Response to: U.S. Copyright Office, Library of Congress, Notice of Inquiry: Artificial Intelligence and Copyright [Docket No. 2023–6]

On behalf of: The Program on Information Justice and Intellectual Property (PIJIP), American University Washington College of Law
https://www.wcl.american.edu/impact/initiatives-programs/pijip/

Introduction

Generative AI raises many new questions about creative practices and the copyright law that applies to them. In this response we attempt to address those issues. However, we would also like to note our effort to address the factual basis for understanding how copyrighted works are used in training generative artificial intelligence models here: *A Copyright-Relevant Primer on Generative AI* https://www.youtube.com/watch?v=2nARmcWZfKE.

The Program on Information Justice and Intellectual Property (PIJIP) is the internationally-recognized intellectual property and information law research and academic program of American University Washington College of Law. The faculty consists of experts in copyright law and technology law, including Professor Michael Carroll, Professor Peter Jaszi, Professor Charles Duan, Professor Christine Farley, and Professor Sean Flynn. PIJIP also works extensively with communities of practice in creative industries, research, and education to document, draft, and implement Best Practices in Fair Use in work conducted by Professor Jaszi and Meredith Jacob.

As part of this scholarship, PIJIP organized a two-day public conference on Copyright and Generative AI, in September 2023. Among other sources, these comments reflect insights from the scholarly presentations and discussion over those two days.

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 potential benefits of this technology include opening up access to authorship of many types to new creators and increasing the benefits of authorship to some existing creators. We are at the very early stages of using these tools and should focus on understanding and exploring new uses. Copyright law has been adaptable to other new technologies, including photography, computer software, and the internet. We should move cautiously and gather more experience with, and perspective on, the impact of generative AI tools for existing creators, new creators, students, educators, and researchers.

While generative AI may be disruptive to some creative communities, it is important to understand the impact more fully instead of trying to freeze in place pre-AI practices and understandings. Historically, new technologies, such as photography, the internet, and digital publishing have increased access to authorship, as well as public access to cultural and knowledge goods.

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?

Within the education and research sector, there are some distinct benefits to access to generative AI tools, including text and data-mining based research techniques, development of educational supports and tools, translation of educational resources, and new flexibility for students with disabilities and neurodiverse students.

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 our research on copyright and AI, we found the following resources useful:

Vox, *AI art, explained*
https://www.youtube.com/watch?v=SVcsDDABEkM

Computerphile, *How AI Image Generators Work (Stable Diffusion / Dall-E) - Computerphile*
https://www.youtube.com/watch?v=1CIpzeNxIhU

Google, *Overview of GAN Structure*
https://developers.google.com/machine-learning/gan

Google Cloud Tech, *Transformers, explained: Understand the model behind GPT, BERT, and T5*
https://www.youtube.com/watch?v=SZorAJ4I-sA

[add transformer articles]
https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

5. Is new legislation warranted to address copyright or related issues with generative AI? If so, what should it entail? Specific proposals and legislative text are not necessary, but the Office welcomes any proposals or text for review. Training If your comment applies only to a specific subset of AI technologies, please make that clear.

It is far too early in the development of these technologies to contemplate new legislation. The technologies are in a period of rapid development. At this point, there is not clear evidence that new legislation is needed to address copyright issues with generative AI. Existing copyright law has proved

adaptable for new technologies in the past, and we should proceed with the assumption that it will be adaptable for this situation.

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

According to public sources, ChatGPT 3.5 was trained on

- Common Crawl – a non-profit that has been making datasets available by copying everything it can find on the Internet once a month since 2008
- https://commoncrawl.org/
- WebText 2 – the filtered text from all web pages referred to from all posts on Reddit that received 3 or more "upvotes" https://openwebtext2.readthedocs.io/en/latest/
- Books 1 –
- and Books 2 – web-based corpora of books -- questions of provenance are in litigation
- Wikipedia

The model underlying Bard uses the following sources:

According to a research paper on LaMDA, the training data is:

- 12.5% C4-based data – (filtered from Common Crawl)
- 12.5% English language Wikipedia
- 12.5% code documents from programming Q&A websites, tutorials, and others
- 6.25% English web documents
- 6.25% Non-English web documents
- 50% dialogs data from public forums

Source: https://www.searchenginejournal.com/google-bard-training-data/478941/#close

Even less is known about Meta's Llama 2 model
Some data available about the original Llama model

- CommonCrawl 67.0% 1.10 3.3 TB
- C4 15.0% 1.06 783 GB
- Github 4.5% 0.64 328 GB
- Wikipedia 4.5% 2.45 83 GB
- Books 4.5% 2.23 85 GB – subject of current litigation
- ArXiv 2.5% 1.06 92 GB
- StackExchange 2.0% 1.03 78 GB

Source: https://www.businessinsider.com/meta-llama-2-data-train-ai-models-2023-7

Image-generating AIs primarily rely on LAION – a non-profit that provides large datasets that pair images with associated text. These are derived from the Common Crawl

LAION-5B, released in March 2022 has 5.85 billion image-text pairs https://laion.ai/

7. To the extent that it informs your views, please briefly describe your personal knowledge of the process by which AI models are trained.

Our full response is in the presentation from our conference Mapping Copyright's Application to Generative Artificial Intelligence, September 29, 2023, American University Washington College of Law, https://www.wcl.american.edu/impact/initiatives-programs/pijip/events/mapping-copyrights-application-to-generative-artificial-intelligence/

The Office is particularly interested in:

Our full response to these questions is here, *A Copyright-Relevant Primer on Generative AI* https://www.youtube.com/watch?v=2nARmcWZfKE

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 core technology used by generative artificial intelligence is artificial neural networks.

The architectural model for such a network is built first, without reference to any training data. The network processes numeric values to produce predicted outcomes. Initially, these values are entirely random.

Training data is copied for purposes of analyzing statistically relevant relationships in the **overall** corpus of text, images, music or other in the training dataset. The goal of this use of the training data is to assess the whole, not individual works within that whole.

These mathematical patterns serve as a reference, or the "rule book", to assess how well the generative AI model responds to prompts when predicting a useful output. The copyrighted works are not "in" the model.

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

The use of the word "inference" in this question is ambiguous and potentially founded on an erroneous factual premise. In law, we might infer from one fact an additional fact. If the term is used in that manner, to suggest that the model infers values from specific works used in training, it is misplaced. Instead, the training process of repeated trial-and-error predictions is akin to linear regression, in which one analyzes a data set and ends ups with two numbers that define a line that provides information

about the data. The line is not attributable to any single piece of input there. It's the result of understanding a pattern out of the whole thing.

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?

This is a subject of active research in the computer science field, but in general, changing the training dataset would require retraining but would not alter a model's ability to predict the same outputs that it predicts with that particular work in the training data.

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?

In general, no. If there are repeated instances of a work of authorship in the training data, then the model can become "overfitted" to that data and produce outputs that appear to copy that material.

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.

As discussed above, most of what gets "extracted" from the training data is factual, consisting of information about grammatical relationships between words. This statistical data is often uncopyrighted factual information.

Further, unauthorized use of copyrighted works to train AI models is a transformative use that is fair use. The cases cited and much of the analysis of fair use and text and data mining applies equally to fair use and generative AI training. *See* Michael W. Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful*, 53 UC Davis Law Review 893 (2019); *see also id.* at 938, recognizing the connection. The article is available here, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3531231

The Supreme Court's opinion in Google LLC v. Oracle America, Inc., 141 S. Ct. 1183 (2021) is in full accord and provides additional support for the analysis. The Court's opinion in Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, 598 U.S. ___ (2023) further supports this conclusion. In Part II.A. of its opinion, in which the Court discussed transformative use, its citations with approval to Authors Guild v. Google, Inc., 804 F.3d 202, (2d Cir. 2015) further reinforce the foundational role that transformative use analysis plays in fair use analysis.

In the popular press, many characterizations of how copyrighted works are used in training data fundamentally misunderstand the use. Broadly, in these models, copyrighted works are used to train prediction models at the core of generative AI by acting as reference texts and images during the process of model training, rather than being ingested and stored within the tool itself. This process happens at a

very atomized level, dealing with tiny discrete chunks of a work, rather than at the level they are normally created or perceived.

This use of these works is transformative, using them for the data about the structure of the text, rather than for their original expressive purpose. As such, our initial analysis is that the use of copyrighted works for AI model training is fair use.

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.

In general, no, for the reasons given in our response to the section 7 questions.

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

The creation of a licensing requirement at this early stage would limit research, not-for-profit uses, and would lock in advantages for large commercial actors, who can negotiate licensing agreements.