Ken Oshima

To Whom at the U.S. Copyright Office This May Concern,

I am submitting the following response as an individual visual illustration hobbyist and open source software developer whose work has been copied to data sets without my consent to train generative AI models:

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?

Considerations of the benefits and risks of generative AI/ML software should be grounded in the fundamental foundation of such systems; the (currently non-consensual) exploitation and derivative amalgamation of existing works, which applies to systems like diffusion-based image generators[1] and large language models (LLMs). Some promote the misleading narrative that generative models display human-level understanding and therefore ought to be treated as legal persons; they are more akin to oft-referenced "stochastic parrots"[2] that output a reflection of their training data sets, with no grounding to any sort of reasoning. Benefits and risks, as well as their consequences to creators and IP holders, therefore should be considered in their use as tools. I do not see any benefit to the technology as long as they are currently infringing on the rights of creators and copyright holders. Risks include said infringement and economic harms as a result, as well as reduced barrier of entry to the creation of misinformation that could contribute to the pollution of the information ecosystem of the internet.

1 The diffusion technique was originally demonstrated to reconstruct an input image. *See https://arxiv.org/abs/1503.03585*

2 *https://dl.acm.org/doi/10.1145/3442188.3445922*

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.

Please see: Jiang et al., *AI Art and its Impact on Artists*, AIES '23 (Aug 29, 2023), *https://dl.acm.org/doi/abs/10.1145/3600211.3604681*

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?

The U.S. Copyright Office should consider the already established internationally-consistent approaches enacted by the Berne Convention, in turn mandated by the TRIPS Agreement, in particular Article 13 of the TRIPS Agreement, where exceptions to exclusive rights "do not conflict with a normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the right holder."[3]

3 *ttps://www.wto.org/english/docs_e/legal_e/31bis_trips_01_e.htm*

5. Is new legislation warranted to address copyright or related issues with generative AI? If so, what should it entail? Specific proposals and legislative text are not necessary, but the Office welcomes any proposals or text for review.

The existing copyright regime is robust enough to generally not warrant additional legislation; judgments rendered to the Thaler and Kashtanova submissions are satisfactory, in my opinion. However, generative AI models do pose new challenges that should be addressed (and are currently being considered in the European Union's AI Act and China's draft AI legislation). One such legislative improvement would be item-level disclosure of training data, which is absolutely imperative if licensing schemes are to be pursued.

Training

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

Copyright-protected training materials include (but not all mentioned here):

- Images/Visual Art: Data set creators including LAION indiscriminately scrape images from the Internet, including copyrighted expressions, without first seeking copyright holders' permission or licensing. This has prompted copyright holders like Getty to pursue legal action[4] against Generative AI companies utilizing these datasets, for reasons including removing or altering copyright management information (CMI).
- Text: Practically all data sets utilized by popular LLMs use web-scale data sets, that are bolstered with data sets containing hundreds of thousands of in-copyright books. These data sets include ones such as Books3 aka "The Pile"[5]. This has prompted several lawsuits against generative AI companies including OpenAI, Meta, Anthropic, etc by known authors.

4 https://www.courtlistener.com/docket/66788385/getty-images-us-inc-v-stability-ai-inc/
5 https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/

- Software code: Software code has been scraped for use in tools like Github Copilot. In the case of Github Copilot, the code has been scraped from Github's own repositories. While software code can be copyrighted, software developers can and do choose to license their code under more permissive terms like GPL. That said, such terms can still require attribution or inclusion of CMI; in the case of Copilot, CMI is alleged to have been stripped when copied for training, according to the plaintiffs of DOE 1 v. GitHub, Inc.[6]

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?

In the case of image generation software, there are only two companies that offer services that "license" works from copyright owners in some way: Getty and Adobe, both of which source from their image stock service to train their generative AI services. However, the licensing of these images is dubious; in the case of Adobe, the company maintains a non-exclusive license over submitted stock images[7], which allows the owner to retain ownership of their copyright[8]. Prior rulings has indicated that existing licenses do not confer additional rights to novel uses[9]; Adobe has not sought additional licensing from copyright holders to permit use of Adobe Stock images for use for generative AI training[10]. This may also be the case for Getty's image generation service.

6 https://www.courtlistener.com/docket/65669506/doe-1-v-github-inc/
7 https://contributor.stock.adobe.com/en
8 https://copyrightalliance.org/education/copyright-law-explained/copyright-transfers/exclusive-vs-non-exclusive-licenses/
9 https://openjurist.org/845/f2d/851/cohen-v-paramount-pictures-corp
10 https://petapixel.com/2023/06/26/photographers-upset-that-adobe-firefly-ai-is-competing-against-them/

6.3. To what extent is non-copyrighted material (such as public domain works) used for AI training? Alternatively, to what extent is training material created or commissioned by developers of AI models?

In the case of image generation software, it is alleged by the plaintiffs of Andersen v. Stability AI Ltd. that Stability AI commissioned LAION, a separate non-profit organization based in Germany, to produce a data set of 5.85 billion web-scraped images known as LAION-5B, which would serve as the training data for the Stable Diffusion image generation software[11]. It should be noted that images in the public domain may only number in the tens of millions[12]. Other companies like Getty and Adobe, may instead "license" images from stock image contributors, but without notifying them or seeking additional licensing, as described in the response to question 6.2.

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.

Most generative AI developers are opaque as to the training process of their models, including OpenAI and Meta. Open-source models provide the best insights into the usage of training materials. The most prominent example of this is Books3, part of the larger "Pile" data set. Books3 was created by developer Shawn Presser in 2020 in response to the secret Books1 and Books2 data sets used by OpenAI to train its GPT models[13]. Books3 contained nearly 200,000 books, including those in copyright, in full plain-text format and was available on the Hugging Face website until only recently[14][15]. The data set has been removed on account of copyright infringement complaints.

11 https://www.courtlistener.com/docket/66732129/1/andersen-v-stability-ai-ltd/
12 https://en.wikipedia.org/wiki/Wikipedia:Public_domain_image_resources
13 https://interestingengineering.com/innovation/anti-piracy-group-shuts-down-books3-a-popular-dataset-for-ai-models
14 https://huggingface.co/datasets/the_pile_books3
15 https://huggingface.co/datasets/the_pile_books3/discussions/7#6523c2a98fc48849920808af

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?

The "purpose and character" of the use of copyrighted works to train an AI model needs to consider that the purpose includes commercial ones. The services and software provided by companies like Stability AI and Midjourney confer full ownership over the assets created using their software, with no restrictions as to the type of usage[16]. The conflict in this first factor of fair use was reaffirmed to be in the copyright holder's favor in Andy Warhol Foundation v. Goldsmith if the purpose is to compete in the same market[17]. Even if the "transformative" aspect were to be considered, fair use could not be met because in all sorts of generative AI models, new expressions are not created; this is a view held by prominent AI scientists such as Yann LeCunn, Chief AI scientist at Meta, who states that models are "unable to invent new things" and "merely regurgitate stuff they've been trained on"[18]. Infringement ultimately occurs at the copying stage of the training process; the value of generative AI models rely on copied and trained works, and if no works were copied, then the model would simply not exist.

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?

According to 17 U.S. Code § 106, copyright holders maintain the *exclusive* right to reproduce and prepare derivatives[19]. Works have to be wholly copied to train AI models, which could preclude the "limited" usage permitted by fair use[20]. Training should not be permitted, if license or permission has been given to do so.

16 https://docs.midjourney.com/docs/terms-of-service
17 https://ipwatchdog.com/2023/05/18/dissent-scotus-ruling-warhol-foundation-fair-use-stifle-creativity/id=161004/
18 https://twitter.com/ylecun/status/1718263147591573949
19 https://www.law.cornell.edu/uscode/text/17/106
20 https://www.copyright.gov/help/faq/faq-fairuse.html

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?

If AI models or data sets are later adapted for use of a commercial nature, that suggests a shift in the "purpose and character" of the model, and as affirmed in Andy Warhol Foundation v. Goldsmith, should be considered as infringing.

8.4. What quantity of training materials do developers of generative AI models use for training? Does the volume of material used to train an AI model affect the fair use analysis? If so, how?

As noted previously, image generation software requires upwards of billions of images for training, such as LAION-5B, a data set of 5.85 web-scraped images. LLMs rely on data sets like Books3, which contain nearly 200,000 books, for the quality of its models. The volume of material required to train a model should not matter in a fair use analysis if the materials are wholly copied. The models requires the material to exist at all, whether it is trained on one book or hundreds of thousands.

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?

If the inquiry is focused on the outputs, then direct copyright infringement should be considered if prompted for a reproduction of the work, as was the case for Midjourney being able to reproduce the "Afghan Girl" photograph[21]. Infringement should also be considered if the AI media generated "in the style of" a particular artist had a causal connection by having training data referencing said artist. However, the inquiry should also consider the input stage during the phase in which a model is trained on wholly copied works. As previously noted, copyright owners hold the exclusive right to reproduction and creation of derivatives.

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)?

Copyright owners must have the right to affirmatively consent and opt in. Recent events have show that an opt out is not a workable approach; OpenAI recently provided an opt out form to enable artists to have their images opted out of being trained in their DALL-E generative models. The process is not workable, requiring individual submission of every single image in an artists' portfolio to be opted out[22]. The problem here is two-fold: First, the opt out process only applies to future models; DALL-E is now in its third widely-available iteration. Second, the opt out process only applies to OpenAI specifically. With multiple entities pursuing their own image generation technology, it is impossible for an individual to opt their work out of every single data set, if an opt out process is even offered at all.

21 https://twitter.com/Zn2plusC/status/1635382361235021824
22 https://www.businessinsider.com/openai-dalle-opt-out-process-artists-enraging-2023-9?op=1

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

Yes, the copyright owner should provide explicit consent for all uses of copyrighted works.

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?

An opt out basis, as previously explained, is likely unfeasible for the average person. Opt in should be the approach adopted, with measures taken towards transparency in data sets and models to ensure compliance.

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?

It should be feasible to get prior consent. For example, due to the GDPR, websites are required to request consent from visitors for their cookie policies. Also note that the basis of U.S. Copyright law provides the exclusive right of creatives to their respective works, and that system has proven to be robust despite the “volume” of works produced throughout the history of the U.S.

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?

Due to the difficulty and potential costliness of "untraining" AI[23], in the case infringement in a model has been found, it should be destroyed. This is not without precedent; the FTC has previously ordered algorithms to be destroyed[24]. Blocking inputs is not a viable option either, as "prompt injection" is becoming an increasingly viable way to circumvent superficial keyword blocks.

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?

Yes, the creator should retain the right to object to an AI model being trained on their work. The provisions of the Visual Artists Rights Act of 1990 extend the Berne moral rights to works of visual arts[25], and at least visual artists should be able to object to the mutilation and distortion of their work in being used to train models that can generate derivatives in their name. In general, I request that the Copyright Office base their rules in Berne moral rights.

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

Licenses should be obtained as they are obtained for any other kind of use under the current copyright regime.

23 https://venturebeat.com/ai/machine-unlearning-the-critical-art-of-teaching-ai-to-forget/
24 https://spectrum.ieee.org/ai-concerns-algorithmic-destruction
25 https://library.osu.edu/site/copyright/2017/07/21/moral-rights-in-the-united-states/

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

Yes, it is feasible in all sectors. As one might voluntarily agree to an online service's terms of service or cookie policies, it should be possible to have a developer provide a scalable way to directly obtain license from a copyright owner.

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?

Yes, a voluntary collective licensing scheme would also be a feasible approach. This approach has been done before by collective licensing societies like ASCAP and BMI, and has been endorsed by the Electronic Frontier Foundation[26].

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 compulsory licensing regime for AI training as it would undermine an artist's choice to refuse licensing if the use prejudices their moral rights.

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

No, for the same reason compulsory licensing regimes are not agreeable: a creator must have the right to choose how their works are appropriated.

26 https://www.eff.org/wp/better-way-forward-voluntary-collective-licensing-music-file-sharing

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

No, a licensing regime should be evenly applied.

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

Licensing revenue contribute a significant portion of the global economy. According to Licensing International, consumers spent approximately $315 billion worldwide on licensed merchandise, and royalty revenues measured $17.4 billion[27]. According to the U.S. National Endowment for the Arts and the Bureau of Economic Analysis, copyright-intensive arts and industries (including 47 million American jobs, according to the U.S. Department of Commerce) contributed $876.7 billion to the national GDP in 2020[28]. If creators are unable to secure a living from their works when supplanted by AI-generated works (that cannot exist without said creators' works as input), then creators will leave their industries and diminish the U.S.' cultural production as a whole. A licensing requirement would benefit both creators and the U.S.

27 https://www.forbes.com/sites/joanverdon/2022/07/26/licensing-still-a-big-business-during-the-pandemic-as-sales-top-315-billion/

28 https://www.uschamber.com/intellectual-property/five-ways-copyright-laws-encourage-personal-expression-and-creativity

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?

No, under current copyright law, there are no instances for when a human using a generative AI system should be considered the "author". Copyright explicitly provides authorship to *human* authors who produce fixed expressions. As the U.S. Copyright Office correctly ruled in their registration for Kristina Kashtanova's *Zarya of the Dawn*, since generative AI systems' "specific output cannot be predicted by users"[29], users of such systems should not be considered authors of the expressions produced. The Copyright Office also noted the important distinction between generative AI users and photographers in their mention of *Burrow-Giles*, that is, photographers can take specific choices (framing, shooting during a particular time, etc.) to ensure their own "original mental conception" which they can view before fixation.

29 https://www.copyright.gov/docs/zarya-of-the-dawn.pdf

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. Generative AI tools by nature produce whole, random expressions and can potentially remove human involvement almost entirely (e.g. automate prompt generation and feed into AI models), thus precluding them copyright exclusively provided to *human* authors. No additional standards are needed for content including AI-generated material; since generation of AI works is more akin to "discovery" than expression, they can fall into the category of unprotectable material like ideas or recipes, as described in U.S. Copyright Office Circular 33[30].

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?

No, legal protection for AI-generated material is not desirable as a policy matter. Exclusions or special rights should not be conferred specifically for AI-generated material. Legal protection is also not necessary to encourage development of generative AI systems. For instance, open source software thrives despite utilizing a wide range of permissive licenses, including MIT which relinquishes copyright protections except for the requirement of inclusion of the license notice in copies.

30 https://www.copyright.gov/circs/circ33.pdf

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?

No, the Copyright Clause does not permit copyright protection for AI-generated material. Rights are noted to be secured by “Authors and Inventors”, which implies a human being and has been clarified by the Copyright Office in Compendium of U.S. Copyright Office Practices § 313.2[31]. Copyrightable AI-generated material would likely not promote the progress of “useful arts” as generative AI systems can create material at a pace far beyond that of a human being, which could crowd out human authors out of markets. This would have the effect of destroying the pool of human-produced works that generative AI data sets rely on and also depriving the creative industries of original output, as briefly mentioned in the response to question 13.

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?

While certain sets of outputs specifically may not infringe on copyrighted works, direct infringement is certainly possible by requesting a reproduction of a highly represented work in a data set which has been overfitted, e.g. “Afghan Girl” in Midjourney. Causal connections suggesting infringement may also be found if requesting a work in the style of an artist and if that artist’s work is found in the training data set. Again, please consider that the inputs of a model’s training data set, without which the model would just not exist, suggest that all outputs are derivative of said inputs.

31 https://www.copyright.gov/comp3/chap300/ch300-copyrightable-authorship.pdf

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

The scale at which generative AI systems copy and appropriate works (often in the millions or billions) makes it difficult to apply the substantial similarity test to the remixed outputs. Instead, examination of the inputs should be the standard pursued when addressing claims of infringement.

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

Civil discovery rules may not be sufficient to enable availability of records of training material. In addition to the burden incurred by the pursuit of legal action, the process may not be expeditious enough to prevent significant economic harm to the copyright owners if infringement is found. I believe it is likely that many generative AI companies purposely obscure the training methods and data used to prevent civil lawsuits. Because of this, I think it is necessary to legislate federal AI data set transparency laws, including item-level data disclosures with preservation of original CMI.

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?

Developers of the generative AI model should be held directly liable for infringement when making unauthorized copies of works for use in training a model; the model for determining liability could be derived from the laws used for governing unauthorized copies of copyrighted sound recordings[32]. The developers of systems incorporating that model should have contributory liability if said developers do not perform their due diligence when ensuring the legality of models used. End users should be found liable for direct infringement if deliberately seeking to reproduce a copyrighted work or leverage an author's trained presence in a model by requesting a stylistically similar work.

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?

The nature of generative AI systems means that CMI cannot be reliably maintained and provided with the outputs produced by said system, all but ensuring infringement unless every single piece of data has been properly licensed. The plaintiffs of DOE 1 v. GitHub, Inc. allege that CMI removal at least constitutes DMCA violations, among other complaints[33]. In other instances, CMI may manifest itself as false reproductions in outputs, suggesting intent to induce, enable, facilitate, or conceal infringement, an argument made by the plaintiffs of Getty Images (US), Inc. v. Stability AI, Inc.[34] Without transparency and item-level disclosures of training data sets, it is impossible to ascertain the extent in which CMI has been removed or altered.

32 https://www.riaa.com/resources-learning/about-piracy/
33 https://www.courtlistener.com/docket/65669506/1/doe-1-v-github-inc/
34 https://www.courtlistener.com/docket/66788385/1/getty-images-us-inc-v-stability-ai-inc/

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?

While I cannot comment on the protection of artistic style, such questions could be resolved by considering the input stage in which works belonging to a particular author are copied to be trained on and then later reproduced or remixed by using the author as an identifier (i.e. requesting works “in the style of”). No additional rights or protections are needed if the exclusive right for an author to produce reproductions or derivatives is preserved.