://blog.ericgoldman.org/archives/2023/06/u-s-supreme-court-vindicates-photographer-but-destabilizes-fair-use-andy-warhol-foundation-v-goldsmith-guest-blog-post.htm ("Despite those caveats, the opinion is likely to be enormously consequential, far more so than the Court's similarly narrow and context-specific ruling two years ago in the Google v. Oracle software case."); Tyler Ochoa, *U.S. Supreme Court Upholds Fair Use in Google-Oracle Software Battle (Guest Blog Post)* (Apr. 8, 2021), https://blog.ericgoldman.org/archives/2021/04/u-s-supreme-court-upholds-fair-use-in-google-oracle-software-battle-guest-blog-post.htm ("Any Supreme Court opinion is important, and this one no doubt will be quoted often in future briefs and opinions. But other than clarifying the standard of review, I doubt the decision will have much impact on fair use cases that do not involve software.").

code at issue.[134] Considering Google's use of the unique form of software declaring code, the Supreme Court found that the context specific nature of a computer programs' functional purpose affects a fair use analysis. Specifically, the *Google* decision did not conform to the justification requirement of criticism, commentary, or information about a copied work discussed in *Warhol* and *Campbell*, but that is only because "[t]he fact that computer programs are primarily functional makes it difficult to apply traditional copyright concepts in that technological world."[135] The Court was careful to distinguish computer code from highly expressive works that have no functional elements that may impact a factor-two analysis, saying that "computer programs differ from books, films, and many other 'literary works' in that [software] programs almost always serve functional purposes."[136] When considering the clearly distinguishable circumstances surrounding generative AI systems' use of creative, expressive works of authorship for ingestion purposes, there is little doubt that the *Google v. Oracle* decision is inapplicable.[137]

The Court's clear distinction between computer code and the exact type of highly expressive works that are being ingested by AI systems confirms that any subsequent discussion of the value of developing new products is cabined to the use of functional code to develop software.

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?

Any analysis of whether a party may be directly liable (at any stage of AI development) should first question whether that party is involved in acts of making and using copies, distributing those copies, creating derivative works of those copies, and/or making those copies or derivative works available. The focus should not be solely on the "training" stage of AI systems or the exact point of ingestion.

---

[134] *See* Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1208 (2021) (the Court was clear that its decision "do[es] not overturn or modify [its] earlier cases" involving fair use).

[135] *Id*. at 1186, 1208.

[136] *Id*. at 1198.

[137] The fact that the underlying AI tools are themselves software is irrelevant when analyzing the use of expressive copyrighted works to generate new works.

Any entity that collects and/or curates copyrighted material for purposes of ingestion into generative AI systems reproduces copyrighted works and is directly liable for those acts–unless they have a valid defense or a license. This is true regardless of whether they are the entity engaged in the act of ingestion or training. Because fair use is an affirmative defense, before a fair use analysis is conducted, a determination of infringement must be made. That means first determining whether an unauthorized reproduction or derivative work has been made, whether there is distribution, and/or whether the copied works have been made available. Once that determination has been made and the entity engaged in the direct infringement has been identified, that party can raise a fair use defense which will be analyzed based on the particular facts. But again, simply because an entity is "not themselves engaged in the training" of an AI system has no impact on an infringement or fair use determination.

Additionally, a party that is not directly liable for engaging in copying or distribution may be liable under the various doctrines of secondary liability if they meet the requirements of any of those doctrines. The doctrines of secondary liability apply to parties involved in the development of generative AI just as they do to other types of uses and users. For example, an entity involved in the development of generative AI may not directly infringe the rights of a copyright owner but may nonetheless induce, cause, or materially contribute to infringement. We therefore urge caution when considering any rules or policies that encourage disaggregation of collecting, curating, and training to avoid liability, as it will promote gamesmanship and data laundering.

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?

Noncommercial or Research Purposes

As discussed in our responses above, fair use is applied on a case-by-case basis and any determination will depend on the specific circumstances and facts. The enumerated uses in section 107 of the Copyright Act—"criticism, comment, news reporting, teaching [], scholarship,

or research"—are instructive but not dispositive of a finding of fair use. Thus, simply because a use is purportedly for research purposes is not itself determinative of a finding of fair use. Consequently, categorical assertions that use of copyrighted works for generative AI research purposes are for research purposes and thus, always qualifies as fair use are legally inaccurate. Regardless of whether a use is for research purposes or not, the statute and case law are clear that a complete evaluation of the four factors is required.

The same is true for any use where the purpose is found to be noncommercial or not-for-profit. Factor one considers the purpose and character of the use, which includes considering whether the use is of a commercial nature or is for nonprofit educational purposes.[138] Thus, whether a use is commercial or noncommercial is a subfactor of factor one, the determination of which will not control a fair use analysis.[139] Importantly, even noncommercial or nonprofit uses can result in market harm or the accrual of benefits to an alleged infringer and weigh against fair use under the first factor.[140]

[138] The term "nonprofit educational purpose" from section 107 of the Copyright Act is considerably narrower than the broad term "noncommercial," which is often used as shorthand.

[139] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 143 S. Ct. 1258, 1276 (2023) ("First, the fact that a use is commercial as opposed to nonprofit is an additional 'element of the first factor.' The commercial nature of the use is not dispositive."); Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 584 (1994) ("The language of the statute makes clear that the commercial or nonprofit educational purpose of a work is only one element of the first factor enquiry into its purpose and character . . . the commercial or nonprofit educational character of a work is 'not conclusive,' but rather a fact to be 'weighed along with other[s] in fair use decisions.'"). *See also* Weissmann v. Freeman, 868 F.2d 1313, 1324 (2d Cir. 1989) ("Monetary gain is not the sole criterion . . . The absence of a dollars and cents profit does not inevitably lead to a finding of fair use.").

[140] *See, e.g.*, Worldwide Church of God v. Phila. Church of God, Inc., 227 F.3d 1110, 1118 (9th Cir. 2000) ("[H]aving in mind that . . . religion is generally regarded as 'not dollar dominated,' MOA 's use unquestionably profits PCG . . . by attracting through distribution of MOA new members who tithe ten percent of their income to PCG, and by enabling the ministry's growth . . . [PCG] gained an 'advantage' or 'benefit' from its distribution and use of MOA without having to account to the copyright holder."); *Weissmann*, 868 F.2d at 1324 ("The absence of a dollars and cents profit does not inevitably lead to a finding of fair use . . . the profit/non-profit distinction is context specific, not dollar dominated."); Soc'y of Holy Transfiguration Monastery, Inc. v. Gregory, 689 F.3d 29, 61 (1st Cir. 2012) ("But removing money from the equation does not, under copyright law, remove liability for transgressing another's works . . . 'Profit,' in this context, is thus not limited simply to dollars and coins; instead, it encompasses other non-monetary calculable benefits or advantages.").

When AI Models or Datasets are Later Adapted for Use of a Commercial Nature

As we discuss in previous responses, a scenario in which there is a distinct commercial use of a corpus of copyrighted works that was initially developed purportedly for noncommercial research purposes is data laundering, and it would influence a fair use analysis differently and weigh against fair use when considering the purpose of the use. The Supreme Court in *Warhol* made clear that it is the specific use of a work or works that matters, and *if a use is noncommercial in one instance and commercial in another the fair use analysis applies differently*. It is critical that any fair use analysis take into account what the purpose and character of the use is at the time of the alleged infringement, not what the nature or purpose of the use might have been at some earlier time.

Funding

Again, any fair use analysis must focus on the specific use at issue, and not whether the entity funding the project is a not-for-profit entity. That said, while funding from a noncommercial source is not dispositive of fair use under the first factor, funding from a commercial source may be evidence of a commercial purpose that would tip the first factor against fair use. This is especially true if funding for the creation of a corpus of copyrighted works for research-related, noncommercial generative AI purposes comes from a for-profit entity that later makes use of the dataset for commercial purposes or would gain an advantage or benefit from the use (i.e., data laundering). There must be a high level of scrutiny when for-profit entities fund seemingly noncommercial generative AI-related research projects to ensure that there is no subsequent data laundering or benefit (commercial or otherwise) to the funding entity. Similarly, there should be a high level of scrutiny when nonprofit entities create datasets and then later enter into deals with for-profit AI companies, regardless of how the nonprofit entities are funded.

A recent example of a court finding that an entity accrued benefits from an alleged noncommercial use that it funded (which weighed against fair use under the first factor) occurred in the book publishers' lawsuit against the Internet Archive (IA) for unauthorized reproduction

and distribution of copyrighted literary works.[141] In that case, the U.S. District Court for the Southern District of New York rejected the Internet Archive's fair use defense that its online "library" was "wholly noncommercial" because the Internet Archive is a non-profit organization and did not charge readers to access the literary works it made available.[142] The decision explains that "IA uses its Website to attract new members, solicit donations, and bolster its standing in the library community," all of which were benefits the court found directly resulted from its unauthorized use of the publishers' copyrighted works.[143] Citing to the Supreme Court's *Harper & Row* decision, the court makes clear that monetary gain is not the sole question under a factor one commercial/noncommercial analysis and that any benefit flowing towards the alleged infringer can weigh against fair use under the first factor. Applying these standards to generative AI companies, it is essential to scrutinize their funding of and relationship to any purported noncommercial or nonprofit entities involved in any stage of generative AI development. It may be prudent to presume an ultimate commercial purpose when a for-profit AI developer funds a non-profit entity's research or when a non-profit entity develops datasets and then commercializes them at a later time.

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?

The quantity of training material developers of generative AI ingest into their systems varies depending on the type of AI model and the type of output. Many popular generative models ingest massive amounts of copyrighted works (and other material scraped from the internet). OpenAI's ChatGPT tool trains on untold numbers of copyright works through the ingestion of massive corpuses of literary works compiled in datasets like Common Crawl and Books1 and Books 2. Similarly, Stable Diffusion's image generator tool was trained on billions of images scraped from the internet and compiled in LAION 5B, which was made from a subset of the

[141] Hachette Book Grp., Inc. v. Internet Archive, No. 20-CV-4160 (JGK), 2023 WL 2623787, at *9 (S.D.N.Y. Mar. 24, 2023) ("IA receives these benefits as a direct result of offering the Publishers' books in ebook form without obtaining a license. Although it does not make a monetary profit, IA still gains 'an advantage or benefit from its distribution and use of' the Works in Suit 'without having to account to the copyright holder[s],' the Publishers.").

[142] *Id*.

[143] *Id*.

Common Crawl dataset. There are also AI tools that only ingest proprietary works, works that are licensed, or works that are in the public domain. These models, like Adobe's Firefly, likely train on a far smaller quantity than the models that ingest works scraped from the internet indiscriminately without authorization.

Ultimately, the quantity or volume of works being used to train the AI model does not alter a fair use analysis. To argue otherwise would turn copyright law on its head. It can't be the case that massive infringement is permitted, while smaller scale infringement is actionable. Using large volumes of copyrighted works would affect a much larger market, making a fair use defense less likely to succeed. As discussed in our response to question 8.1 above, claims by some that unlimited amounts of works are needed to develop the most optimally functional AI models is not a justification that weighs in favor of fair use. The successful function and popularity of tools that are trained solely on licensed or public domain works refutes any notion that an AI developer must use massive amounts of unlicensed copyrighted works in order for it to function. As noted in other responses, arguments that fair use should categorically excuse unauthorized use of copyrighted materials for training purposes because obtaining licenses is difficult or would hinder the development of the most desirable (i.e., profitable) AI model must be rejected.

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?

When there is an existing or potential licensing market for the same training purpose, the harm to that licensing market and the value of copyrighted works included in licenses is undeniable. It is critical to recognize that the fourth factor does not require a copyright owner to have already entered a market (or offered a license). Section 107 makes explicitly clear that the fourth factor considers whether there is an effect on a *potential* market that a copyright owner may enter in the future. Additionally, a fourth factor analysis considers the market harm that would occur *were the use to become widespread and unrestricted*. What that means is that even if an unauthorized use is not currently causing cognizable harm to a market for or value of a copyrighted works, courts can consider whether such harms would be likely if the use was to proliferate. Further, as we

explain in response to question eight above, an additional element that courts should consider under factor four is the compounded harm to copyright owners that occurs when AI developers scrape works from illegal pirate websites or services.

As we discuss in our response to question eight above on the fourth factor, the output of generative AI systems competes in the same market as the ingested copyrighted works, which would harm the market for the original works and weigh against fair use under the fourth factor. The relevant inquiry under a factor-four analysis should be whether generative AI outputs act as a substitute or supplant the market for a particular work, the body of works of the same author, or the market for that general class of works. Although cases in the past have largely focused on harm to a particular work—because that was the nature of the harm—with generative AI, the harm is often to a creator's overall body of work or even the market more broadly. These harms all impact the creator's incentives, and they should be considered under a factor-four analysis.

In addition, there is also a market for copyright-owner curated datasets that have high quality content, that are well organized, and that include high quality metadata, and harms to the market for such datasets should be considered. Similarly, the relevant inquiry when considering a fair use defense related to input-stage infringement is whether the unauthorized ingestion of copyrighted works would supplant the existing or potential licensing market for the works.

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)?

Under the Copyright Act, a copyright owner must affirmatively consent to the use of their work for training unless a defense applies. Allowing an AI system to use the work unless the copyright owner objects (i.e., opts out), would require enactment of legislation. As noted elsewhere in these responses, there is a burgeoning licensing market for AI training, which demonstrates that no exception is necessary or deserved. Thus, other than possibly the situation described in our response to question 9.2, the Copyright Alliance would vehemently oppose such legislation.

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

Consent of the copyright owner should be required for all uses of copyrighted works to train AI models unless a defense (like fair use) applies regardless of whether the use is commercial or non-commercial—it already is. That is the law and should remain the law. To the extent commercial or non-commercial use is relevant to consent, the fair use defense already adequately takes that into account.

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 in our response to questions 9 and 9.1, we adamantly oppose an opt-out approach, with the one possible exception described below.

As noted elsewhere in these comments, we recognize that, so far, the licensing market for AI training has largely eluded individual creators. It is our hope that leading AI companies will soon begin to also license the works of individual creators for training purposes, just as they have begun doing so for businesses with large corpuses of high-value copyrighted works. In fact, recently, we have seen evidence that that might be the case.[144] However, if that does not transpire and, as noted in more detail in our response to question five, if (i) there exists a general consensus of organizations and individual creators within a particular industry (for example, the book publishing industry) who are willing to accept "opt outs" solely in the context of enacting an extended collective license (ECL)[145] provision; (ii) such provision is narrowly targeted to a

[144] For example, visual communication company, Canva, launched a suite of generative AI tools called Magic Studio that allow its users to generate videos, presentations, and other designs from text prompts. In its announcement, Canva announced that it is dedicating $200 million over the next three years in creator and AI royalties. Charlotte Trueman, *Canva Bolsters AI Offerings, Providing Copyright Indemnity for AI-Generated Images*, COMPUTERWORLD (Oct. 5, 2023), https://www.computerworld.com/article/3708249/canva-bolsters-ai-offerings-providing-copyright-indemnity-for-ai-generated-images.html.

[145] When we refer to extended collective licensing in these comments, we adopt the Copyright Office's definition from its *Legal Issues in Mass Digitization: A Preliminary Analysis and Discussion Document*, which defines

particular industry and a particular type of work(s); and (iii) such license would not directly or indirectly affect (through inadvertent consequences or otherwise) those industries and works not intended to be covered by the license, the Copyright Alliance would not oppose such an approach. Any such licenses must include robust notice practices to ensure all creators are notified and given an opportunity to opt out, and the process for opting out must be very simple.

There continues to be a debate whether an AI model that has been trained on a work can be retrained to "unlearn" that work. To the best of our knowledge, technologies do not yet exist that can effectively remove entire works at scale from an AI model after it has been trained – though they might be coming.[146] Some indicate that untraining models is challenging.[147] Others indicate that it can be done, but it could be expensive.[148] In the event an AI model cannot practically be retrained or a particular ingested work cannot practically be "forgotten," that serves as further evidence of why an opt-out system would not work since the harm caused to the copyright owner cannot be undone once the work has been ingested (and many of the biggest models in current use have already been built). In the event an AI model can be retrained, the fact that it may be expensive to do so further establishes that licensing, rather than an opt out, is the better and most cost-effective option.

With regard to technical tools that might facilitate an opt-out process, such as a technical flag or metadata indicating that a work should not be ingested for AI training, even though we oppose an

---

extended collective licensing as an approach where "the government passes legislation authorizing a collective organization to license all works within a category, such as literary works, for particular, limited uses, regardless of whether copyright owners belong to the organization or not. The collective then negotiates agreements with user groups, and the terms of those agreements are binding upon all copyright owners by operation of law." U.S. COPYRIGHT OFF., LEGAL ISSUES IN MASS DIGITIZATION: PRELIMINARY ANALYSIS AND DISCUSSION DOCUMENT 30–31 (2011), https://www.copyright.gov/docs/massdigitization/USCOMassDigitization_October2011.pdf.

[146] At least one group of researchers has done so with one work but are not sure they could scale it. *See* Ronen Eldan & Mark Russinovich, *Who's Harry Potter? Approximate Unlearning in LLMs*, CORNELL UNIV. ARXIV (Oct. 4, 2023), https://arxiv.org/pdf/2310.02238.pdf (acknowledging that large language models (LLMs) "are trained on massive internet corpora that often contains copyright infringing content" and they propose a "novel technique for unlearning a subset of the training data from an LLM, without having to retrain it from scratch").

[147] *See* Jie Xu et al., *Machine Unlearning: Solutions and Challenges*, CORNELL UNIV. ARXIV (Aug. 14, 2023), https://arxiv.org/pdf/2308.07061.pdf ("Machine unlearning faces challenges from inherent properties of ML models as well as practical implementation issues.").

[148] *See id*.

opt out approach, that does not mean that opt out technologies do not have a role to play and should not be developed and used. There are various promising technical tools, such as C2PA and Project Origin, and, in the context of visual works, IPTC (which is compatible with C2PA) is becoming the standard both for opting out of ingestion as well as labeling generative AI output as synthetic.[149] Many standard-making bodies are also considering this issue, such as the W3C's TDM opt-out protocol (developed in response to the EU's DSM Directive).

There are also some tools that are not, at least yet, adequately effective for AI ingestion opt out purposes. Robots.txt protocol is one example. While robots.txt does alert scraping tools not to ingest the associated copyrighted work, it has significant limitations because it is only effective to the extent it is recognized and respected, and it was not designed to be targeted to scraping for generative AI ingestion. Robots.txt would also prevent a search engine from scraping and categorizing the work. A copyright owner may want their work to be scraped for search engine purposes—so they can be found on the internet—but not for AI ingestion. Even if robots.txt is used it will not work effectively by itself because it operates at the URL or website level. Copyrighted works will be available on pirate sites outside of the content owner's control, and if those sites don't opt-out, then broad-scale web scraping means the content will end up in the training set anyway.

When opt outs are used, they must be respected by AI companies—regardless of whether they are in text (such as in a website's terms of use) or accomplished through the use of a technical tool, like metadata, a flag, or one of the tools noted above. Any AI company that is ingesting copyrighted works must first determine whether the work contains an opt-out notification. If an AI developer ingests a copyrighted work that is protected by opt-out measures, the act of ingestion should be considered to be willful infringement and potentially result in heightened damage awards under the Copyright Act.

[149] For example, Copyright Alliance member, the PLUS Coalition, has partnered with the IPTC to publish a draft on proposed revisions to the PLUS License Data Format standard that proposes a standard for expressing image data mining permissions, constraints, and prohibitions. This includes declaring in image files whether an image can be used as part of a training data set used to train a generative AI model. *See PLUS Publishes Draft Standard for Image Data Mining Restrictions*, IPTC (June 28, 2023), https://iptc.org/news/plus-publishes-draft-standard-for-image-data-mining-restrictions/.

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?

As an initial matter, it must be noted that independent creators are at a disadvantage whenever considering technological solutions and monitoring for theft of their works. Most independent creators do not have the resources or technical expertise to regularly monitor for theft of their works or to take technical and other steps to prevent piracy. Adding to these burdens by expecting them to now monitor for AI ingestion of their works and to use technological solutions to prevent ingestion is an insurmountable obstacle for most independent creators. And even for the small fraction of creators who are able to overcome these hurdles, how can they be expected to opt out in the first instance if they don't know when, where, how, and by whom their works are being used?

As to the technical and practical obstacles relating to unlearning and robots.txt, please see our response to question 9.2.

Regarding the second part of this question, it is feasible to get consent on a mass scale. In fact, it happens frequently. Examples include licensing for music streaming services and voluntary collective licensing through organizations like the Copyright Clearance Center, both of which involve vast amounts of copyrighted works. Mass copyright infringement should not be rewarded by creating a special copyright exception for AI ingestion. Like others that have proceeded them, AI companies should be required to get authorization for any and all works they use (unless a defense applies under the law or if the works are in the public domain). The idea that just because it may be harder to get consent from copyright owners when large volumes of works are being used, it is therefore not infringement, would simply incentivize infringers to illegally copy more as a means for avoiding infringement—that cannot possibly be the law.

Getting a license, even when a large volume of works is being used is not difficult; there are flexible licensing models available and many different copyright management organizations to implement them. (See the response to question 10.2 for discussion of CMOs.) As noted in the *Warhol* decision, "licenses… are how [creators…] make a living. They provide an economic

incentive to create original works, which is the goal of copyright."[150] Thus, there is every incentive for copyright owners to make it easy for AI companies to license their works.

Not only is it feasible to get consent in advance, it is also often beneficial to the AI company to do so. In addition to not having to worry about liability, licensors can provide datasets that are well organized, have high-quality content, and importantly have consistent and high-quality metadata. (This is discussed in more detail in our responses to questions six and its subparts.) In most instances, licensors can also convey worldwide rights, which (largely) obviates the need for AI companies to navigate the plethora of different rules and requirements across the globe.

Lastly, for some copyrighted works, like music, it is not necessarily the case that a "large volume of works" is necessary to train an AI model, when in reality that is far from a certainty. Different AI models operate differently. Certain music AI models may not need to copy any copyrighted songs at all. For example, in the Office's listening session on music and sound recordings, Alex Mitchell, CEO and Co-Founder of Boomy, explained that Boomy does not ingest any copyrighted music.[151] And even for those AI models that do ingest music, there are reports that suggest a model trained with higher-quality music, but lower amounts of licensed music (MusicGen) actually work better than music AI models trained with more, but likely less quality, music.[152] That logic also applies to other models, like large visual models (LVM) and LLMs. While an LLM arguably needs to copy a large volume of text to function optimally, if the text is high-quality text licensed from copyright owners, that volume can be considerably less.

---

[150] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 143 S. Ct. 1258, 1278 (2023).

[151] *See* Listening Session on Music and Sound Recordings, held by U.S. Copyright Off. (May 31, 2023), https://www.copyright.gov/ai/listening-sessions.html#sound-recordings.

[152] *See* Matt Mullen, *AI Music Wars: Meta Takes on Google and Releases Its Own AI Music Generator – But Whose Is Better?*, MUSICRADAR (June 16, 2023), https://www.musicradar.com/news/meta-google-ai-music-wars-musicgen (concluding that Meta's product, which is trained on significantly less music than Google's product, creates better music); *see also* Jade Copet, et al., *Simple and Controllable Music Generatio*n, CORNELL UNIV. ARXIV 2, 7 (June 8, 2023), https://arxiv.org/pdf/2306.05284.pdf ("[H]uman evaluation suggests that MusicGen yields high quality samples which are better melodically aligned with a given harmonic structure, while adhering to a textual description . . . Results suggest that MusicGen performs better than the evaluated baselines as evaluated by human listeners, both in terms of audio quality and adherence to the provided text description.").

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?

If an opt-out request (whether delivered via a flag, watermark or by other means) is not honored, at the very least, this should be evidence of willful infringement in awarding statutory damages. And, if, in the future, there is a government issued-license that is necessary for AI companies to operate (as has been suggested by some stakeholders and policymakers[153]), and if there is failure to honor opt-out requests, then that license should be revoked and the AI company should be fined by the government agency, as well as the AI company being liable for willful copyright infringement(s).

With regard to potential other causes of action beyond copyright infringement, as is the case with any wrongful act, just because existing remedies exist under one law (e.g., the copyright law), does not preclude additional or future causes of action from being brought in the same action for the same act.

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?

Developments in AI should not alter established principles of copyright, including rights, ownership, and standing.[154] In the context of this question, we take no position on other existing or future laws.

[153] Khari Johnson, *Senators Want ChatGPT-Level AI to Require a Government License*, WIRED (Sept. 9, 2023), https://www.wired.com/story/senators-want-chatgpt-ai-to-require-government-license/.

[154] The determination who the owner is and what rights that person has may depends not only on the Copyright Act but also on contract law.

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

First, this question is phrased in a way that implies that copyright owners' consent may not be required for use of their works to train generative AI models. Consent is required, absent a valid defense or exception. That is the law. This question should instead ask: "*Because* copyright owners' consent is required to train generative AI models, how can or should licenses be obtained*.*"

Licenses for use of copyrighted works for generative AI training can be obtained in the usual way—by contacting the copyright owner, the owner's agent, or in some situations a collective rights management organization. There are numerous examples of AI licenses already being negotiated and completed in this manner, which we discuss below and in other responses. In addition, licenses are being developed now that likely would have been available before massive amounts of works were ingested without permission if AI developers had reached out to copyright owners.

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

Yes, voluntary licensing is feasible, as evidenced by existing agreements between AI developers and copyright owners for generative AI training (and other previous technological innovations in the way copyrighted content is used and distributed) and licenses that are being developed by rights owners. As history has shown, creators and copyright owners are usually willing to license their works; that is, of course, how creators typically earn a living (in addition to sales). Many creators and rightsholders already license their copyrighted works for commercial AI uses, and many of those that do not are in the process of doing so. Copyright law incentivizes those creators and rightsholders to lawfully enhance and aggregate their copyrighted works for that purpose—such as through semantic enrichment, metadata tagging, content normalization, and data cleanup.

There is already a high demand for corpuses of copyrighted works to train AI systems, and copyright owners have already entered into licensing agreements (or are offering licenses) for

text and data mining (TDM) uses.[155] The licensing activity in the TDM markets (which we discuss in more detail in responses to question six and its subparts) is evidence of existing markets for the use of copyrighted works for AI training and development, and it is important that the conditions of those licenses are respected and that they are not undermined by new, unwarranted exceptions that excuse unauthorized uses. In contrast to earlier forms of AI, these new generative AI models are ingesting copyrighted works to generate works that compete in the same market as the ingested works. In some cases, the output could qualify as derivatives of the ingested, copyrighted works.[156] Preserving opportunities for licensing and authorization of copyright works for ingestion could help to mitigate or prevent harm to copyright owners and creators arising from AI output that supersedes or supplants the market for the ingested works.

Where a copyright owner offers licenses for uses relating to the training of AI systems, it is essential that the licenses be respected by any copyright or AI legal regime. The existence of a licensing market is one factor that may weigh against a finding that copying without the permission of the copyright owner is excused by the fair use defense.[157] The marketplace should continue to properly value and incentivize creativity, and AI policy should not interfere with the right of copyright owners to license, or choose not to license, their works for AI purposes.

As discussed in our response to question 8.1, some have argued that licensing copyright protected material for AI ingestion is not practical because, in order to be successful, AI systems

---

[155] *Associated Press, OpenAI Partner to Explore Generative AI Use in News*, REUTERS (July 13, 2023, 1:08 PM), https://www.reuters.com/business/media-telecom/associated-press-openai-partner-explore-generative-ai-use-news-2023-07-13/; *see Elsevier Text and Data Mining (TDM) License*, ELSEVIER, https://beta.elsevier.com/about/policies-and-standards/text-and-data-mining/license?trial=true (last visited Oct. 27, 2023); COPYRIGHT CLEARANCE CENTER, https://www.copyright.com/solutions-rightfind-xml/ (last visited Oct. 27, 2023); Kyle Wiggers, *This Startup Wants to Train Art-Generating AI Strictly on Licensed Images*, TECHCRUNCH (Apr. 13, 2023, 8:30 AM), https://techcrunch.com/2023/04/13/this-startup-wants-to-train-art-generating-ai-strictly-on-licensed-images/?guccounter=2.

[156] In one of the copyright infringement lawsuits filed against Stability AI, evidence was presented of output that was substantially similar to a specific photograph that was ingested without authorization. *See* Getty Images, Inc. v. Stability AI, Inc., 23cv00135 (D. Del. Feb. 3, 2023).

[157] *See* Am. Geophysical Union v. Texaco Inc., 60 F.3d 913, 929 (2d Cir. 1994) ("[I] is indisputable that, as a general matter, a copyright holder is entitled to demand a royalty for licensing others to use its copyrighted work, see 17 U.S.C. § 106 (copyright owner has exclusive right "to authorize" certain uses), and that the impact on potential licensing revenues is a proper subject for consideration in assessing the fourth [fair use] factor . . . ").

must use *every* piece of available content. It might be desirable to train an AI tool on as much content as practically can be scraped from the internet, but—as evidenced by the many successful AI developers that do not scrape the internet for ingestion purposes—that does not make it necessary.

Just because AI developers *want* to use everything to train their systems doesn't mean that it is necessary or justified. There is evidence that generative AI can succeed when ingesting only copyrighted materials that are properly licensed or in the public domain. We understand that the large general purpose LLMs do require a lot of ingestion material, but there is no reason that they cannot rely on public domain, open access and licensed texts. In the image space, Adobe's Firefly suite of generative AI tools—which are trained on proprietary stock images, licensed images, and public domain images whose copyrights have expired—has seen broad consumer adoption.[158] Additionally, leading AI developer Nvidia has partnered with Getty Images to build new generative AI technologies that ingest only fully licensed works, and IBM recently announced a collaboration with Adobe to assist customers in implementing generative AI models based on Adobe's Firefly technology.[159] These are just some examples that demonstrate that the foundation of an AI model can be built on licensed works.

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?

As noted in our response to question two, generative AI impacts each of the different copyright communities differently. Each industry has unique business models and different approaches to licensing. Because a voluntary collective licensing model is desirable for some, any extended

[158] Rashi Shrivastava, *Adobe Brings Its Generative AI Tool Firefly To Businesses*, FORBES (June 8, 2023). https://www.forbes.com/sites/rashishrivastava/2023/06/08/adobe-brings-its-generative-ai-tool-firefly-to-businesses/?sh=53f4dd94582b.

[159] Rick Merritt, *Moving Pictures: NVIDIA, Getty Images Collaborate on Generative AI*, NVIDIA (March 21, 2023). https://blogs.nvidia.com/blog/2023/03/21/generative-ai-getty-images/; *IBM Expands Partnership with Adobe To Deliver Content Supply Chain Solution Using Generative AI*, IMB NEWSROOM (June 19, 2023) https://newsroom.ibm.com/2023-06-19-IBM-Expands-Partnership-with-Adobe-To-Deliver-Content-Supply-Chain-Solution-Using-Generative-AI.

collective licensing or mandatory/statutory collective licensing model should not be considered outside of a narrow, industry-specific context (the conditions of which are discussed in our answer to question five and in other responses).

As discussed below and in other responses, there are already voluntary collective licensing schemes that have emerged to supply targeted licensed uses in certain industries. It is important to note that these licensing mechanisms are the result of the free market and copyright owners' freedom to license (or not license). As such, direct licensing (including, where desired, on a collective basis) should *always* be the default approach. Exceptions to direct licensing should only apply when market inefficiencies make direct licensing impossible or virtually so, such as when there are a great number of individual rights holders and, even in those situations, those exceptions should *never* be compulsory.

There are numerous voluntary collective management organizations (CMOs) that administer copyright owners' rights. Perhaps the most well-known and established CMOs are in the music industry, namely, ASCAP, BMI, SESAC, and GMR.[160] Other CMOs in the music space include Vydia,[161] AWAL,[162] and Merlin.[163] Additionally, there have been agreements reached between music publishers and platforms using copyrighted works that establish licensing systems through which royalties are distributed.[164] And there are many others.

---

[160] *A Comprehensive Comparison of Performance Rights Organizations (PROs) In the US*, Paul Resinkoff, DIGITAL MUSIC NEWS (Feb. 20, 2018), https://www.digitalmusicnews.com/2018/02/20/performance-rights-pro-ascap-bmi-sesac-soundexchange/.

[161] VYDIA, https://vydia.com/ (last visited Oct. 27, 2023).

[162] AWAL, https://www.awal.com/ (last visited Oct. 27, 2023).

[163] MERLIN, https://merlinnetwork.org/ (last visited Oct. 27, 2023).

[164] *See NMPA and Peleton Announce Settlement of Litigation, Joint Collaboration Agreement*, PR NEWSWIRE (Feb. 27, 2020, 9:00 AM), https://www.prnewswire.com/news-releases/nmpa-and-peloton-announce-settlement-of-litigation-joint-collaboration-agreement-301012233.html; *NMPA and YouTube Reach Agreement to Distribute Unclaimed Royalties*, NMPA (Dec. 8, 2016), https://www.nmpa.org/nmpa-and-youtube-reach-agreement-to-distribute-unclaimed-royalties/; *NMPA and Roblox Strike Industry-Wide Agreement*, NMPA (Sept. 27, 2021), https://www.nmpa.org/nmpa-and-roblox-strike-industry-wide-agreement/.

Outside of music, there are many other CMOs for other types of works. These include, but are not limited to:

- American Society for Visual Arts Licensing (ASCRL), which collects foreign payments for works of visual art that are mandated by foreign law and distributes those payments to its members.[165]

- Artists Rights Society (ARS), which collects foreign payments for works of fine art that are mandated by foreign law and distributes those payments to its members.[166]

- Copyright Clearance Center (CCC), which collects and distributes license royalties for literary works.[167]

- Motion Picture Licensing Corporation (MPLC) and SWANK, which license on a non-exclusive basis the public performance of copyrighted motion pictures, television programs and other audiovisual works that were originally intended for personal use only.[168]

---

[165] AMERICAN SOCIETY FOR COLLECTIVE RIGHTS LICENSING, https://ascrl.org/ (last visited Oct. 27, 2023) ("Established by authors and rights holders, ASCRL collects foreign payments for visual works that are mandated by foreign law and distributes those payments to ASCRL members. ASCRL - The recognized leader in the administration and distribution of collective revenue for U.S. illustrators and photographers and U.S. published works.").

[166] ARTISTS RIGHTS SOCIETY, https://arsny.com/about/ (last visited Oct. 27, 2023) ("Artists Rights Society is a unique and inclusive alliance of forward-thinking visual artists who seek to actively participate in the broader economic and cultural vitality of their time. Founded in 1987, we harness our 30+ years of experience with the power and prestige of our 122,000-strong global collective in order to create exciting new projects and collaborations. ARS plays a vital role in protecting the intellectual property of artists, at a time when the ability to control the rights of their work is often challenged.").

[167] ABOUT CCC, https://www.copyright.com/company-about/ (last visited Oct. 27, 2023) ("A pioneer in voluntary collective licensing, CCC is a leading information solutions provider to organizations around the world.").

[168] *See* MOTION PICTURE LICENSING CORPORATION, https://www.mplc.org/ (last visited Oct. 27, 2023); SWANK, https://www.swank.com/ (last visited Oct. 27, 2023).

There are many other CMOs and more CMOs are in the process of being developed as result of AI.[169]

Many CMOs already operate without an antitrust exemption. To the extent there are antitrust concerns expressed by parties interested in negotiating collective licenses, one possible approach is to request a Business Review Letter (BLR) from the Department of Justice (DOJ) to evaluate any antitrust concerns and to obtain "guidance from the Department with respect to the scope, interpretation, and application of the antitrust laws to particular proposed conduct."[170] The problem with this approach is that it can take up to two years to receive a BLR, and thus would seem to move too slowly to address antitrust issues relating to collective licensing to AI companies. Providing the DOJ with additional resources so that BLRs can be obtained more quickly may be one possible solution to this problem.[171]

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?

No. Congress should not establish any new compulsory licensing regimes, and that includes a compulsory license regime related to generative AI. The Copyright Office[172] and Copyright

[169] Experienced executives in royalty collection and AI are developing platforms to solve the attribution and royalty-payment problem. For example, Dave Davis, former Chief Commercial Officer and the Motion Picture Licensing Corporation, and Jim Golden, former Chief Digital Officer at The Rockefeller Foundation are working on an initiative to address this as a business opportunity.

[170] *Business Reviews*, U.S. DEPT. OF JUST.: ANTITRUST DIVISION, https://www.justice.gov/atr/what-business-review#:~:text=Section%2050.6.,laws%20to%20particular%20proposed%20conduct (last visited Oct. 27, 2023).

[171] For business reviews concerning export trade, DOJ issue a response within 30 business days from the date that the Division receives all relevant data. The same time frame should be in place for responses to AI-related collective licensing inquiries.

[172] *See, e.g.*, U.S. Copyright Off., Analysis and Recommendations Regarding the Section 119 Compulsory License 6–7 (2019) ("Repeatedly, the Copyright Office has recommended that Congress phaseout the section 119 compulsory license for secondary transmissions of distant television programming by satellite . . . Copyright Office [maintains a] long-held view that a compulsory license 'should be utilized only if compelling reasons support its existence . . . .'"); *Music Licensing Reform: Hearing Before the Subcomm. on Cts. & Intell. Prop. of the Sen. Comm. on the Judiciary*, 109th Congress (2005) (statement of Marybeth Peters, Reg. of Copyrights) ("The Copyright Office has long taken the position that statutory licenses should be enacted only in exceptional cases, when the marketplace is incapable of working . . . After all, the Constitution speaks of authors' "exclusive rights to their Writings," and in general authors should be free to determine whether, under what conditions and at what price they will license the

Alliance[173] have long opposed government mandates or initiatives that would diminish competition in the marketplace and the right of creators and copyright owners to control the dissemination of their works to the public and the terms and conditions thereof. As noted in our responses above, free markets should be respected and direct licensing in the free market should be the default.

The U.S. Constitution recognizes that creators' contributions and investments and the public's interest in accessing these works are best realized through the rights and freedoms afforded by copyright. Government mandates and initiatives severely upset the balance of interests in allowing public access to creative works and rewarding the inspired efforts of their creators. Under such edicts, copyright owners would be unable to freely utilize the full scope of their exclusive rights.

Copyright owners need to recoup their investments in the creation and marketing of their works. If copyright owners cannot recoup these investments, they will not be able to sustain their businesses and careers, will be discouraged from creating and distributing new works for the public to enjoy, and will not be able to uphold the highest standards of quality and integrity in the copyrighted works they produce.

Competitive markets result in better products and services, as well as increased choices for consumers. But undue government interference with these markets has the opposite effect. Markets cannot remain competitive and efficient when federal or state governments interfere in

---

use of their works."); Cable Compulsory License; Definition of Cable Systems, 56 Fed. Reg. 31,580, 31,590 (proposed July 11, 1991) (to be codified at 37 C.F.R. pt. 201) ("Compulsory licenses are limitations to the exclusive rights normally accorded to copyright owners and, as such, must be construed narrowly to comport with their specific legislative intention."); U.S. Copyright Off., The Cable and Satellite Carrier Compulsory Licenses: An Overview and Analysis 127 (1992) (citing Cable Compulsory License; Definition of Cable Systems, 56 Fed. Reg. at 31,590); *Copyright Broadcast Programming on the Internet: Hearing Before the Subcomm. on Cts. & Intell. Prop. of the H. Comm. on the Judiciary,* 106th Cong. 25–26 (2000) (statement of Marybeth Peters, Reg. of Copyrights) ("The Copyright Office has long been a critic of compulsory licensing for broadcast retransmissions. A compulsory license is not only a derogation of a copyright owner's exclusive rights, but it also prevents the marketplace from deciding the fair value of copyrighted works through government-set price controls.").

[173] *See* COPYRIGHT ALLIANCE, POSITION PAPER ON GOVERNMENT-MANDATED BUSINESS MODELS AND INITIATIVES 1 (2022), https://copyrightalliance.org/wp-content/uploads/2022/09/Gov-Mandated-business-model-position-paper.pdf ("The Copyright Alliance supports competition in the marketplace and the right of creators and copyright owners to control the dissemination of their works to the public.").

ways that unfairly or otherwise inappropriately favor certain types of business models, products, services, and providers over others. Such interference discourages corporate copyright owners from developing innovative business models and creates significant obstacles in their abilities to do so, which leads to fewer options for the public to access new and quality copyrighted works as products and services in the marketplace.

When the government puts its thumb on the scale to favor certain business models or mandates the terms under which works are made available to the public, it undermines the Constitutional purposes and goals of federal copyright law and destroys the existing incentives for copyright owners to create and disseminate a diverse array of creative works to the public. Neither the Constitution nor the Copyright Act authorizes federal or state governments to restrict, diminish, or eliminate copyright owners' exclusive rights as a condition for the receipt of federal funding or as the basis for achieving any other regulatory objective.

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

As noted in our responses, we recognize that, so far, the AI training licensing market has largely eluded individual creators. It is our hope that AI companies will soon begin to also license the works of individual creators for ingestion purposes, just as they have begun doing for some businesses with large corpuses of high-value copyrighted works. In fact, recently, we have seen evidence that that might be the case.[174] As we note in other responses, we believe that direct or

[174] For example, visual communication company, Canva, launched a suite of generative AI tools called Magic Studio that allows its users to generate videos, presentations, and other designs from text prompts. In its announcement, Canva announced that it is dedicating $200 million over the next three years in creator and AI royalties. Charlotte Trueman, *Canva Bolsters AI Offerings, Providing Copyright Indemnity for AI-Generated Images*, COMPUTERWORLD (Oct. 5, 2023, 6:15 AM), https://www.computerworld.com/article/3708249/canva-bolsters-ai-offerings-providing-copyright-indemnity-for-ai-generated-images.html. Adobe has started paying "bonuses" to artists whose images were used in training though it "maintains that it has the legal right to train its Firefly image-synthesizing AI model on works uploaded to its platform . . . The payout is a function of (a) the number of images submitted and (b) the number of times someone licensed those images in the preceding 12 months. *See* Jeremy T. Elman et al., The AI Update: October 5, 2023, DUANE MORRIS: THE A.I. BLOG (Oct. 5, 2023), https://blogs.duanemorris.com/artificialintelligence/2023/10/05/the-ai-update-october-5-2023/#more-157; Kyle Wiggers, *How Much Can Artists Make from Generative AI? Vendors Won't Say*, TECHCRUNCH (Sept. 30, 2023, 10:30 AM), https://techcrunch.com/2023/09/30/how-much-can-artists-make-from-generative-ai-vendors-wont-say/ ("Adobe, which trains its family of generative AI models, called Firefly, . . . says that it'll pay out a once-a-year 'bonus' that's 'different for each contributor.'. . .Kneschke's survey found that the average revenue from the Contributors Fund was $0.0078 per image while the median was $0.0069 per image. Assuming those numbers are accurate, a photographer with around 2,000 images would make roughly $15 . . .").

voluntary collective licensing should be the default. However, if that does not transpire and, as noted in more detail in our responses to questions five and 9.2, there may be a general consensus of organizations and individual creators within a particular industry (for example, the book publishing industry) who are willing to accept an extended collective license approach that is narrowly targeted to a particular industry and a particular type of work(s), and would not directly or indirectly effect (through inadvertent consequences or otherwise) those industries and works not intended to be covered by the legislation.

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

Licensing regimes should be tailored to the type of work at issue. (See our responses to questions 2, 5, 9.2, 10.2, and 10.4.) However, in all cases, voluntary, free-market licensing should be the default.

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)?

In response to the first part of this question, licensing for generative AI purposes is no different than any other type of licensing—the legal and technical issues are the same as anything else. As for practical considerations, it depends on the type of work and AI model. Licensing might be on a larger scale, but that does not mean that licenses are not required. (See responses to "large scale" issues in questions 8.1, 9.3, and 10.1.) This is evidenced by licensing deals being entered into by AI developers and copyright owners for use of vast archives of copyrighted material, such as the agreement between OpenAI and the Associated Press discussed in response to question 6.

The second part of this question asks who should be responsible for securing a license. The bottom line is that any entity that reproduces, distributes, creates a derivative of, or engages in any activity that implicates copyright owners' exclusive rights needs a license (absent a clear exception in the law or valid defense), and it is that entity's responsibility to secure it (or a permissible sublicense). Indeed, the issues raised in the question can be dealt with most efficiently by contract as early as possible in the generative AI value chain. That includes any entity that is curating the training dataset or ingesting the copyrighted material into the model, and the license that they obtain should include the right to ingest and reproduce copyrighted works for training purposes, along with the right to make the model available, distribute the model, etc.

Importantly, just because an initial or previous use was considered fair use or there was a license, doesn't necessarily mean that a downstream use is fair use or licensed. This concept was confirmed by the Supreme Court's *Warhol* decision, which found that Vanity Fair's initial license to use Goldsmith's photo for the "Orange Prince" silkscreen that Andy Warhol subsequently created did not immunize the Andy Warhol Foundation from infringement liability when it licensed the work to Conde Nast following Prince's death in 2016.[175] Ultimately, each specific use of a copyrighted work by each participant in the AI supply chain must be analyzed independently regardless of an initial license or fair use determination.

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 explained earlier, there are at least two instances of infringement that can occur when copyrighted works are ingested by AI systems for the purpose of generating new content. The first is infringement during the ingestion process that occurs when an unauthorized reproduction is made. The second type occurs when a specific output of the generative AI system infringes

[175] *See generally* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 143 S. Ct. 1258 (2023) ("Taken together, these two elements—that Goldsmith's photograph and AWF's 2016 licensing of Orange Prince share substantially the same purpose, and that AWF's use of Goldsmith's photo was of a commercial nature—counsel against fair use, absent some other justification for copying.").

rights in a particular work—for example, by generating material that is substantially similar to the copyrighted elements of an ingested work.

Because this question is in the training section of the NOI, and not the infringement section, we answer this question as follows: The question of whether a particular work contributes to a particular output from a generative AI system is wholly irrelevant to the infringement analysis for training/ingestion. When a copy is made for ingestion purposes there is an infringement unless the use is licensed, or an exception (like fair use) applies. Therefore, the answer to the question of whether it is technically feasible to identify the degree to which a particular work contributes to a particular output from a generative AI system makes no difference with regard to ingestion and training.

If this question was in the infringement section of the NOI, then we would answer as follows: For purposes of determining whether an AI-generated *output* is infringing the test to determine infringement is substantial similarity (and access). If works are substantially similar to one another and access to the copyrighted work is established, then the question of whether a particular work "contributes" to a particular output is technically irrelevant to any copyright analysis, since copying is then presumed. While there are existing technologies that compare ingested works to outputs for similarities, the results are not proxies for substantial similarity.

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

First of all, this question should ask the inverse: *What would be the economic impacts of* ***not*** *requiring licensing on the creative community?* The way the question is phrased implies that a "licensing requirement" is something that is being considered, when it is already the law (again, absent a clear exception or fair use defense). The answer to the inverse is that it would completely undermine existing and potential markets and cause immeasurable harm to copyright owners. Further, in the absence of licensing, there will be (and already is) considerable litigation, which is expensive, time-consuming and diverts attention away from AI development and the creation of copyrighted works.

To answer the question as it is presented, requiring a license for the ingestion of copyrighted works by generative AI systems would not adversely impact development and adoption of AI technologies. Licensing copyrighted works is a normal cost of doing business, and licenses are entered into across the spectrum of copyright industries. Whether compulsory, collective, or direct, licenses dictate the use and distribution of every type of copyrighted work from software to music to literary works and more. While the type and terms of licenses may differ from industry to industry, they are an established part of the greater creative ecosystem, and their application to generative AI should be no different.

It is the choice of any AI developer as to what and how many copyrighted works it ingests into a model for training purposes, and any argument to the contrary that an AI system must ingest as many works as possible (and that licensing is impossible) is a red herring used to justify massive infringement that has already occurred. (See our responses to questions 8.1, 9.3, and 10.1.) The choice that AI developers have was illustrated wonderfully by Adobe's Dana Rao in his testimony before the Senate Judiciary Committee, Subcommittee on Intellectual Property's hearing on *Artificial Intelligence and Intellectual Property—Part II: Copyright.*[176] Rao explained that Adobe "chose a path that supports creators and customers by training on a dataset that is designed to be commercially safe."[177] That meant training its Firefly model only on licensed images from its own Adobe Stock photography collection, and if needed, to expand its dataset to include openly licensed content and public domain images where copyright has expired.[178] As we note in response to question 8.1, Adobe's Firefly suite of generative AI tools have seen broad consumer adoption and represent how AI technology can successfully augment human artistic expression when trained on proprietary or licensed copyrighted works.

---

[176] *Artificial Intelligence and Intellectual Property—Part II: Copyright, Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. 3–4 (2023) (written testimony of Dana Rao, Exec. Vice President, Gen. Couns., & Chief Trust Officer, Adobe Inc.), https://www.judiciary.senate.gov/imo/media/doc/2023-07-12_pm_-_testimony_-_rao.pdf.

[177] *Id*. at 4.

[178] *Id*.

Finally, when considering the impact of licensing (or *not* licensing), the fourth fair use factor is paramount. Notably, the fourth factor *does not require* courts to consider the economic impact of securing a license, but it does explicitly require them to consider the economic impact of *not* securing a license. The text of section 107(4) of the Copyright Act does not say anything about the potential impact to *the user or the user's market*. Instead, courts shall consider "the effect of the use upon the potential market for or value *of the copyrighted work*." This distinction is critical, and it is one that must be taken into account under any fair use analysis related to the unauthorize ingestion of copyrighted works by AI developers.

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

One issue related to ingestion of works that was not addressed in the questions is the stripping of metadata that occurs in the process of ingestion. Metadata and other copyright management information (CMI) helps to identify works that have been ingested. When that CMI is removed during the ingestion process, it makes it difficult, if not impossible, for copyright owners to determine whether their works have been ingested. Removing this CMI also conceals the infringement and makes infringement more difficult to prove. Additionally, removing CMI may strip metadata that includes opt-out flags. As such, it compounds the harms caused by infringement and should be punishable under section 1202 of the Copyright Act. It is imperative that any CMI associated with a work not be removed or altered during the ingestion process. Removal or alteration of CMI should be considered to be evidence of willful infringement, and result in larger damage awards against AI developers found liable for infringement.

TRANSPARENCY & RECORDKEEPING

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

There are numerous benefits of transparency, many of which are not specific to copyright. We limit our comments to the benefits of transparency to copyright in the context of AI. Developers of AI models that are made available directly or indirectly to the public that ingest copyrighted works owned by third parties without a license should be required to satisfy transparency standards related to the collection, retention, and disclosure of the copyrighted works they use to train AI. Adequate transparency regarding ingestion of copyrighted works goes a long way in helping to ensure that copyright owners' rights are respected. Best practices from corporations, research institutions, governments, and other organizations that encourage transparency around AI ingestion already exist, and they enable users of AI systems or those affected by its outputs to know the provenance of those outputs.[179] There is no reason these same responsibilities should not also apply to the ingestion of copyrighted works. However, it's also important to note that there is a big difference between voluntary best practices and binding legal requirements, which is why we support the imposition of legal obligations related to transparency and record keeping.

As discussed in greater detail in response to question 15.1, it is vital that AI developers be legally required to maintain adequate records of what materials were used to train the AI (and how those materials are used) and to make those records publicly accessible and searchable as appropriate, subject to two important exceptions. First, this obligation should not apply to any ingested works of which the AI developer is also the copyright owner.[180] And, second, where there is a license between the AI developer and the copyright owner(s) of the works ingested that authorizes such

[179] *See, e.g.*, CONTENT AUTHENTICITY INITIATIVE, https://contentauthenticity.org/ (last visited Oct. 25, 2023) ("Our tools make it easy to indicate when AI was used to generate or alter content. Information about specific AI models used and more can be conveyed to viewers, helping to prevent misinformation and increase transparency around the use of AI.").

[180] Unless contrary to obligations under other laws, contracts, or collective bargaining agreements.

ingestion for AI development purposes, this obligation should be subject to whatever reasonable confidentiality provisions those parties have negotiated in that license.[181]

As noted earlier in our comments, adequate and appropriate transparency and record-keeping benefits both copyright owners and AI developers in resolving questions regarding infringement, fair use, and compliance with licensing terms. Those practices have the added benefit of also promoting safe, ethical, and unbiased AI systems.

15.1. What level of specificity should be required?

As noted above, except where the AI developer is also the copyright owner of the works being ingested by the AI system or where a license between the copyright owner and the AI developer dictates specific terms, when AI models that ingest copyrighted works owned by third parties without a license are made available directly or indirectly to the public, AI developers should be required to maintain certain records relating to the works ingested. These records should indicate:

- which copyrighted works are ingested;
- how those works are used;
- when the works were ingested;
- the legal basis for collection;
- how the work was acquired (e.g., through a license);
- whether any modifications, additions, or deletions have been made to a training dataset acquired from a third party;
- whether copies have been disseminated to third parties; and
- whether copies of the works are retained.

[181] As discussed in more detail in our responses to questions 6.1 and 8.3, data laundering is a major issue in the AI context. The practice of data laundering is an attempt to avoid accountability. *See* Andy Baio, *AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability*, WAXY.ORG (Sept. 30, 2022), https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/.

Where copies of the works ingested are retained, records should also indicate how long copies are retained and what security measures are in place to prevent the copies from being leaked through a cyberattack or otherwise or inadvertently disclosed. Such records should be maintained for a minimum of seven years from the time at which the AI system is no longer being publicly deployed. Caution in the manner of disclosure should be exercised so that these public disclosures do not further propagate the spread or use of unlicensed copyrighted works.

15.2. To whom should disclosures be made?

The disclosures outlined in our response to question 15.1 should be made publicly available and searchable as appropriate—as noted above, subject to whatever reasonable confidentiality provisions the parties to a license may negotiate. Again, caution in the manner of disclosure should be exercised so that these public disclosures do not further propagate the spread or use of unlicensed copyrighted works.

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

Developers of AI systems that incorporate models from third parties should have the same obligations as the developer of the underlying models, which should include keeping appropriate records and publicly disclosing the AI model that is being incorporated, subject to the conditions noted above. In certain instances, it may also make sense to require developers to also disclose information regarding any modifications, additions, or deletions they have made to the training dataset acquired from the third party. Further, if the developer is operating under an upstream party's license, it should have a sublicense and the upstream party should have right to sublicense.

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?

Recordkeeping should not be onerous or expensive. It simply requires keeping track of what datasets are used when, and if the AI company creates their own datasets, what sources they

used. While the volume of material used in training is often large, systems are not trained on unorganized raw files. Because the datasets are organized and cleaned before training, a commercial market already exists to help AI developers keep such records.[182] Recordkeeping costs are simply a cost of doing business that is necessary in order to promote safe, responsible, respectful, ethical, and unbiased AI systems. These costs must be borne by developers of AI models who are neither owners nor licensees of the copyrighted works at issue.

The cost to copyright owners of *not* imposing a record-keeping system is enormous.

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

As an initial matter, and as we detail in other responses, the ingestion of copyrighted material by AI systems implicates the reproduction right. When an unauthorized copy is made of a work protected by copyright, there is a violation of the copyright owner's right to reproduce the work unless the copier has a valid defense or a license. Where an AI developer obtains a license prior to ingestion, that license serves the function of notice. The specific terms of a license agreement can further specify any additional obligations related to notice. To be clear, notice alone—i.e., without permission—is not enough and would run afoul of the law.

As a practical matter, AI developers are best situated to know what works have been ingested and to make that information available. If AI developers use publicly available datasets, they should be able to identify that dataset. If they create their own datasets, they need to be transparent about what works are included in the dataset, subject to the conditions we identify in response to question 15. The burden should not be on the copyright owner to determine if their works have been ingested.[183]

---

[182] *See, e.g.*, WHYLABS, https://whylabs.ai/ (last visited Oct. 26, 2023) ("Structured or unstructured. Monitor raw data, feature data, predictions and actuals . . . Integrate seamlessly with existing data pipelines and multi-cloud architectures."); SUPERANNOTATE. https://www.superannotate.com/ (last visited Oct. 26, 2023) ("Build, fine-tune, iterate, and manage your AI models faster with the highest- quality training data.").

[183] It also goes without saying that it would be exceedingly difficult and time-consuming for them to do so.

Where works have already been ingested without a license, for transparency purposes,[184] except for the instances we described earlier where transparency requirements should not apply, AI developers should be required to:

- make publicly available searchable databases of copyrighted works (text, music, images, etc.) that copyright owners can use to determine if their works have been trained by that company's AI model using standard metadata (e.g., ISRC or artist and track name for musical recordings);

- make publicly available a searchable database of URLs of webpages that have been scraped publicly available;[185] and

- ensure that, when prompted, their AI models disclose whether the model was trained on a particular work.

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?

There are existing laws, like federal and state privacy laws, that may require developers of AI models or systems to retain or disclose records about the materials they used for training. However, a discussion of those laws is beyond the scope of the Copyright Alliance's mission, so we do not address them here.[186]

[184] We are only discussing transparency in this section and nothing in our response should be construed to absolve the AI developer of any responsibility to license works or "unlearn" works that are unlicensed.

[185] While a list of URLs is helpful, a list of URLs by itself is not sufficient to determine if a work has been infringed, since the copyright owner may not know if their content is available on a specific website.

[186] It should be noted, however, that, based on ongoing discussions within the Biden Administration and in Congress, it seems likely that before too long the United States will enact laws that require developers of AI models or systems to retain or disclose records about the materials they used for training, and such laws will and should be applicable to copyrighted works.

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?

The factors for determining copyrightability have been well developed and articulated throughout copyright law jurisprudence and continue to withstand the advent of new technologies. In *Feist Publications v. Rural Telephone*, the Supreme Court explained that a work of authorship must possess "at least some minimal degree of creativity" to sustain a copyright claim.[187] And in *Burrow-Giles Lithographic Co. v. Sarony*, the Supreme Court noted that the question of copyrightability is to be determined based on "the existence of those facts of originality, of intellectual production, of thought, and conception on the part of the author."[188] Finally, in *Thaler v. Perlmutter,* the district court reiterated that copyright only protects the unique value of human creativity, noting that courts have "uniformly declined to recognize copyright in works created absent any human involvement".[189] In determining whether a work generated using AI is copyrightable, these longstanding standards of copyrightability will apply no differently than they do in other contexts. The question of copyrightability must be determined on a case-by-case basis, based on the particular facts at issue.

19. Are any revisions to the Copyright Act necessary to clarify the human authorship requirement or to provide additional standards to determine when content including AI-generated material is subject to copyright protection?

No revisions to the Copyright Act are "necessary to clarify the human authorship requirement," especially in view of the recent decision of the U.S. District Court for the District of Columbia in

[187] *Feist Publications, Inc. v. Rural Telephone Service Co.,* 499 U.S. 340, 345 (1991).

[188] *Burrow-Giles Lithographic Company v. Sarony*, 111 U.S. 53, 59-60 (1884).

[189] *See* Thaler v. Perlmutter, No. 22-1564, 2023 U.S. Dist. LEXIS 145823, at *15–17 (D.D.C. Aug. 18, 2023) (citing, among other cases, Urantia Found. v. Maaherra, 114 F.3d 955, 958–59 (9th Cir. 1997); Kelley v. Chicago Park Dist., 635 F.3d 290, 304–06 (7th Cir. 2011); Naruto v. Slater, 888 F.3d 418, 420 (9th Cir. 2018)).

Thaler v. Perlmutter granting the U.S. Copyright Office's motion for summary judgment and confirming that "human authorship is an essential part of a valid copyright claim" and "a bedrock requirement of copyright."[190] When material is wholly generated by AI and there is no human authorship involved, as was the case in *Thaler*, that material should not be protected by copyright. The Copyright Office and at least one court are in agreement here, and thus no change to the Copyright Act on these issues is warranted.

We found the second part of this question, which asks—"[a]re any revisions to the Copyright Act necessary to provide *additional* standards to determine when content including AI-generated material is subject to copyright protection"—to be somewhat unclear. Presently, the Copyright Act does not contain any "standards" for determining authorship. Section 102 of the Act makes clear that "[c]opyright protection subsists…in original works of *authorship* fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device." (emphasis added) But no place in the Act itself is a "standard" for determining "authorship" provided (and thus the term "additional" here is confusing). To the extent there exist "standards" for determining authorship in works that contain both elements of AI-generated output and human creativity, those "standards" are found in the Copyright Office's recent *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence*,[191] not the Copyright Act.[192] (We discuss the *Guidance* in our answer to question 34). Thus, as stated above, because the Copyright Office and the courts, in decisions such as *Thaler* and *Naruto v. Slater*,[193] are reaching the correct conclusions in cases where material is wholly generated by AI, we do not think any change to the Copyright Act is necessary.[194]

---

[190] Thaler v. Perlmutter, No. 22-1564, 2023 U.S. Dist. LEXIS 145823, at *2 (D.C. Cir. 2023).

[191] *See generally* Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 88 Fed. Reg. 16190, 16190–94 (Mar. 16, 2023), https://www.copyright.gov/ai/ai_policy_guidance.pdf?loclr=eanco.

[192] The Copyright Office has indicated that it will make updates to the human authorship requirement section in the Compendium of U.S. Copyright Practices to reflect the AI registration guidance, and we support those updates.

[193] 888 F. 3d 418 (9th Cir. 2018).

[194] We note that we do not believe it appropriate to include copyrightability standards in the Act. Any such standards are better to be discussed in guidance, circulars, the Compendium (or perhaps regulations), but not statutory law that will require an Act of Congress to update or change in the future.

In cases where material is an amalgam of both AI-generated output and human creativity, please see our answers to questions 18 and 34.

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?

AI tools have the potential to assist human creativity, much like other creative tools that have come before it. However, Copyright protection for *wholly* AI-generated material is *not* desirable as a policy matter. As noted throughout our answers, wholly generated AI material that is based on copyrighted works ingested by AI developers without compensating the creator or obtaining their permission to ingest their works has the potential to supplant the market for the ingested works. Policymakers should be discouraging such activities, not incentivizing them by granting legal protection to material manufactured outside of the realm of human authorship.

In a world where human creators are competing with machines, the incentives established by copyright law are more important than ever. So why would policymakers want to level the playing field by incentivizing machine creation by affording copyright protection to wholly AI-generated output? If the Copyright Office and other policymakers give incentives to generate AI content, the sheer volume and speed with which AI material is generated could obliterate the markets for much human creation. Our popular culture will be overtaken by low quality, AI-generated works because the cost of human creation would be deemed too burdensome in comparison to using AI.

As noted elsewhere in these comments, our views are limited to copyright law. We take no position on whether other types of existing or future legal protection are or may be desirable.

For similar reasons to those discussed in the first paragraph of this response, additional copyright protection is not necessary to encourage development of generative AI technologies and systems. It is important to understand that AI companies are selling a service—they are not selling the AI-

generated outputs. Thus, they don't need copyright incentives for AI-generated outputs. And as to the computer programs and other aspects of their businesses, AI companies are already able to rely not just on copyright but also a combination of various other intellectual property protections (e.g., patents, trade secret) for legal protection.

The pace of AI development demonstrates that there are already adequate incentives in place. Today, there exist a large number of AI developers and systems.[195] That number has grown exponentially over the past year and is likely to continue to increase in the coming months and years. Similarly, the number of AI users and customers has also expanded significantly.[196] It is abundantly clear that no additional copyright-related incentives are needed to encourage AI developers and systems to enter the marketplace and prospering.

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 noted above, we do not believe any new form of copyright protection is necessary or 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?

The Copyright Clause of the Constitution grants Congress the power "[t]o promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries." We do not believe the Clause can be interpreted to support the claim that the Constitution permits copyright protection for non-humans.

---

[195] *See* Mark Webster, *149 AI Statistics: The Present and Future of AI At Your Fingertips*, AUTHORITYHACKER (Oct. 6, 2023), https://www.authorityhacker.com/ai-statistics/.

[196] *See id*.

Central to the Copyright Clause is the concept of creator incentivization, which is not applicable to machines that do not need or comprehend incentivization. As the U.S. District Court for the District of Columbia recently explained in *Thaler v. Perlmutter*:

> "The act of human creation—and how to best encourage human individuals to engage in that creation, and thereby promote science and the useful arts—was thus central to American copyright from its very inception. Non-human actors need no incentivization with the promise of exclusive rights under United States law, and copyright was therefore not designed to reach them."[197]

The court's opinion adopts the Copyright Office's position (responding to Thaler's complaint) that "the Constitutional purpose of copyright is to incentivize *humans* to create expressive works" and that "human creativity is the *sine qua non* at the core of copyrightability, even as that human creativity is channeled through new tools or into new media." Both the Copyright Office and District Court explain that the history and language of the Copyright Act, Supreme Court precedent, and the Copyright Office Compendium support the position that only human authorship qualifies for copyright protection.[198]

[197] Thaler v. Perlmutter, No. 22-1564, 2023 U.S. Dist. LEXIS 145823, at *13 (D.C. Cir. 2023).

[198] *See e.g.*, Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53, 58 (1884) (limiting copyright law to protecting only the creations of human authors); Mazer v. Stein, 347 U.S. 201, 214 (1954) (holding that a work "must be original, that is, the author's tangible expression of his ideas"); Goldstein v. California, 412 U.S. 546 (1973) (defining "author" as "an 'originator'" and "he to whom anything owes its origin").

Infringement

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

Yes, AI-generated outputs may implicate the exclusive rights of reproduction and the derivative work right. Below are examples of scenarios where output of AI systems could be infringing.

- *Overfitting*: In some instances, AI tools exhibit a machine learning flaw known as overfitting where the output of the system closely matches its training data.[199] Overfitting can occur for many reasons, including training the AI model for too long on a limited amount of ingested material or when the ingested material contains large amounts of irrelevant information (also known as "noisy data").[200] One result of overfitting is that the AI model's output will sometimes closely resemble a work in the set of material it ingests, which could violate a number of a copyright owner's exclusive rights in a work.

- *Prompting for Copyright-Protected Material*: Users of AI image generators can enter prompts that include the names of copyright protected characters or works, and the resulting output might include protected elements of those works. For example, prompting an image generator with a request for an image of a popular superhero or cartoon character would likely result in output that would infringe the copyright in the character. Similarly, prompting an LLM for a specific copyrighted work, like song lyrics or a poem, would likely result in AI responding with the copyrighted lyrics or poem.

[199] *What is Overfitting?*, IBM, https://www.ibm.com/topics/overfitting (last visited Oct. 19, 2023).

[200] *What Is Overfitting?*, AMAZON WEB SERVS., https://aws.amazon.com/what-is/overfitting (last visited Oct. 19, 2023).

- *Style Prompts*: Copyright does not protect style.[201] That said, when a user prompts an AI system to generate new material "in the style of" a particular artist, there is a risk that the output will be substantially similar to a particular ingested work by that artist. In determining whether infringement has occurred, the ultimate issue will remain whether the defendant copied protectable elements of the plaintiff's work, not whether it merely imitated the plaintiff's "style."[202]

In addition to ensuring that their use of copyrighted materials for AI ingestion was done lawfully/with authorization, it is essential that AI companies implement effective safeguards to ensure that these and other types of output-related infringements do not occur. Requiring that secondary users implement effective safeguards to prevent the likelihood of infringement is not a new concept. For example, in *Authors Guild v. Google*, as part of its fair use analysis, the court extensively discussed the need for Google to implement measures to prevent the likelihood of infringement from the output of the Google Book system and from a user of the system who might manipulate the system.[203]

Importantly, when copyrighted works are ingested pursuant to a license, the parties can negotiate what these safeguards will be and how they should work. This is yet another reason that licensing of copyrighted works by AI companies is so crucial and is generally superior to other options.

The last point we'd like to make in response to this question relates to the scope of the derivative-work right (commonly also referred to as the adaptation right). Some have argued that

---

[201] *See generally* 2 Patry on Copyright § 4:14.

[202] Steinberg v. Columbia Pictures Industries, Inc., 663 F. Supp. 706, 713–14 (S.D.N.Y. 1987) (finding that a movie poster copied expressive elements of an artist's style and that the similarity between the works were not based on unprotectable scènes à faire).

[203] *See* Authors Guild v. Google, Inc., 804 F.3d 202, 227–28 (2d Cir. 2015) ("Plaintiffs argue that Google's storage of its digitized copies of Plaintiffs' books exposes them to the risk that hackers might gain access and make the books widely available, thus destroying the value of their copyrights . . . Google's prudent acknowledgment that 'security breaches could expose [it] to a risk of loss ... due to the actions of outside parties, employee error, malfeasance, or otherwise,' however, falls far short of rebutting Google's demonstration of the effective measures it takes to guard against piratical hacking.").

in order for there to be an infringement of the derivative-work right the alleged infringer must have made a contribution of new original expression to the alleged infringing work. That is not correct, as it erroneously imports the standard for copyrightability in a derivative work into a determination of whether there is an infringing derivative. For the derivative-work right to be infringed there is no requirement that: (i) there be a copyrightable contribution of new material added to or changed within the alleged infringing material, or that (ii) any new material must be by a human.

The definition of derivative work in section 101 of the Copyright Act provides:

> A "derivative work" is a work based upon one or more preexisting works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which a work may be recast, transformed, or adapted.[204]

There is no requirement within the definition that the derivative work contains a copyrightable contribution.

The Copyright Office seems to agree. In its recent comments to the American Law Institute the Copyright Office states:

> "The Office believes that the test for copyrightability and the test for infringement of the derivative-works right are distinct. With respect to the former, copyright only extends to "original works of authorship," and thus only the products of human creativity are eligible for copyright protection. In contrast, the derivative-works right is framed in terms of 'preparation,' indicating that non-human actions may be sufficient to infringe the right."[205]

---

[204] 17 U.S.C § 101.

[205] Letter from Suzanne V. Wilson, Gen. Couns. & Assoc. Reg. of Copyrights, U.S. Copyright Off., on Preliminary Draft No. 9 of the A.L.I.'s Restatement of the Law of Copyright, to A.L.I. 2 (Sept. 26, 2023), https://www.copyright.gov/rulings-filings/restatement/comments/2023-09-26-Preliminary-Draft-No-9.pdf.

Taking a position that a derivative work must have a contribution of new original expression and that such contribution must be made by a human, would result in a situation in which no AI-generated material would ever qualify as an infringing derivative of a work ingested by the AI system. That position would create a huge loophole in the law. For example, if this were correct, when a human creates an animated motion picture version of a book without permission of the book's copyright owner that would clearly be a violation of the copyright owner's derivative-work right, but when a human simply prompts an AI tool to produce the animated version that would not result in a violation of the derivative-work right. That inconsistent and unfair result cannot be the law. The better rule—and one that is supported by the law—is that when an act violates the derivative work right when performed by a human, that same act should be a violation when it is performed by an AI tool. We discuss liability issues more in response to question 25 below.

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?

Substantial similarity is the existing test under the copyright law and, at this time, should be sufficient to address claims of copyright infringement based on the output from AI systems. However, we will be monitoring the case law in this area to determine whether any modifications to traditional notions of substantial similarity may be warranted. While this question is about output, we feel obliged to reiterate that, when there is evidence of copying complete works—as in the case of ingestion/input—there is no need to resort to a consideration of whether the output is substantially similar to the ingested work.

We take no position on, and our answer above should not impact in any way, considerations relating to the need for or appropriateness of standards that might be imposed under a potential *sui generis* law.

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?

Existing civil discovery rules are insufficient to address this situation. For starters, existing discovery rules may not be adequate or efficient to prove the elements of copying when such records are not maintained by the AI developer. Moreover, litigation in federal court is far too expensive and time consuming to be a practical option for most individual creators.[206] In addition, sophisticated litigants can easily weaponize the discovery process in federal court against parties with fewer resources to drive up the cost, prolong the process, and make litigation untenable.

Another problem with this approach is that because copyright liability is joint and several, there can be more than one direct infringer, each involved in a different stage of the development and/or use of the generative AI model. Obtaining the requisite information from each of them may prove to be daunting.

As noted in other areas of our response, we support calls for appropriate transparency and record keeping. Transparency and the costs associated with it are necessary to promote ethical AI systems and should be borne by AI developers. Creators should not be required to subsidize the costs of doing business incurred by AI developers. Please also see our responses to questions 15 and 15.1, in which we detail the records that AI developers should be required to collect, retain, and disclose to ensure adequate transparency.

[206] And even where it is a viable option as in the case with infringements brought to the Copyright Claims Board (CCB), the CCB's limited discovery process would likely be unsuitable for uncovering the elements necessary to prove infringement.

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?

Like all questions regarding infringement liability, the answer to this question will be fact dependent. Because copyright liability is joint and several, there can be more than one direct infringer, and each may be involved in a different stage of the development and/or use of the generative AI model. In addition, under existing theories of direct and secondary copyright infringement, liability could attach to one or more of the relevant actors identified in the question depending on the specific facts at issue.[207] Thus, copyright owners should be able to seek remedies from any and all parties that play a role in the infringement related to the development and use of generative AI models. This includes collectors and curators of datasets, model developers, developers of systems that incorporate the model, the end user of the system, and any other party that facilitates and/or contributes to infringement or benefits from the use of the model.

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

Open-source AI models do not raise unique considerations with respect to infringement based on their outputs.[208] Open-source AI models need to be treated the same as non-open-source model. Therefore, we would likely oppose any laws or policies that would exempt open-source AI models from the same legal or regulatory obligations imposed on non-open-source AI models.

---

[207] Of note, some companies working in AI development, including Microsoft and Adobe, have announced plans to indemnify certain end users against copyright infringement claims arising from the output generated by their generative AI models when certain conditions are met. *See e.g.*, Brad Smith & Hossein Nowbar, *Microsoft Announces New Copilot Copyright Commitment for Customers*, MICROSOFT (Sept. 7, 2023), https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/; Stephen Nellis, *Adobe Pushes Firefly AI Into Big Business, with Financial Cover*, REUTERS (June 8, 2023, 3:37 PM), https://www.reuters.com/technology/adobe-pushes-firefly-ai-into-big-business-with-financial-cover-2023-06-08/. Significantly, the indemnity does not cover infringing inputs.

[208] We recognize that the issue of open-source models is becoming central to the debate surrounding transparency obligations for foundation models under the European Union's AI Act. The concerns surrounding open-source AI models were also raised during the Senate Judiciary Committee, Subcommittee on Privacy, Technology, and the Law, hearing on *Oversight of A.I.: Principles for Regulation,* during which witnesses testified to the dangers of opening up the operation of generative AI technologies to bad actors. *See Oversight of A.I.: Principles for Regulation: Hearing Before the Subcomm. on Priv., Tech., & the L. of the S. Comm. on the Judiciary*, 117th Cong. (2023), https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-principles-for-regulation.

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?

Ever since its enactment as part of the Digital Millennium Copyright Act (DMCA), section 1202 has been a very important, but relatively infrequently used, part of the Copyright Act. With the advent of AI technologies, section 1202 has taken on even greater importance to the copyright ecosystem. This is because copyright management information (CMI) plays a crucial role in ensuring that AI technologies are adopted, implemented, and used in a manner that is responsible, ethical, and respectful. CMI is essential to infringement evaluations as well as to transparency and labeling, which we discuss in greater detail in those respective sections.

With regard to infringement, CMI may be used to determine whether a particular work has been ingested, and stripping the CMI makes recordkeeping much more difficult. When CMI is intact and not altered or removed, it would be possible to automate the task of generating records of use. Section 1202(a) and (b) apply both to the ingestion of copyrighted materials that may be stripped of copyright management information (CMI) and the generation and distribution of potentially infringing output that contains altered or false CMI (or from which CMI has been removed). Claims of violations of section 1202 are raised in several of the class actions suits as well as the Getty lawsuits against Stability AI, such as *Anderson v. Stability AI*, *Doe v GitHub*, *Tremblay v. OpenAI, Inc., Silverman v Open AI, Inc.* and *Getty Images v. Stability AI*. A detailed summary of these cases can be found in Appendix A to these comments. In particular, we highlight the Getty cases in which claims of section 1201(a) violations related to modification of watermark that provides false CMI and claims of section 1202(b) violations related to removal of metadata and watermarks are both raised.

It is important to recognize that in the digital age, section 1202 violations can take many forms related to the falsification, removal, or alteration of CMI that compromises attribution and integrity of copyright-protected works. These problems, which have been largely related to the distribution of works over the internet since the enactment of the DMCA, are now compounded

by AI developers that alter or remove CMI to "prepare" works for ingestion. Further, when a generative AI model's output reproduces copyrighted material (or is a derivative of an ingested work), the removal, alteration, or falsification of CMI would rob copyright owners of attribution, as well as another way to prove copying. In enacting 1202, Congress recognized that CMI is essential to "establishing an efficient Internet marketplace" by tracking and monitoring copyright uses and facilitating licensing agreements.[209] When AI developers violate section 1202, it makes already burdensome monitoring practices that much more difficult for copyright owners, and it hinders the development of market-based licenses (discussed throughout our responses).

As we note in response to question five, we believe that section 1202 should be amended to only require that a copyright owner prove that the information was removed or altered knowingly or recklessly, not that the copyright management information (CMI) was removed or altered with the knowledge that it would induce, enable, facilitate, or conceal infringement. Additionally, because metadata and CMI are often stripped from uploaded works by online service providers in order to reduce file sizes and decrease transmission and storage costs during the caching process, we support changes to the statute that would clarify that failing to return CMI to a work following its removal for resizing or storing purposes is a violation of 1202(b). In the AI context, these amendments are crucial to ensure that metadata is maintained that can be used to determine whether a work has been ingested by an AI system, and possibly to indicate the provenance of derivative works containing both AI and human-authored elements. Finally, as we discuss in response to question 28, to the extent any material is labeled as AI-generated, that information should fall within the definition of "copyright management information" in section 1202. This is so that someone who falsely labels something as AI-generated output (when that work is not in fact AI-generated) in order to induce, enable, facilitate, or conceal infringement of the work can potentially be held liable for such actions under section 1202.

[209] U.S. COPYRIGHT OFF., AUTHORS, ATTRIBUTION AND INTEGRITY: EXAMINING MORAL RIGHTS IN THE UNITED STATES 84 (2019), https://www.copyright.gov/policy/moralrights/full-report.pdf (citing S. REP. NO. 105-190, at 16 (1998)).

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

At this time, we do not have any additional issues to raise with regard to copyright liability relating to AI-generated output.

LABELING OR IDENTIFICATION

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

Questions regarding whether, to what extent, and how to impose requirements to label or otherwise publicly identify AI-generated outputs are not directly related to copyright law. However, such requirements may have copyright-related implications. Therefore, while this is not a question we can answer outright, as it goes beyond the scope of copyright, any disclosure requirements should take into account the impact on copyright law and copyright owners, as well as principles of free expression, and remember that a "one size fits all" approach may not be feasible.

To the extent someone falsely labels a human-created copyrighted work as AI-generated output in order to induce, enable, facilitate, or conceal infringement of that work, that information should fall within the definition of "copyright management information" in 17 U.S.C. 1202 and that person should potentially be liable for such actions under section 1202.

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

To the extent that labeling is warranted (see our answer above), it seems sensible to require the party best situated to know that the material is AI generated and best situated to appropriately identify the work as such to be responsible for complying with the labeling requirement. In many cases, this will be the person who generated the output (or their employer if the person is acting within the scope of their employment). In addition, AI developers should be able to implement a

persistent function within their AI models that would automatically label output as being AI-generated when the output was generated using their AI models. In such case, users should be prohibited from removing or otherwise obstructing the use such labels or labeling tools. (See answers to questions 28 and its subparts.)[210]

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

No comment.

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?

We are reluctant to comment on potential consequences until specific labeling requirements have been proposed, as the appropriateness of any consequences for the failure to label a particular work or the removal of a label (e.g., fines or suspension of AI operating licenses granted by government) must take into account the nature of the requirements themselves. Therefore, we will reserve comment until there are specific proposals, except to say that any consequences must not in any way impact or threaten a rightsholder's ability to retain copyright protection in their works, brings actions to enforce their copyrights, or otherwise run afoul of any international treaty obligations of the United States.

[210] As noted in or response to question 28, any labeling requirements should take into account the impact on copyright law and copyright owners, as well as principles of free expression, and a "one size fits all" approach may not be feasible or appropriate. Thus, there may be circumstances where removing or obstructing the use of a label, should not be prohibited. As stated previously in our responses, approaches to AI will vary industry by industry and will need to be tailored to the type of industry and type of copyrighted work.

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?

Tools, such as ChatGPT Zero are already used to identify AI-generated material, and IPTC Digital Source Type,[211] Coalition for Content Provenance and Authenticity (C2PA),[212] Content Authenticity Initiative (CAI),[213] and Project Origin,[214] are in the process of being developed by a wide group of stakeholders.

ADDITIONAL QUESTIONS ABOUT ISSUES RELATED TO COPYRIGHT

We are not answering questions 30-33 because they do not deal with copyright issues and therefore fall outside the scope of the Copyright Alliance mission.

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

There are three issues we would like to raise that were not directly referenced in the first 33 questions: (i) the Copyright Office guidance on the registration of works that contain AI-generated elements;[215] (ii) the need for bulk registration of dynamic web content so that press

[211] The IPTC develops and promotes "efficient technical standards to improve the management and exchange of information between content providers, intermediaries and consumers." *About IPTC*, INT'L PRESS TELECOMMS. COUNCIL, https://iptc.org/about-iptc/ (last visited Oct. 25, 2023).

[212] As noted on the C2PA website, "The Coalition for Content Provenance and Authenticity (C2PA) addresses the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of media content. C2PA is a Joint Development Foundation project, formed through an alliance between Adobe, Arm, Intel, Microsoft and Truepic." COAL. FOR CONTENT PROVENANCE & AUTHENTICITY, https://c2pa.org/ (last visited Oct. 25, 2023).

[213] The goal of the CAI is fight prevent misinformation by adding "a layer of verifiable trust" to all types of digital creativities through provenance and attribution solutions. *How It Works*, CONTENT AUTHENTICITY INITIATIVE, https://contentauthenticity.org/how-it-works (last visited Oct. 25, 2023).

[214] The objective of Project Origin is to create a process "where the provenance and technical integrity of content can be confirmed [by] [e]stablishing a chain of trust from the publisher to the consumer." PROJECT ORIGIN, https://www.originproject.info/ (last visited Oct. 25, 2023).

[215] Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence (Mar. 16, 2023). https://www.copyright.gov/ai/ai_policy_guidance.pdf?loclr=eanco.

publishers can protect against generative AI-related infringement, and (iii) countermeasures, like Glaze, to combat unauthorized ingestion.

Comments on the Copyright Office Registration Guidance for Works Containing AI-Generated Elements

Earlier this year, the Copyright Office issued guidance on the registration of works that contain AI-generated elements titled *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence*.[216] In the guidance, the Office explains that applicants have a duty to disclose the inclusion of AI-generated content in a work submitted for registration and to provide an explanation of the human author's contributions to the work.[217] Other notable requirements are that for AI-generated content, registrants must use the standard application, and in a situation where registration has already been granted (but AI-generated content was not disclosed), the applicant should correct the public record by submitting a supplementary registration.[218]

We appreciate the Copyright Office's effort to provide much-needed guidance on the complex issues surrounding the copyrightability of works that contain AI generated elements, but there remain many unanswered questions and some confusion on how the standards set forth in the guidance will be applied in practice. In particular, we believe it is not a good use of Copyright Office resources to engage in investigations into the boundaries of what is disclaimed as AI-generated and whether there is sufficient human involvement in each case. Nor should the Office make inquiries into whether there are AI-generated elements in a work when there is no indication of such by the applicant on the registration form. The Copyright Office, as it does for disclaimed pre-existing works incorporated in a new work, should at most merely require the

[216] *See generally id*.

[217] *Id*. at 16193.

[218] *Id*.

applicant to generally disclose that the work incorporates materials wholly generated by AI and identify the nature of that material in the registration application.[219]

The guidance also includes the requirement that registrants must use the standard application when registering a work with AI-generated content, which raises the following concerns for creators and copyright owners:

- It prohibits the use of group registrations and the benefits that flow from them, making it more challenging and economically infeasible for certain creators to register their works with the Office.[220] We ask that the Copyright Office amend its guidance to permit group registrations in this context.

- When a registrant used a form other than the standard form in the past to register the AI-assisted work but now needs to go back and revise their registration to disclaim AI-generated content, there are many unanswered questions surrounding how they would do so and what the consequences would be. Specifically, it's unclear what effect a change in forms would have on the effective date of registration if a group registration was broken up into many standard forms. It would be helpful to have further guidance here.

- There is also confusion amongst many in the copyright community about how material generated in part using AI should be disclosed in a registration application. The guidance applies obligations to disclose AI-generated material included in works without drawing a clear line around what those are. There was a recent webinar that

[219] *See id*.

[220] Many individual creators are not policy or legal experts and may fail to realize that works with AI elements cannot be registered in a group registration application. This means that if they unknowingly choose the group registration option, they are inevitably set up for failure as they will be unaware of disclosure requirements which leads to sunk costs of time and resources spent in the registration process in addition to the Office's invalidation of the registration application.

sought to clarify some things, but also raised additional issues.[221] It would be helpful to have further written guidance and clarification of these registration issues.

- There are a number of inconsistencies between the guidance and parts of the Copyright Office Compendium on registration guidelines that must be clarified. One example is that the Compendium says that unclaimable material should be disclaimed when it represents an "appreciable portion" of the whole work, whereas the guidance says that AI-generated content that is more than *de minimis* should be explicitly excluded from the application. These are two different standards that must be reconciled. In the webinar, the Office attempted to define what it meant by *de minimis* and described how it compared to "appreciable amount," but in the process raised additional questions, which we look forward to working with the Office on clarifying. Another example is that when a work contains AI-generated material, both the Compendium and the guidance indicate that there are three fields in the registration application that need to be completed. However, while the three fields referenced are part of the online application, only two are part of the standard paper form. It is unclear if that means that the online application must exclusively be used to register works with AI-generated content.

- There is concern amongst many in the copyright community about retroactive application of the Copyright Office's guidance. For creators and organizations with a vast portfolio of registrations, the threat of invalidation or cancellation is a major concern, especially when the guidance on where to draw the line regarding what to disclaim is unclear. Specific concerns include: (i) whether and, if so, how the U.S. Copyright Office will go back and revoke applications that did not accurately disclose AI use; (ii) whether the new guidance will be misused by overly aggressive litigators to challenge the validity of every copyright registration if they believe AI was used even slightly and was not disclosed—in turn this might make litigation more expensive; and (iii) the cost of registration is expensive for many individual artists,

[221] *Webinar: Registration Guidance for Works Containing AI-Generated Content*, U.S. COPYRIGHT OFF. (June 28, 2023), https://www.copyright.gov/events/ai-application-process/.

and the confusion of registering works that incorporate AI created by the guidance will be discouraging for artists and become a barrier to registration.

Finally, we recognize that the Office's registration division has its own internal policies and procedures that govern its registration practices to a more finely tuned degree. Because registration of works containing AI generated material is such a novel issue, we encourage the Office to increase its transparency regarding relevant internal policies and procedures. We look forward to working with the Office on answering these questions and updates to the guidance and Compendium that the Copyright Office has indicated will be made in the future. The copyright community looks forward to the opportunity to review and comment on those updates before they take effect.

Registration of Dynamic Web Content

As we have explained previously, we urge the Office to update its registration practices to allow for registration of dynamic web content.[222] It is essential for the effective enforcement of press publishers' rights that the Copyright Office's registration practices keep pace with market realities and new industry business practices and improve the process for registering dynamic website content. This is even more critical in the generative AI space, where, without an efficient system to register dynamic web content, press publishers are unable to register and therefore unable to enforce their copyrights against aggregators and others AI developers who take their content.

Countermeasures

In response to AI developers ingesting copyrighted works without requesting permission or securing a license, some copyright owners and creators have begun to use technological

[222] *See* Copyright Alliance, Additional Comments Submitted in Response to U.S. Copyright Office's Nov. 9, 2021, Notice of Inquiry at 10–12 (Jan. 5, 2022), https://copyrightalliance.org/wp-content/uploads/2022/01/Copyright-Alliances-Additional-Comments-on-Publishers-Protections-Study.pdf ("The Office must update its registration practices to allow for bulk registration of dynamic web content.").

countermeasures to fight back. The most well-known and effective of these countermeasures are Glaze and Nightshade, developed by researchers at the University of Chicago. As described on the Glaze website:[223]

> Glaze works by understanding the AI models that are training on human art, and using machine learning algorithms, computing a set of minimal changes to artworks, such that it appears unchanged to human eyes, but appears to AI models like a dramatically different art style. For example, human eyes might find a *glazed* charcoal portrait with a realism style to be unchanged, but an AI model might see the glazed version as a modern abstract style, a la Jackson Pollock. So when someone then prompts the model to generate art mimicking the charcoal artist, they will get something quite different from what they expected.

Glaze developers claim that it is effective because it cannot be circumvented[224] "because it is not a watermark or hidden message (steganography), and it is not brittle." While Glaze presently works for works of visual art, in the future it may also work for other types of copyrighted works.

Closely related to Glaze is another countermeasure, referred to as a "data poisoning tool" called Nightshade. Nightshade will be integrated into Glaze in the future. Nightshade works by "poisoning samples that are incorporated into a model's dataset and cause it to malfunction." [225] Glaze and Nightshade are just two countermeasures and other technological solutions to help individual creators fight back against AI systems who ingest their works without permission.[226]

---

[223] *What Is Glaze?*, UNIV. OF CHI. GLAZE, https://glaze.cs.uchicago.edu/what-is-glaze.html (last visited October 27, 2023).

[224] Glaze's effects cannot be circumvented by such actions as taking a screenshot/photo of the art, cropping the art, filtering for noise/artifacts, reformatting/resizing/resampling the image, compressing the image, smoothing out the pixels, adding noise to break the pattern. *See id*.

[225] Melissa Heikkilä, *This New Data Poisoning Tool Lets Artists Fight Back Against Generative AI*, MIT TECH. REV. (Oct. 23, 2023), https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/.

[226] Ruixiang Tang et al., *Did You Train on My Dataset? Towards Public Dataset Protection with Clean-Label Backdoor Watermarking*, CORNELL UNIV. ARXIV (Apr. 10, 2023), https://arxiv.org/pdf/2303.11470.pdf.

It's important to recognize that these tools can have a significant negative effect on AI development. In very basic terms, where Nightshade is used on an image of a cat, a human will see a cat, but the AI model may see a giraffe. "Tricking" AI models in this way is harmful to those who use AI systems, as well as AI development and the public more generally, because it provokes the spread of misinformation and threatens the integrity of AI models. We support creators and copyright owners who use these tools because these tools enable them to protect themselves against AI developers who choose not to license their works. But it is our hope that, in the near future, creators and copyright owners will not have to resort to using these tools and other countermeasure tactics because AI developers have chosen to act responsibly, ethically, and respectfully by licensing their works instead.

Conclusion

We would once again like to thank the Copyright Office for the comprehensive Notice of Inquiry and the overall thought and attention it has given to the intersection of generative AI and copyright. We appreciate the opportunity to submit these comments and look forward to working with the Office and other stakeholders on these issues in the future.

Respectfully submitted,

Keith Kupferschmid
CEO Copyright Alliance
1331 F Street, NW, Suite 950
Washington, D.C. 20004

October 30, 2023

Appendix A

AI LITIGATION SUMMARY

COPYRIGHTABILITY

***Thaler v. Perlmutter* (D. DC)**

In early 2022, the U.S. Copyright Office Review Board affirmed a denial of registration for a two-dimensional artwork "authored" by an AI algorithm called the "Creativity Machine," explaining that the registrant, Steven Thaler, failed to show requisite human authorship in the work and that the work could not qualify as a work-made-for-hire. Thaler then filed a complaint in the District Court for the District of Washington, DC, alleging that the Office's denial of Thaler's registration application was an arbitrary and capricious agency action. On August 18, 2023, the District Court for the District of Columbia issued an opinion granting the U.S. Copyright Office's motion for summary judgment, explaining that "defendants are correct that human authorship is an essential part of a valid copyright claim" and "a bedrock requirement of copyright." The court also denied Thaler's claim that the Copyright Office's refusal to register the work was "arbitrary and capricious"—and therefore a violation of the Administrative Procedures Act (APA)—finding that "the Register did not err in denying the copyright registration application." On October 11, Thaler filed a notice of appeal.

INFRINGEMENT

***Anderson v. Stability AI* (N.D. Ca) – Class Action**

On January 13, 2023, artists Sarah Andersen, Kelly McKernan, and Karla Ortiz filed a class-action lawsuit against Stability AI, Midjourney, and DeviantArt in the Northern District of California, alleging copyright infringement and right of publicity violations for the use of the

plaintiffs' works in training data sets for the AI image-generating platforms Stable Diffusion, the Midjourney Product, DreamStudio, and DreamUp. On July 19, 2023, the court held a hearing on motions to dismiss by Stability AI, Midjourney, and DeviantArt, during which Plaintiffs' counsel conceded that two of the named plaintiffs have not registered the copyright in their works. Judge William Orrick expressed skepticism that each of the defendants' products incorporated plaintiffs' works in their entirety and said the plaintiffs are unlikely to succeed on their secondary liability claims, noting that he did not believe that a "claim regarding output images is plausible at the moment, because there's no substantial similarity" between images created by the artists and the AI systems and that more was needed to clarify the differences in the infringement claims against the various defendants. Judge Orrick indicated that he would dismiss most of the claims due to these concerns, but that the plaintiffs "can take comfort in the leave to amend."

***Chabon et al v. OpenAI, Inc. et al* (N.D.Ca.)**

On September 8, 2023, a group of authors, including Michael Chabon, filed a class action lawsuit against OpenAI in the district court for the Northern District of California, alleging that the company used the authors' books without authorization to train ChatGPT. The complaint alleges that ChatGPT itself is an infringing derivative work and states that when prompted, ChatGPT provides extremely detailed summaries, examples, and descriptions of the authors' works, and that the authors' writing styles can be accurately imitated. The plaintiffs are suing for copyright infringement and removal of copyright management information, as well as state-related claims including unfair competition and negligence. On September 12, 2023, the same group of plaintiffs filed a similar lawsuit against Meta.

***Concord Music Group, Inc. v. Anthropic PBC* (M.D.TN)**

On October 18, 2023, music publishers Universal Music Publishing Group, Concord Music Group, and ABKCO, filed a lawsuit in the district court for the Middle District of Tennessee against AI company, Anthropic, alleging direct, contributory, and vicarious copyright infringement claims as well as copyright management information removal claims. The plaintiffs allege that Anthropic unlawfully copied and distributed plaintiffs' musical works, including lyrics, to develop Anthropic's generative AI chatbot, Claude. The plaintiffs allege that when prompted, Claude generates output that copies the publishers' lyrics. The complaint also alleges

that 500 works have been infringed and request statutory damages—resulting in a total damage award demand of $75 million for copyright infringement.

Doe v. GitHub **(N.D. Ca.) – Class Action**

On November 3, 2022, a group of GitHub programmers filed a class action lawsuit against Microsoft and Open AI for allegedly violating their open source licenses and scraping their code to train Microsoft's AI tool, *GitHub Copilot*. On January 26, 2023, Microsoft, GitHub, and Open AI filed a motion to dismiss the case, arguing that the complaint "fails on two intrinsic defects: lack of injury and lack of an otherwise viable claim." On May 11, 2023, the district court issued an order granting in part and denying in part defendants' motions to dismiss. The court allowed plaintiffs to amend claims related to the violation of Section 1202(a) of the DMCA, tortious interference, fraud, false designation of origin, and violation of the California Consumer Privacy Act (CCPA). Plaintiffs filed a first amended complaint on July 21, 2023, which included examples of alleged verbatim copying of plaintiffs' code. Defendants filed renewed motions to dismiss on August 10, 2023.

Getty Images v. Stability AI **(D. De.) &** ***Getty Images v. Stability AI*** **(UK)**

In early 2023, Getty Images filed a copyright and trademark infringement suit against Stability AI in the U.S. District Court for the District of Delaware, along with a lawsuit against Stability AI in the High Court of Justice in London. The U.S. complaint alleges that Stability AI "copied *more than 12 million* photographs from Getty Images' collection, along with the associated captions and metadata, without permission from or compensation to Getty Images, as part of its efforts to build a competing business." In addition to willful and intentional copyright infringement claims, Getty alleges that Stability AI removed or altered copyright management information (CMI), provided false copyright management information, and infringed Getty Images' trademarks. On May 2, 2023, Stability AI filed a motion to dismiss arguing that the court lacks personal jurisdiction over Stability UK; Stability UK is a necessary and indispensable party; and by "lumping [together] allegations" against Stability U.S. and Stability UK under the collective designation "Stability AI," the complaint fails to identify which defendant is responsible for the alleged infringing acts and therefore fails to state a claim. In the alternative, Stability moves to transfer the case to the Northern District of California.

***Huckabee et al v. Meta Platforms, Inc. et al* (S.D.N.Y)**

On October 17, 2023, a group of authors including former Arkansas governor, Mike Huckabee, and best-selling Christian author Lysa TerKeurst filed a class-action lawsuit in the district court for the Southern District of New York against Meta, Microsoft, EleutherAI, and Bloomberg for direct and vicarious copyright infringement, removal of copyright management information, and various other state-law claims. The plaintiffs allege that the defendants infringed by using plaintiffs' books to develop defendants' large language AI models (LLMs) using the "Books3" training dataset. The lawsuit also alleges that AI research company, EleutherAI, is liable for copyright infringement for hosting and distributing "The Pile" dataset, which includes Books3.

***J.L. v. Alphabet Inc.* (N.D. Ca.)**

On July 11, 2023, a group of anonymous plaintiffs filed a class-action lawsuit against Google for the use of personal information and various copyrighted works to train its AI models. The plaintiffs allege privacy law violations, violation of California's unfair competition law, other state law violations, direct and contributory copyright infringement, and DMCA violations. The complaint alleges that Google's large language model, Bard, generates summaries of copyrighted books or output that reproduces verbatim excerpts from copyrighted books. In addition to damages, the plaintiffs are requesting an injunction compelling the establishment of an independent AI council to monitor and oversee Google AI products and the destruction and purging of class members' Personal Information, which includes copyrighted works and creative content. On October 16, 2023, Alphabet filed a motion to dismiss, claiming that plaintiff fails to plausibly alleged that Bard is an infringing derivative work, that any of Bard's output is substantially similar to plaintiff's works, or that there was intentional removal of specified CMI from copies of plaintiff's book.

***Planner 5D v Facebook (Meta)* (N.D. Ca.)**

In 2019, UAB Planner 5D brought a complaint for copyright infringement and trade secret misappropriations against Facebook, Inc., Facebook Technologies, LLC, and The Trustees of Princeton University. Planner 5D, a Lithuanian company that operates a home design website that allows users to create virtual interior design scenes using a library of virtual objects, alleged

that Facebook and Princeton copied its database of objects and scenes for the commercial potential of scene recognition AI technology. On February 17, 2023, Meta (formerly Facebook) filed a motion for summary judgment, arguing that because the Copyright Office correctly concluded that the works at issue are "data files" and not "computer programs," Planner 5D failed to satisfy the pre-suit application-and-deposit requirement of Section 411(a). Meta also argues that Planner 5D's insistence that the works are computer programs "in the face of both controlling case law and administrative guidance to the contrary allowed it to circumvent mandatory administrative review and gave it a litigation advantage," and that if the court endorses such an approach, it would create an "untenable loophole in the registration system." On April 6, 2023, Planner 5D filed an opposition to Meta's motion for summary judgment, arguing that its "object and scenes" works qualify as computer programs and are copyrightable.

***Silverman v. OpenAI* (N.D. Ca.) & *Silverman v. Meta* (N.D. Ca.)**

On July 7, 2023, Sarah Silverman, Christopher Golden, and Richard Kadrey filed a class-action lawsuit against OpenAI (and a separate suit against Meta) in the district court for the Northern District of California, accusing the AI developers of copyright infringement related to the unauthorized use of plaintiffs' books to train the proprietary large language models (LLMs) ChatGPT and LLaMA. The complaints allege that OpenAI and Meta harvested mass quantities of literary works through illegal online "shadow libraries" and made copies of plaintiffs' works during the training process. Also included are DMCA claims for the removal of copyright management information under section 1202(b), as well as claims for unfair competition, negligence, and unjust enrichment. On August 28, 2023, OpenAI filed a motion to dismiss. The motion addresses both the Silverman and Tremblay complaints (see summary below) and was filed concurrently on both dockets.

***Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.* (D. Del.)**

In 2020, Thomson-Reuters sued Ross Intelligence, which is a competitor legal research service, for copyright infringement, alleging that Ross obtained copyrighted legal content from a Westlaw subscriber to develop its own competing product based on machine learning. The claims allege that an AI bot systematically mined, collected, and downloading content from the Westlaw database. In March 2022, Thomson-Reuters survived a motion to dismiss by Ross. However,

Ross then brough anti-trust counterclaims alleging that Thomson-Reuters uses anticompetitive practices to maintain its monopoly over the legal research market. In early 2023, the parties filed cross motions for summary judgment, with Thomson-Reuters arguing that Ross has not shown that its creation of a competing database qualifies as fair use. On September 25, 2023, a memorandum opinion was issued, largely denying both motions for summary judgment. The opinion explains that there is still a genuine factual dispute over the copyrightability of Westlaw's headnotes, and that although Ross actually copied portions of bulk memos, the question of substantial similarity must be decided by a jury. The opinion also says while Ross's first sale and first amendment defenses fail, all of Thomson Reuters' theories of infringement liability and Ross's fair use defense must be decided by a jury.

***Tremblay et al v. OpenAI, Inc. et al.* (N.D. Ca.)**

On June 28, 2023, two authors of literary works—representing a proposed class of plaintiffs—filed a lawsuit in the U.S. District Court for the Northern District of California against OpenAI. The complaint accuses the AI developer of copyright infringement related to the unauthorized use of plaintiffs' works to train its proprietary large language model (LLM), ChatGPT. The complaint alleges that OpenAI harvested mass quantities of literary works through illegal online "shadow libraries" and made copies of plaintiffs' works during the training process. Also included are DMCA claims for the removal of copyright management information under section 1202(b), as well as claims for unfair competition, negligence, and unjust enrichment. On August 28, OpenAI filed a motion to dismiss the "ancillary claims" of vicarious infringement, violation of the DMCA, unfair competition, negligence, and unjust enrichment. The motion does not respond to the direct infringement claim, which OpenAI says it "will seek to resolve as a matter of law at a later stage of the case."