Before the
UNITED STATES COPYRIGHT OFFICE

Library of Congress

Notice of Inquiry

Artificial Intelligence and Copyright:

Docket No. 2023-6

Notice and Request for Public Comment

REPLY COMMENTS OF ANONYMOUS

The rapid growth of generative AI technology has provided great surprises and possibilities for the creative industry, as well as raised significant questions about copyright legal systems. The Notice of Inquiry published by the Copyright Office covers a great deal of important questions under heated discussions between different stakeholders.

We notice the major rights holders in the fields of music, image, and literature have shared their valuable viewpoints based on their position in the industry, which includes the Universal Music Group ("UMG"), American Society of Composers, Recording Industry Association of America, Inc. (“RIAA”), Authors and Publishers (“ASCAP”), National Music Publishers Association ("NMPA"), Broadcast Music, Inc. ("BMI"), International Confederation of Societies of Authors and Composers ("CISAC"), Getty Images, Authors Guild, and New York Times, etc.. We would like to participate in an anonymous form and share our opinions, particularly in reply to some comments from the aforementioned major rights holders.

General Questions

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

copyright owners, technology developers, researchers, and the public?

We notice that comments from major rights holders all recognize the benefits of generative AI, while still holding serious concern about the risk that generated outputs will compete with and displace human creations, and even finally destroy the creative industry. We view this issue from a different perspective.

In the context of generative AI technology, copyright holders and AI technology developers are not antagonistic to each other. Generative AI's positive and significant contribution to the content creation industry is great because it lowers the threshold of creation, improves efficiency, and reduces costs. Generative AI allows creators to invest their energy in more creative, meaningful work while focusing less on simple, repetitive work. For example, AI generated content (AIGC) technology has been widely used in the game industry to carry out repetitive creations, which has helped significantly to improve production efficiency and reduce production costs.

At a high level, the dynamic balance between technological innovation and content protection must be considered. With reference to the era when Internet technology was just emerging, "safe harbor" rules played a great role in encouraging technological development, and only through development could a winning balance between technological innovation and content protection be achieved.

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?

In reply to comments by UMG and RIAA that "US should align with EU regulations", "Japan and Singapore set an overly broad exception for TDM in their copyright act", we hold different opinions. When looking into the detailed provisions under Japan and Singapore copyright laws, we noticed they are not merely providing the "fair use" exemptions for data training, but also setting out comprehensive requirements to invoke such exemptions.

Article 30-4 of Japanese Copyright Act provides that data analysis of copyrighted content can constitute fair use and exploitation of works is permitted "in any way and to the extent considered necessary" in prescribed cases, unless "the action would unreasonably prejudice the interests of the copyright owner in light of the nature or purpose of the work or the circumstances of its exploitation."

Section 244(1)-(3) of Singapore Copyright Act expressly recognizes data analysis as fair use and details on limitations. Specifically, use is allowed when the purpose of the use is limited to data analysis, the copy will not be supplied to others, access to the data is lawful, and the first copy is not an infringing copy under this provision.

Article 30-4 of Japanese Copyright Act	Sections 244 (1)-(3) of Singapore Copyright Act
Article 30-4 (Exploitation without the Purpose of Enjoying the Thoughts or Sentiments Expressed in a Work) *It is permissible to exploit a work, **in any way and to the extent considered necessary**, in any of the following cases, or in any other case in which it is not a person's purpose to personally enjoy or cause another person to enjoy the thoughts or sentiments expressed in that work; provided, however, **that this does not apply if the action would unreasonably prejudice the interests of the copyright owner in light of the nature or purpose of the work or the circumstances of its exploitation**:* *...* *B. **if it is done for use in data analysis (meaning the extraction, comparison, classification, or other statistical analysis of the constituent language, sounds, images, or other elemental data from a large number of works or a large volume of other such data; the same***	***Copying or communicating for computational data analysis*** *244.—(1) If the conditions in subsection (2) are met, it is a permitted use for a person (X) to make a copy of any of the following material:* *(a) a work;* *(b) a recording of a protected performance.* *(2) The conditions are —* *(a) the copy is made for the purpose of —* ***(i) computational data analysis; or*** ***(ii) preparing the work or recording for computational data analysis;*** *(b) X does not use the copy for any other purpose;* *(c) X does **not supply (whether by communication or otherwise) the copy to any person** other than for the purpose of —* *(i) verifying the results of the computational data analysis carried out by X; or* *(ii) collaborative research or study relating to the purpose of the computational data analysis carried out by X;* *(d) X has **lawful access to the material** (called in this section the first copy) from which the copy is made; and* *(e) one of the following conditions is met:* *(i) the **first copy is not an infringing copy;*** *(ii) the first copy is an infringing copy but —* *(A) X does not know this; and* *(B) if the first copy is obtained from a flagrantly infringing online location (whether or not the location is subject to an access disabling order under section 325) — X **does not know and could not reasonably have known** that;* *(iii) the first copy is an infringing copy but —* *(A) the use of infringing copies is **necessary for a prescribed purpose**; and*

applies in Article 47-5, paragraph (1), item (ii));	*(B) X does not use the copy to carry out computational data analysis for any other purpose.*

We would suggest avoiding statutory or regulatory approaches similar to the Article 4 of EU Copyright Directive 2019 which provides that member states must provide an exception for “reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining”, which may only be obtained for as long as necessary for the purposes of text and data mining. This exception, although appearing widely applicable, is narrowed down by providing that rightsholders can expressly reserve their rights, thus will not be practical to follow and likely result in situation that no text or data mining can take place.

Article 4 Exception or limitation for text and data mining

1. *Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining. 17.5.2019 EN Official Journal of the European Union L 130/113*
2. *Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining.*
3. *The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph* ***has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.***

In light of the above, as already flagged out by some major rights holders, the uniformity of global compliance requirements is crucial, to the further development of the creative industry and the AI technology. Firstly, AI model training requires a vast amount of data from a wide range of sources, which often involves multiple countries and regions. Secondly, the final model can be accessed in multiple products to serve users across the world, which will result in cross-jurisdictional issues prompting an array of legal considerations accordingly.

Additionally, if the US places limits on the use of training data that would make it cost prohibitive, in reality, companies and its servers will have to move offshore to countries which have laws that are more friendly for the use of training data. Just as tax liabilities drive corporate restructuring to countries that are more tax-friendly than the US, some companies may do the same with AI laws.

Training

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.

Many major rights holders, including UMG and ASCAP, assume that training materials are necessarily retained by developers of AI models. Based on our understanding of generative AI technology, it's possible to delete training data after completion of the training process. However, the training data may still need to be retained for some reasonable consideration.

After training is completed, the knowledge and parameters obtained by the AI model have been determined. Saving the parameter is more important, and the AI model becomes less dependent on the original data that had been used in the training.

However, in practice, the original training data will not be deleted, considering that such original training data is crucial for training when the AI model is iterated. Especially, for large language AI models, the amount of training data needed can be huge. If not retained after the training is completed, the training data will need to be re-collected again during the iteration, which leads to significant economic and time costs, and would be inefficient and uneconomical.

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.

In reply to major rights holders' comments that model training will always infringe on the exclusive rights of copyright owners, we have different opinions and would like to provide the following detailed training processes to elaborate:

- downloading and storing the data to local server;
- cleaning and filtering the data to filter out data of poor quality;
- marking/tagging the data manually or by machine; and
- reading the data in the official data training process.

The only exclusive right of the right holder implicated under the data training process as specified above is the reproduction of the materials. There is not any involvement of adaptation, performance, creation of derivative works, or communication to the public of the works.

Moreover, all the above processes, even though may constitute reproduction of the materials, are simulating the process of human’s knowledge learning, which is filtering, classifying, refining, memorizing of knowledge, information, ideas, facts, and styles that not always qualify for copyright protection. The factual impact on the rights and interests of copyright owners, especially in copyright context, could be rather limited.

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?

In reply to Authors Guild's comment that "*Unlearning inferences may be possible*”, at the current technical level, there is no effective way to "unlearn" inferences gained from training without retraining.

Retraining is technically possible, but it comes with significant time and economic costs. Retraining a large language model can take several months (the time of which may be sufficient already for ChatGPT to complete iteration of a new AI model) and entail a large amount of machine and labor costs.

For security and compliance purposes, it may be necessary to "unlearn" some "unacceptable" data. However, this means should only be adopted for specific models or data having significant impacts on state security or public safety. It is not recommended to initiate such an uneconomical remedy for copyright infringement, as it would result in great uncertainty and additional costs to AI model developers and therefore hinder further industry development.

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.

We notice that many comments from major rights holders consider data training does not qualify for fair use, because there's possibility that AI-generated output could compete directly with human-created content. However, we don't agree with such assumptions and reasoning, which deliberately confuse generative AI technology with the particular output it generated, and ignore the transformative nature of the training process itself.

AI model training should be considered fair use because the use of the original work during AI model training is highly transformative. Different from the entertainment and appreciation purposes provided by the original work, the use of copyrighted content in the process of AI model training is a non-expressive and non-communicating use, which is only aimed at helping the computer program to learn the underlying ideas, methods, facts, and patterns, enable the computer program to discover new knowledge from the training data, and finally enable the program to generate new content. The use of copyrighted content in the process of AI model training is not intended to duplicate content contained in the training data.

The following cases and the analysis of "transformative use" are relevant to this question.

***Authors Guild, Inc. v. Google Inc.*, No. 13-4829-cv (2d Cir. Oct. 16, 2015)**

In Authors Guild v. Google, Google was accused of infringing the copyright of over 20 million books by scanning, digitizing and allowing the books to be searchable through Google's tools. The 2(nd) Circuit found that Google's copying was transformative. The Court reasoned that Google's product "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." The Court further alluded that even if such works were used by libraries in an infringing way, that does not make Google a contributory infringer of those works.

One can point toward the transformative aspect of using training data for machine learning models. The use of copyrighted works for training data is to provide usable models for use in AI technology. It does not supplant the original copyrighted work or even create a derivative of the work. Any portion of the original work is also used with numerous other data points to create the generative work. It is highly unlikely that any generative work would be substantially similar to a copyrighted work. Holding that use of the underlying works as training data augments public knowledge by providing AI models informed by collective human knowledge and experience, all of which are necessary to create realistic models. Further, just because a work made by the training data could be used in an infringing way, does not mean that the AI tool provider is a contributory infringer of those works.

Academic discussions around fair use in generative AI focus chiefly on the Google Books case, at least with respect to how to address the first fair use factor (i.e., the "nature and use" of the work, and the level of transformation of the original work). In that sense, this case is likely going to be determinative, if not highly influential, in rendering a judicial opinion.

***Kelly v. Arriba Soft Corp.*, 336 F.3d 811 (9th Cir. 2003)**

In Kelly, the defendant had a tool that crawled the Internet for images and generated thumbnail-sized images based on those images and displayed those in search engine results. The defendant copied 35 of the plaintiff's photographs and transformed them into thumbnail images for its search engine. The Court found that the copies were transformative and served an entirely different function than the underlying works. The thumbnail copies were "unrelated to any aesthetic purpose" and instead used for indexing the Internet and images hosted on the Internet.

By analogy, use of copyrighted material in training data for machine learning models is necessary as they are part of the collective human knowledge and experience. While the data is not used to index the knowledge of the world, it is used to index the knowledge of the world to make it accessible and easier to synthesize.

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?

We disagree with the opinion, as implied in the comments from major rights holders, that the *Andy Warhol Foundation v. Goldsmith* case can be seen as a precedence that generative AI does not qualify for fair use. We note that *Andy Warhol Foundation v. Goldsmith* may appear similar to an AIGC scenario, but it is in fact quite different. We agree with the Supreme Court's comments on transformative use: "the first fair use factor considers whether the use of a copyrighted work has a further purpose or different character."

In this Case, the Court found that there was no fair use defense in Andy Warhol's licensing of a Warhol work that incorporates Goldsmith's photograph of Prince. The court placed a large emphasis on the fact that the licensing of that image to Conde Nast was commercial in nature, and directly competed with Goldsmith's business in licensing their images to publications. In addition to Andy Warhol Foundation's image on the magazine cover, the magazine contained numerous concert and studio photographs of Prince. In that context, the purpose of the image is substantially the same as that of Goldsmith's photograph. Both are portraits of Prince used in magazines to illustrate stories about Prince.

The Court in this case placed a greater emphasis on the "purpose and nature" of the infringement than in past cases. The Court also mentioned that there could be a different fair use judgement for other works of Warhol, such as his Soup Can Series. "The purpose of Campbell's logo is to advertise soup. Warhol's canvases do not share that purpose. Rather, the Soup Cans series uses Campbell's copyrighted work for an artistic commentary on consumerism, a purpose that is orthogonal to advertising soup. The use, therefore, does not supersede the objects of the advertising logo."

The holding is helpful to demonstrate that use of copyrighted material for purposes of training data should constitute fair use. The purpose of the use of the training data is not to sell or market works that compete with the underlying works. From the perspective of the AIGC industry, both the abovementioned types of derivative works could be generated. That being said, developers of AI models train the AI models only to provide the computer program with the ability to generate content, rather than intended to generate one particular derivative work of one particular original work. The content generated could be either similar to the not-transformative-enough "Orange Prince" or to the transformative-enough "Soup Can Series". The specific content generated and whether or not the purpose of use is consistent with the original work, all depend on the users who use the AI model. Therefore, it cannot be arbitrarily concluded that AI model training does not constitute fair use simply because there are possibilities that user-generated content may have an alternative effect on the original work.

It is necessary to analyze and consider issues in separate stages. The pre-training stage uses more data and the data is used in a rougher manner. The data trained at this stage has a more limited impact on the final generation, and compared with the use at fine-tuning stage, the data training at this stage has a higher degree of transformation.

9.3. What legal, technical, or practical obstacles are there to establishing or using such a[n opt out] process? Given the volume of works used in training, is it feasible to get consent in advance from copyright owners?

We partially agree with comments from some major rights holders that "*an opt-out system does not work*". As mentioned above, the scale of data required for data training is huge. As such, relevant legal implications must be examined for the reasons laid out below. Firstly, copyright is not obtained on the basis of the application, and there are difficulties and uncertainties in identifying the right holders of copyrighted works. Secondly, if AI model developers are required to confirm whether the right holders have made an opt-out statement, it would be difficult to define by law what form of opt-out statement is valid and how to publish it.

If the right holder makes a new opt-out statement during the training process or after the training is completed, the knowledge of the AI model would have been already fixed by then as explained above. It would be difficult and technically impractical to "unlearn" such training data.

The above highlights the legal, technical and practical obstacles relating to right holders making opt-out statements.

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

In reply to comments of UMG, NMPA and RIAA that "*Licensing is the only lawful and practical means to permit the use of musical content for purposes of generative AI training*" and "*copyright owners' consent through voluntary marketplace agreements is required*", we believe that there are still some impediments for voluntary licensing:

- Training materials are of huge volume, various types, and their rights holders are geographically dispersed.
- Transaction costs of the materials could be huge, and transactions with each of the right holders could be inefficient.
- It is difficult for the AI model developers to obtain consent by legitimate transaction means in practice.

In reply to comments of ASCAP and Authors Guild that "*Similar collective licensing arrangements [for technology that has disrupted the music sector] can be negotiated and executed for generative AI training*" and "*Collective licensing is an effective means of providing licenses*", we believe there are still some difficulties that need to be considered in generative AI training:

- An extremely strong supporting system, which can cover a wide range of types of works, is needed. However, effective collective management organizations ("CMOs") do not exist for every type of work in the market, and even those existing CMOs for music, literature, and photographic works are relatively decentralized.

- The related regulations need to be clarified, including the pricing standard, royalty collecting method, royalty distribution mechanism, as well as the transparency of CMOs. For example, under compulsory licensing, the pricing standard should be reasonable. The reason being that the required volume of data is huge and if the pricing standard is too high, it will be unaffordable for AI model developers.

Generative AI Outputs

20. 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?

In reply to major rights holders' comments that "*there is no policy advantage to affording copyright protection to AI-generated content*" and "*Broad legal protections for purely AI-generated works is contrary to the purpose of copyright law*", we hold different opinions.

Legal protection for AI-generated material is necessary, even if it is not at the same level as human created works. Copyright protection for computer software only covers the code of the AI model, which could protect the investment of the AI model developer but still cannot protect the content generated by or via AI models. Failing to protect AI-generated content will result in the following consequences:

- AI-generated content is identical to human created works in terms of the form of expression, and it is difficult to make a distinction in the marketplace. A big difference in the attribution of rights will have an adverse effect on content distribution and the stability of licensing transactions.
- From a fair competition perspective, if AI-generated material is not protected and can be used free of charge while human created works need to be authorized and paid, human created works would become less appealing from an economic perspective. The market share of human created works will be squeezed. This would also discourage the creation and development of human created works in the longer term.

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?

We agree with the comments that end users who prompt an AI system to generate an infringing output should be liable, when AI-generated material is found to be infringing on copyrighted work. While we cannot agree with the comments that developers of the AI model and the system should also be liable.

- Whether AI-generated material constitutes infringement would largely depend on the specific use scenario of the material. After specific material is generated by AI, AI model developers' control over the material is very limited. As such, it is reasonable for the party who uses AI-generated material to take responsibility for any infringement.
- In some AI models, the user could apply the AI model to generate content in accordance with his/her own expectations by inputting prompts or other more specific information. Thus, the model developer's involvement in and control over the content generation process is even more limited.

Additional Questions About Issues Related to Copyright

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

We disagree with UMG's comment that "*style is a diffuse similarity in artistic approach*". From a policy perspective, copyright law does not protect style or genre alone since it would frustrate the very purpose of copyright law – to foster creativity. We hold the opinion that the basic principle of idea-expression distinction under the copyright law should not be breached. Style itself is not protected by copyright law. If the interests of the original author are affected, protection can only be sought by considering other aspects of legal interests, instead copyright infringement.

Dave Grossman Designs, Inc. v. Bortin., 347 F. Supp. 1150 (N.D. Ill. 1972). This seminal case on this issue was filed by the copyright owner of a series of statues of children, and she alleged that certain aspects of these statues formulated a style. The defendant copied certain elements of the plaintiff's work, including the style, coloring, packaging and (in our words) trade dress. The court stated, "The law of copyright is clear that only specific expressions of an idea may be copyrighted, that other parties may copy that idea, but that other parties may not copy that specific expression of the idea or portions thereof. For example, Picasso may be entitled to a copyright on his portrait of three women painted in his Cubist motif. Any artist, however, may paint a picture of any subject in the Cubist motif, including a portrait of three women, and not violate Picasso's copyright so long as the second artist does not *substantially copy* Picasso's *1157 specific expression of his idea.

By analogy, the use of training data to create works using the style of one artist or several does not constitute infringement unless the generated work is substantially similar to the copyrighted work. Most AI models take only a tiny fraction of that work, and transform it with various other works to inform how it creates a work (image, text, sound). Therefore, the likelihood of that a generated work is substantially similar to an underlying copyrighted work is extremely low.