I am a freelance author and professional software developer. This gives me a unique perspective to understand both the concerns of copyright holders as well as some technical details of how the AI models work (though that is not my expertise). These views are my own; I do not represent my employer or any other organization.

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?

Generative AI is an existential threat to all creatives—artists, authors, musicians, and narrators. Here are a few examples of how this technology is already having a tremendous negative impact on society:

Search Engines

Current search engines direct seekers to individual websites. Seekers can choose the sources that are credible and useful to them, and the original authors benefit from the traffic directed to their websites (in terms of credit, visibility, and revenue from ads or other services they may offer).

“Search” engines powered by Generative AI, on the other hand, harvest information from a multitude of sources—without respect to authenticity or copyright—and create composite answers. These composite answers are often riddled with bias and misinformation. They keep traffic within the “walled garden” of the search engine, a commercial enterprise.

Perhaps worst of all, the answers do not cite the sources; in fact they cannot. As ChatGPT will tell you:

“*I don't have access to specific sources, databases, or the internet to provide citations for my responses. My responses are generated based on a mixture of licensed data, data created by human trainers, and publicly available information up to my last training data in September 2021.”*

AI apps can even invent responses out of whole cloth—a process called “hallucination”. While sometimes hilariously nonsensical, these hallucinations can sometimes prove harmful, such as erroneously reporting that a public official has committed a crime or presenting a “recipe” with harmful ingredients. (https://techcrunch.com/2023/09/04/are-language-models-doomed-to-always-hallucinate/, https://arstechnica.com/information-technology/2023/08/ai-powered-grocery-bot-suggests-recipe-for-toxic-gas-poison-bread-sandwich/)

Books

Amazon’s self-publishing marketplace is already being flooded by low-quality books being “written” by Generative AI: https://www.techradar.com/computing/artificial-intelligence/amazon-has-a-big-problem-as-ai-generated-books-flood-kindle-unlimited Literary magazines have also been affected: https://www.nytimes.com/2023/02/23/technology/clarkesworld-submissions-ai-sci-fi.html

This is having a negative impact on authors, as the deluge of junk books is cutting into royalties (same pie - more slices) and displacing human authors from the bestseller lists. (The books aren’t legitimate best-sellers; bots are just artificially inflating the numbers.) It is undermining confidence in the entire self-published book market and making it harder for real people to make a living as an author.

Articles

Last November, the online magazine C-NET began publishing articles from Generative AI rather than human staff writers: https://www.wired.com/story/cnet-published-ai-generated-stories-then-its-staff-pushed-back/ Other online magazines have been accused of doing the same, threatening writer careers.

Covert Art

Publishers have begun to use AI-generated book covers rather than hiring human artists to design covers. These are not just cash-strapped indie authors who couldn’t afford a real cover designer, but big publishing houses putting out books for best-selling authors: https://www.themarysue.com/another-major-publisher-caught-using-ai-generated-cover-image-on-bestselling-authors-work/

This is not a one-time occurrence but a growing trend that will harm human artists.

Stock Images

Photographers and illustrators on stock image sites like Getty, 123RF, Adobe Stock, and more are threatened by Generative AI art tools. Magazines, blogs, etc. can just put a prompt into a tool like Midjourney to get “Man jogging” rather than paying a human artist. See Getty’s copyright infringement lawsuit for details: https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion

Other Concerns

The threat of studios using Generative AI in script writing was a core issue in the recent strike by the Writers’ Guild of America: https://time.com/6277158/writers-strike-ai-wga-screenwriting/

The Authors Guild has filed a class action lawsuit against Generative AI and written an open letter to AI developers citing copyright concerns: https://actionnetwork.org/petitions/authors-guild-open-letter-to-generative-ai-leaders

There’s also a growing trend to replace human narrators with Generative AI in audiobooks: https://www.washingtonpost.com/books/2023/08/17/audiobooks-artificial-intelligence/

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 an independent author and copyright holder, this technology has a direct impact on me. I have already seen cases where it has harvested copyrighted work from my website, mangled it into an error-riddled mess, and used it to answer questions about my work without credit or compensation.

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.

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?: https://dl.acm.org/doi/10.1145/3442188.3445922

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.

I feel that the existing copyright laws apply. Generative AI is a derivative work. While it may not outwardly resemble the original work, the original work is fundamental to the generated output. Without it, there would be no output. This is not dissimilar to a copyrighted piece of software used without permission in a new product's algorithm. You might not be able to see it outwardly in the app, but it's still there being leveraged illicitly under the hood.

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

- Web pages and articles

- Photographs and Illustrations
- Books and Book Covers
- Music
- Audiobook narration

Much of the material appears to have been obtained simply by scanning the open Internet in a process called “scraping”. Automated systems can effectively download an entire website, similar to how a search engine bot “crawls” the web to index search results.

The key difference is that search engines crawl the web to direct people to the original works. Generative AI scrapers harvest the material for their own commercial projects, without regard to whether the material is copyrighted.

Some of the books used in the training of Generative AI came from pirated book sites; specifically the “Books3” database: https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/

On its website, OpenAI admits to using “publicly accessible” information in its training data:

“For this set of information, we only use publicly available information that is freely and openly available on the Internet – for example, we do not seek information behind paywalls or from the “dark web.” We apply filters and remove information that we do not want our models to learn from or output, such as hate speech, adult content, sites that primarily aggregate personal information, and spam. We then use the information to teach our models.” (https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed)

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?

There are some attempts at licensing images used in AI generation. See Getty’s AI engine for one example: https://www.theverge.com/2023/9/25/23884679/getty-ai-generative-image-platform-launch

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.

A Generative AI model does not need to maintain a bit-for-bit copy of the original work in order to generate a copy of it. You can see this in Getty Image's copyright infringement lawsuit. The images generated by the AI system very closely resemble the originals, and in some cases even include Getty's watermark logo https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion You can also see this in the Authors Guild lawsuit with book details: https://www.nytimes.com/2023/09/20/books/authors-openai-lawsuit-chatgpt-copyright.html Significant portions of the trained data remains enmeshed in the trained model.

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

Some have claimed that Generative AI does not infringe copyright because it is merely "learning" from existing work in the same way that humans do. This is simply untrue. Generative AI is closer to a calculator. We don't say that calculators "learn" math. But where a regular calculator leverages general rules of mathematics, Generative AI leverages the hard work of human artists.

When you ask a human "What is D-Day" or "Who is Barack Obama", they can respond based on their accumulated knowledge and lived experience. When you ask ChatGPT, it is not using thought or reasoning; it has no true learning or experience to draw upon. It doesn't really understand WW2 or know who Obama is. It's just regurgitating words based on its algorithm—words that ultimately were written by humans.

Without the trained, copyrighted works to draw upon, ChatGPT could output nothing.

For example: if life on Mars were discovered tomorrow, ChatGPT couldn't generate an article about it until it had been trained on existing articles written by human authors. Even then, all it could do was parrot back the details from those articles. It has no insight or information of its own to offer.

This is why you can see effects closer to outright plagiarism on niche topics. Without a vast host of data to draw upon, ChatGPT has a limited number of sources to parrot. It starts to look more like a grade-schooler copying what they read and shuffling a few words around.

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. (44) How should the fair use analysis apply if AI models or datasets are later adapted for use of a commercial nature? (45) Does it make a difference if funding for these noncommercial or research uses is provided by for-profit developers of AI systems?

A similar sort of issue has been addressed in open-source software with licenses that carry forward to derivative works. For instance: say that Library A is only licensed for research/noncommercial use and Library B includes Library A. You can't then use Library B in a commercial product.

In the same vein, you could allow copyrighted works to be used in training an AI model for research purposes, but then NOT allow that trained model to be used for commercial purposes.

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?

Generative AI works best when trained on a massive data set. However, I don't think this should affect the fair use analysis. As the Authors Guild said in its open letter to AI developers:

"Millions of copyrighted books, articles, essays, and poetry provide the "food" for AI systems, endless meals for which there has been no bill."

Without the copyrighted work, their product would not exist. The fact that they need to steal a lot of copyrighted work doesn't change the nature of the theft.

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? (46) 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?

You need to account for the impact on the market as a whole, not limited to any particular copyrighted work. For example, Generative AI could put thousands of book cover artists out of business without ever generating a cover that exactly mirrors a particular copyrighted cover.

Some argue this is just a natural march of technology. Didn't the printing press put all book-copiers out of business? Sure. But the printing press did not use the labor of the book-copiers against them. Generative AI is putting artists out of work using their own art.

Without the underlying copyrighted material, Generative AI doesn't have a product to sell. Its sole function is to generate derivative 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)?

Affirmatively consent, just as in all other instances of copyright and trademark use.

9.1. Should consent of the copyright owner be required for all uses of copyrighted works to train AI models or only commercial uses? (47)

Some consideration should be carved out for research, but allowing wholesale scanning of everything on the internet without consent (as was seemingly done for the current slate of AI models) seems an overreach. If I were doing a regular research study (not AI-related), Fair Use would allow me to use elements of a book in my research, but I couldn't just reprint the entirety of a book in my study.

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? (48)

Given that copyrighted work can appear in multiple sources/forms across the internet, I don’t see how a technical opt-out method is practical. Look at the “books 3” data set containing thousands of pirated books being used to train existing data sets, for instance. How would someone “opt out” of their *stolen* book being used to train AI? Affirmative consent is essential.

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?

Using the volume of the work is a chicken and egg argument. Organizations like OpenAI chose to use a massive volume of work in training precisely *because* they were not forced to consider anyone’s copyright. That shouldn’t be used after the fact as a justification for not requiring consent.

Put another way—the fact that it would have been difficult for Napster to negotiate with the thousands of musicians/record labels whose work it had stolen should not have been a valid justification for its continued existence. It had an obligation to secure permission to distribute their copyrighted work. (Also services like Apple Music and Spotify show that such a thing is, in fact, possible on a massive scale.)

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?

This question just further highlights why an opt-out system is infeasible. If the system is opt-in rather than opt-out, those responsible for the AI model would be forced to find the copyright holder (whomever that may be) and secure permission.

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

By contacting them and negotiating a license, just as one would for any other use of copyrighted material.

10.3. Should Congress consider establishing a compulsory licensing regime? (50) 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?

Absolutely not. Forcing artists to submit their work to an endeavor that threatens to put them out of business is wrong on so many levels.

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)?

I think valuable lessons can be learned here from the open-source software community. By having appropriate licensing agreements attached to data sets and AI models, those license conditions can be carried forward to any products that incorporate those components. This is similar to the way a software license is attached to a software library and imposes terms of use on other libraries or products that incorporate it.

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.

With the current models, that is not feasible. ChatGPT, for instance, isn’t referencing a specific database of sources, but has built a web of connections between words.

As Stephen Wolfram wrote, “(W)hen ChatGPT does something like write an essay what it’s essentially doing is just asking over and over again “given the text so far, what should the next word be?”—and each time adding a word.” (https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/)

Not only does ChatGPT not know where its information came from, it can often make up information—and sources. This was most famously demonstrated in the case of lawyers who relied on its fabricated legal citations. (https://arstechnica.com/tech-policy/2023/06/lawyers-have-real-bad-day-in-court-after-citing-fake-cases-made-up-by-chatgpt/) Other researchers have found cases where ChatGPT cites fake books or even fake patent registrations.

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

Yes, this information should be associated with the data sets. AI models can reference which data sets and other sources they were trained on, maintaining transparency at all levels.

15.1. What level of specificity should be required?

You’d need a system similar to citing sources for academic writing (title, author, URL, etc.) which reflects the fact that web pages can change over time.

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

This is yet another argument in favor of affirmative consent. If you need to opt-in for your work being used, then you don’t need a separate notification process; you already know where/how it’s being used.

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. There is artistry in choosing prompts/commands, but that's not the same as human creation. It's similar to fractal art. They can be beautiful, and humans do have a hand in tweaking the inputs, but they are not copyrightable because they are algorithmically generated. Someone else with those exact same prompts (and associated 'seed' values) could generate the exact same output—pixel for pixel, word for word—with an AI. And no matter the prompt, the output is still fundamentally derived from the underlying works in the AI model.

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?

I don't know how it could work on a technical level, but I believe such a requirement would be beneficial.

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

Family therapist Jonathan Decker commented on the SAG/WGA strike: "How many of us use these stories for inspiration, for escape, for connection with our loved ones, for connection with our friends, to recharge so we can face our lives again? I don't think we can understate how important art is for our mental and emotional health and has been throughout the history of humanity."
(https://www.youtube.com/watch?v=plDCA3cMN88)

Machines cannot replicate the emotions at the core of artistic expression. They have nothing new to offer the world. They can only create shallow imitations. But that's not going to stop big corporations from trying to replace human artists with cheaper machine "labor". We need to protect the artists from having their livelihoods destroyed by this new technology, just as copyright law originally protected artists from things like the printing press and photocopier.