Answers to Questions

Arising In Relation to the U.S. Senate Committee on the Judiciary Subcommittee on Intellectual Property Hearing on "Artificial Intelligence and Intellectual Property – Part II: Copyright and Artificial Intelligence" held on July 12, 2023

Matthew Sag
Professor of Law in Artificial Intelligence, Machine Learning, and Data Science
Emory University School of Law

Chair Coons, Ranking Member Tillis, Members of the Subcommittee:

Thank you for the opportunity to answer the following Questions for the Record.

Question 1. Given generative AI is developing all over the world and countries are responding to it in different ways, are there policies or regulations being adopted elsewhere that you recommend that the U.S. consider or avoid?

(1) In relation to authorship, there is a strong international consensus that authorship requires some kind of subjective intention to manifest or communicate a belief or a state of mind that is entirely lacking in current and foreseeable computer technology. No changes are required in U.S. law at this time.

A notable exception to this international consensus is Section 9(3) of the UK Copyright, Designs and Patents Act 1988. This provides that for "computer-generated" works, (meaning a work generated by a computer "in circumstances such that there is no human author") "the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken." Section 9(3) was not drafted in contemplation of generative AI and the provision is generally regarded as not particularly useful in that context, or any other.

In fact, the section shows some of the risks in trying to make computer generated works copyrightable. The phrase "the person by whom the arrangements necessary for the creation of the work are undertaken" is deeply ambiguous in the context of generative AI: it could refer to the authors of the works in the training data, the person or persons who selected the training data, the person or persons who trained the model, the person or persons who provided fine-tuning feedback in a process of reinforcement learning through human feedback, the person or persons who provided the necessary computing resources to train the model or to run the model in response to a user query, or to the person or persons who wrote the prompt. This is only the beginning of the uncertainty, but I believe I have made my point.

(2) In relation to the use of copyrighted works and training data, the United States' world leading position in AI research is partly attributable to the flexibility afforded by the fair use doctrine. Given the pace of technological change, there may be some wisdom in continuing to allow the courts to apply the fair use doctrine to particular situations, rather than prescriptively legislating in more detail. The track record of other jurisdictions trying to legislate for issues we deal with through fair use is mixed, at best. Attempt to write technology specific rules often fail to predict the past, let alone the future.

On the other hand, business and research communities may benefit from an express copyright limitation that allows for text data mining, similar to Article 3 and Article 4 of the EU Digital Single Market Directive. I have reservations about this approach if it displaces fair use and I anticipate that any "clarifying" legislation will nonetheless require significant judicial clarification.

However, the U.S. could benefit from a clear safe harbor for the use of copyrighted works as training data for machine learning. I address this in my answer to question 8, below.

Question 2. A recent survey on how consumers view AI found that most consumers – nearly 80% – believe the use of AI should be explicitly disclosed. Do you agree? Why or why not?

In principle, it is hard to disagree with calls for disclosure and transparency. However, the devil is in the details.

It seems simple enough to propose that images generated by an AI without any significant human creative import should be labeled either through water marketing, metadata, or some other description as "AI generated."

However, we need to be clear on why such labeling is important and how critical terms will be defined.

- The line between AI-generated and human-generated may be difficult to draw when a person uses and then edits AI-generated text, or when a person uses AI-powered editing tools to manipulate a work that was initially human-authored.
- In some contexts, disclosure and labeling may be important because we want to assess whether the work should be considered as creative or expressive. Accordingly, the threshold for labeling something as AI-generated might be quite high.
- In other contexts, labeling may be important because the public needs to know whether the content they are being presented with has been manipulated, or even entirely manufactured. If a news report features an image of the Pope in a white puffy

> jacket, the tools used to create the image are far less important than the fact that the image is fake. Accordingly, any manipulation of the image or text should be disclosed.

My point is that the reasons we want disclosure need to align with the details of any proposed requirement, and that there may be tradeoffs between different objectives.

3. What are the benefits and disadvantages of requiring an AI company to keep records of everything that is ingested and to make those records publicly available?

a. Under what circumstances, if any, should an AI company NOT be required to make its records of everything that is ingested by the AI publicly available?

b. Under what circumstances, if any, should an AI company be required to make its records of everything that is ingested by the AI publicly available?

The term "ingested" is imprecise and creates unnecessary ambiguity:

Machine learning models "learn" by exposure to the training data. The training data influences the model, but it does not become part of the model, except in rare cases of overfitting (usually due to a failure to effectively deduplicate the training data). If Congress legislates in relation to "ingestion," technology companies will plausibly respond that the models do not ingest anything.

Record-keeping:

Any entity that relies on the fair use doctrine to assemble a significant quantity of copyrighted works for use as training data for machine learning should be required to keep accurate records that identify those copyrighted works and their provenance.

For each model, those records should also keep track of whether the work was ultimately used in training. Some works initially copied for the purpose of training may never actually be used because they are duplicates of other works, or because they have undesirable properties (for example, the inclusion of pornography or hate speech).

Any record keeping provision should apply to all works copied as part of the potential set of training data, not just those works actually used to train the model. Depending on the work, maintaining accurate records may require archiving a permanent copy of the work. Identifying a work by title and URL, may not provide sufficient information.

Disclosure:

(a) Any entity subject to the recordkeeping requirement suggested above ("a covered entity") should be required to publicly disclose summary information about the composition of copyrighted works in the training data associated with each new publicly released or commercially significant model in a timely fashion. The recordkeeping requirement should not apply to works included with the express permission of the relevant copyright owner, works in the public domain, etc.

(b) Covered entities should be required to disclose examples of specific works in each relevant category that were and were not included in the training data.

(c) Technology companies and researchers may have a legitimate interest in not publicly disclosing the entire contents of the training data. The composition of the training data could be a valuable trade secret in some circumstances. Moreover, information about the training data combined with other information could disclose personal or sensitive information about individuals in some contexts. Covered entities that do not disclose the entire contents of their training data to the public should be required to provide a mechanism to allow individuals to easily determine whether their works were part of the training data for a given model.

(d) Covered entities should be required to make the training data available for auditing by the Copyright Office, or some other government agency. The purpose of the audit would be to determine the adequacy of the summary in (a) and the disclosure mechanism in (c).

<u>Regulatory burden:</u>

The recordkeeping provision proposed above would impose a minimal burden on technology companies and researchers, because the steps outlined above should already be part of any data workflow. The disclosure provisions proposed above need to be carefully considered to ensure that they are reasonable and proportionate. To the extent that such provisions apply to university and not for profit researchers, they should be calibrated to that environment. The Copyright Office could be delegated authority to define exemptions to recordkeeping and/or disclosure requirements.

Question 4. Do you think that generative AI prompts provided by users are copyrightable? And if so, under what circumstances could they be copyrightable?

A user prompt *could* be copyrightable in some specific cases, however, in practice most user prompts will not be copyrightable.

Much like computer software, a prompt could be copyrightable as a literary work if it is original, if it contains a non-trivial amount of creativity, and if it is not simply an uncopyrightable idea or instruction. The mere fact that a literary expression is used as a prompt does not make it ineligible for copyright. However, most prompts will lack sufficient originality and creativity beyond their functional attributes and will in practice be ineligible for copyright protection.

a. Do you think that whether the prompt used is copyrightable or not should impact the copyrightability of the resulting AI output generated as a result of the provided prompt?

The copyrightability of a prompt does not itself determine the copyrightability of the content generated by an AI when the prompt is invoked. The copyrightability of an AI generated image depends on whether the image reflects a person's original conception of that image in sufficient detail. A prompt may be independently copyrightable as a literary work but entirely fail to produce an image that reflects the prompt-author's conception. For example, a copyrightable haiku could be used as a prompt, but the resulting image would not be copyrightable.

It is important to understand that the relationship between words typed into the context window of a generative AI and the output of that AI is quite different to the relationship between instruction and output in other contexts. When a composer writes a set of instructions in the form of sheet music, she does so with a very specific output in mind. Even though performers add a layer of creativity, subjectivity, personal interpretation in implementing sheet music instructions and producing musical sounds, there is a very tight correlation between the work envisaged by the composer and the work ultimately performed by a performer following the sheet music. In contrast, when a user inputs a prompt into a text-to-image generator, such as Stable Diffusion and Midjourney, she often has no specific expectation of the appearance of the resulting image. The process is deterministic in that the same prompt will result in the same image if the initial state of the program is the same, but from the user perspective, (1) specific outputs are generally so unpredictable as to appear random, and (2) platforms appear to be designed such that the initial state of the program is never the same. The same is true with ChatGPT, you never really know what the answer to a prompt will be, and it is never exactly the same because the system varies the initial state every time.

The copyrightability of AI generated content does not depend on the copyrightability of the instructions used to generate that content. As I explained in my written testimony:

> *Generative AI is often used as a tool in the creative process. A person who instructs a Generative AI with enough detail, such that model output reflects that person's original conception of the work, should be regarded as the author of the resulting work. However, simple text prompting is unlikely to meet this standard.*
>
> *… there is no reason in principle why prompts couldn't be detailed enough to meet the traditional threshold of authorship in some cases. Sophisticated prompts that specify details of an image should be sufficient to meet the requirement that the work that results from and reflects a person's original conception of the expression.*

> *Furthermore, refining text prompts and choosing between different outputs should also be recognized as way in which a human using Generative AI could meet the authorship standard.*

Question 5. What does the impact of generative AI have on the creative industry? Specifically, what are your thoughts regarding the concern that the proliferation of generation AI will take over jobs?

Generative AI has enormous potential to make creative people more productive and to allow more people to express their creativity. By reducing the cost of creation, generative AI will enable individuals and companies to do more with less — whether that implies an increase in creative production or any decrease in employment in creative industries is a difficult question to answer in the abstract. I am not an economist or an industry specialist, but I have studied economic history and how the law responds to and shapes new technology for the past 20 years.

In the early days of the World Wide Web (in the mid-1990s) it was easy to foresee how digitization and distributed networking could disrupt existing business models. However, most of the new opportunities and new business models the web made possible were unforeseen, and I suspect, unforeseeable. This is a recurrent story with disruptive technology. The jobs that might be lost loom large because they are foreseeable, the new jobs that the technology makes possible seem like optimistic speculation because we don't know with specificity what they will be. I don't wish to trivialize the legitimate anxieties professionals in the creative industries have in relation to AI, just to place them in a broader context.

The people whose jobs are most at risk from generative AI are those that produce content that is easily fungible with other content. Three obvious examples would be: stock photography, cover art, and narrative content for search engine optimization. In contrast, work that is customized, work that is part of a long-term relationship, or work that is valued because of its connection to the personality of a particular author does not appear to be at risk.

In answering this question, I have assumed that Congress will ensure that individuals are given adequate protection from the use of AI to generate synthetic content that recreates their voice, image, or likeness (i.e., deepfakes). One of the things that emerged most clearly from the July 12 hearing was that some of the most pressing concerns people have in relation to generative AI have nothing to do with copyright, and everything to do with the increased capacity of deepfake technology.

Question 6. If a generative AI system is found to infringe a copyrighted work, who should be liable for the infringement – the AI company, the user providing the prompts to the AI tool, or both?

Before I answer this question, it is important to note that AI produced content that infringes copyright is extremely rare. If we apply our traditional tests of infringement and seek to identify "substantial similarity" between the outputs of generative AI models and the copyrighted works used to training those models, we will almost inevitably come up short. The reason for this, as I explained in my written testimony, is that generative AI models 'usually learn from the training data at a fairly abstract level. Moreover, the output of generative AI usually combines abstract latent features learned from the training data in a way that ensures model outputs look nothing like specific model inputs."

The output of generative AI is most likely to infringe on copyrightable characters, where, practically speaking, the level of similarity required to establish infringement is more abstract.[1] Setting that issue aside, infringing output may result from (1) overfitting, usually a product of failing to effectively deduplicate the training data, or (2) from user instructions.

Who is liable for isolated instances of infringement that result from using generative AI will depend on whether courts apply the volitional act requirement in this context. Although copyright infringement does not require a particular mental state—you can infringe copyright by mistake, or even by subconsciously copying—it is widely assumed that a defendant may be held directly liable only if it has engaged in volitional conduct.[2] Technology companies will plausibly argue that although they provide the infrastructure to create images/music/text, if the user who actually prompts the system to create the output is the one who "makes the copy." Understandably, courts have not addressed the application of the volitional act requirement in the context of generative AI, and it is arguable that this context should be distinguished from the automated copying and transmission systems where the volitional act requirement has been applied in the past.

If the user is treated as the one who "makes" the offending copy, then under current law the technology provider will only be liable if the requirements of the vicarious, contributory, or inducement-based liability are made out.

If the technology provider is deemed to "make" the offending copy, it is essentially subject to strict liability. Congress may wish to consider whether some intermediate standard is desirable, such that technology providers have an obligation to take reasonable measures to

[1] For a more detailed explanation, see Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023)(Available at SSRN: https://ssrn.com/abstract=4438593)

[2] In Religious Technology Center v. Netcom On-Line Communication Services, Inc., 907 F. Supp. 1361, 1370 (N.D. Cal. 1995), the district court held that the defendant Internet service provider was not liable for the automatic reproduction of a copyrighted work by its computer system. The court refused to impose direct liability on the service provider, reasoning that: "Although copyright is a strict liability statute, there should still be some element of volition or causation which is lacking where a defendant's system is merely used to create a copy by a third party."

prevent infringement, but are not held liable for the independent choices of the platform users.

Regardless of who is liable for individual instances of infringing output, the existence of infringing output may have implications for whether the fair use defense applies to the assembly of the training corpus in the first place. As I explain in a forthcoming Law Review article:

> *If ordinary and foreseeable uses of generative AI result in model outputs that would infringe on the inputs no matter what intervening technological steps were involved, then the non-expressive use rationale no longer applies. If training LLMs on copyrighted works is not justified in terms of non-expressive use, then there is no obvious fair use rationale to replace it, except perhaps in the non-commercial research sector. If LLMs just took expressive works and reconveyed that same expression to a new audience with no additional commentary or criticism, or no distinct informational purpose, that would be a very poor candidate for fair use.*[3]

Question 7. In your opinion – currently or in the foreseeable future – can AI generated material ever replace the quality of human created work?

The literal answer to this question is, yes. We have already seen examples of AI content that has been adjudged to be as good as human authored work.

However, I believe the spirit of the question is really about whether there is something special or significant that should make us regard human authored works more highly than AI generated content. In some cases, the answer is clearly yes, in much the same fashion as we often regard an original work of art as more desirable than a copy. But in many cases, there is nothing intrinsically special about human authored content.

Question 8. A balance needs to be struck in terms of how to encourage innovation, how to be responsible, and how to ensure that there is clarity for all using this technology. How do you propose we do this in the copyright space in a way that allows the U.S. to stay competitive and remain the global leader?

The fair use doctrine already gives U.S. technology companies and researchers a substantial advantage over their peers in many other developed nations. Courts in the United States have a strong track record of applying the fair use doctrine in a way that balances innovation with respect for the interests of copyright owners. Properly applied, the fair use doctrine allows for technical acts of reproduction that do not interfere with the copyright owner's interest in controlling the communication of their original expression to the public. The

[3] Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023)(Available at SSRN: https://ssrn.com/abstract=4438593).

courts were correct to rule that peer-to-peer file sharing was not fair use; they were also right to find that Google Books and HathiTrust were fair use.

The U.S. could supplement the fair use doctrine by establishing a safe harbor regime for non-expressive uses, without prejudice to the general application of Section 107 of the Copyright Act. The application of the safe harbor could be conditioned on taking certain affirmative steps to protect both copyright and non-copyright interests of authors and copyright owners of works used in training data. Any such safe harbor should be optional, not compulsory, to avoid First Amendment entanglement. Key requirements for the safe harbor should be modeled on the best practices for deduplication and avoiding overfitting that have already been identified in the computer science and legal literature;[4] but they must also provide room for the development of new best practices that will doubtless emerge. The safe harbors could be designed to protect interests that relate to right of publicity and trademark related concerns.

Question 9. In the copyright context, what differentiates the technology of generative AI from other machine-aided creativity, such as photography, video cameras, electronic music, and the like, all of which allow the public to develop and advance knowledge?

There is a long history of technologies that have enabled new forms of creativity or reduced the cost or skills required to engage in creativity, and each one has disrupted existing market structures. Generative AI as part of this tradition, but there are some important distinctions to be drawn.

Authorship, originality and ownership:

In traditional machine-aided creativity, such as photography or electronic music, the authorship of the tool-user is rarely questioned. Although operating a camera take less skill and training than painting with oil on a canvas, courts have long recognized that the combination of minute anesthetic decisions involving framing, timing, lighting, positioning, etc. are enough to make the photographer the author. In contrast, as discussed above and in my written testimony, much of the content produced by generative AI does not meet the authorship standard and is thus uncopyrightable.

Reproducibility and Volume:

AI can generate vast amounts of new, seemingly creative content at a speed and volume far beyond human capacity. In contrast, traditional machine-aided creative processes are constrained by human capabilities and time.

[4] For an introduction to this literature, see Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023)(Available at SSRN: https://ssrn.com/abstract=4438593).

Copyright law is premised on the fact that information goods are expensive to create but cheap to copy. Generative AI may undermine that premise by making novel information goods cheap to create. I don't think this makes copyright obsolete, but it does explain why the uncopyrightability of generative AI content is no cause for concern.

Question 10. What steps can and should the creative community take today to ensure that their work is more easily attributed to them, regardless of whether their work is used for training an AI model? For example, indicating authorship and contact information via the metadata of the author's digital content.

Creative communities may need to revisit the contractual terms under which their work is distributed. Many open-source and creative commons licenses implicitly allow for works to be used in machine learning training, but these contracts were not drafted with this scenario mind. The users of social media platforms and cloud hosting services may also be surprised by the extent to which they have already agreed to allow their photos, videos, music, and social media posts to be used to train machine learning algorithms. The same goes for any creative professional who contracts with an aggregator, such as a stock photo agency.

Persistent metadata about signals appropriate and inappropriate uses of a work would be advantageous, but I believe other witnesses are better placed to address this issue.

Question 11. Are existing laws and regulations sufficient to deal with the issues relating to transparency and record keeping by AI companies?

No.

I am not aware of any laws or regulations that require companies developing generative AI tools to disclose the details of which copyrighted works were used in training, or even to keep a complete and accurate record of those works.

For a proposal for such a requirement, see my answer to question 3.

Question 12. Have you reviewed the U.S. Copyright Office's Registration Guidance for "Works Containing Material Generated by Artificial Intelligence" and, if so, what are your views on the guidance?

a. Do you think that the Copyright Office got it right? Are there aspects of the guidance that could stand to be clarified or revised?

Please refer to Appendix A of my written testimony, "When Should A Human Be Credited With Authorship Of Something Created Using Generative AI?"

Question 13. Both the U.S. Patent and Trademark Office and the U.S. Copyright Office have engaged in extensive outreach regarding AI. Have you participated in this outreach and, if so, how did you find it? What more can and should these offices do?

The USPTO and the Copyright Office deserve recognition for their outreach in relation to the intersection of AI with intellectual property.

I believe it would be beneficial for one of these agencies to convene a working group to suggest best practices for generative AI.

Although copyright infringement, trademark infringement, and interference with privacy and personality rights, do not pose the same existential risk as Skynet or an out-of-control paperclip factory, these copyright and copyright-adjacent risks are foreseeable in relation to generative AI. These risks are also a lot more likely.

Like other issues in AI safety, addressing the potential for copyright infringement and other related harms will require technical solutions informed by legal, ethical, and policy frameworks. The USPTO or the Copyright Office could make a significant contribution without the need for additional legislation by exploring options for continuing the development of generative AI while reducing potential harms and adverse impacts.[5]

Question 14. Language Learning Machines are increasingly being used to generate source code and help software developers write software. Such models can require a vast amount of source code and thus can turn to open-source software (OSS) for scraping publicly available source code.

a. If AI models are trained on OSS, does that infringe on the copyright of the respective authors?

b. If an AI model is used to generate code, does that generated code constitute a derivative work? And if that AI model was trained on copyleft-licensed OSS, must it also be licensed under copyleft?

There is a very close relationship between the two parts to this question.

(1) An AI model trained on open-source software is likely to qualify as fair use as long as the outputs of the model are not substantially similar to the protectable original expression of

[5] For an initial proposal for best practices for "Copyright Safety for Generative AI," see Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023)(Available at SSRN: https://ssrn.com/abstract=4438593).

the copyrighted inputs. However, making this assessment in the context of computer software is complicated by the anomalous nature of software.

(2) An AI model trained on open-source software may not even need to qualify as fair use if the open-source license permits copying. Whether copying is permitted depends on the exact terms of the license and the specific details of how the training data is used. Some open-source licenses allow for unrestricted reuse, some allow for unrestricted non-commercial reuse, some prohibit the creation of derivative works, some are contingent on the license terms being carried forward to derivative works based on the open-source software.

Let's consider and open-source license that authorizes reproduction and the creation of derivative works, but under the condition that any derivative works must also be licensed under the same terms. I.e., a viral license. In that scenario, an AI developer could reproduce the works as part of the training data and fall within the terms of the license by (a) releasing the trained model under the same open-source license,[6] or (b) ensuring that the model did not amount to a derivative work. This is quite plausible because the learned weights and biases of an AI model trained on open-source software represent an abstraction and generalization of the input data, rather than a copy of the training data. There are some clear examples of generative AI coding tools memorizing and repeating examples from the training data, but if this can be avoided the trained model will not amount to a derivative work.

I should say more about derivative works because this is an aspect of copyright law that people find very confusing. The scope of the Copyright Act's right "to prepare derivative works based upon the copyrighted work" (Section 106(2)) is often misunderstood. A poem inspired by a painting is not a derivative work. An index to a textbook is not a derivative work. A frequency table showing how often words are used in a novel is not a derivative work. A piece of software that is not substantially similar to the software in the training data is not a derivative work.

Making a derivative work necessitates recasting a qualitatively and quantitatively significant amount of the primary work's original expression into a new form or a new version. Assessing whether this threshold has been met requires some understanding of what made the primary work copyrightable in the first place. Suppose I reduced a novel such as *Fifty Shades of Gray* down to a table of individual words and the frequency with which they appeared in the text. I could program a computer to randomly construct an alternative novel, *Gray Fifty Shades Of*, which followed traditional rules of English grammar and used the same individual words in the same proportions (plus or minus 5%, to give it some flexibility). A few things should be obvious about, *Gray Fifty Shades Of*: (i) it would be terrible; (ii) it would not exist, but for *Fifty Shades Of Gray*, (iii) but it would not convey any of the original expression of the primary work. Without some nontrivial overlap in original expression *Gray*

[6] Complying with attribution requirements in some of the creative commons licenses could be tricky.

Fifty Shades Of would not be a derivative work. On the other hand, a sequel to the primary work that uses the same characters and settings would be very likely to be a derivative work.

Question 15. Some AI developers have said that the ingestion of copyrighted works is transformative and qualifies as fair use. What impact does the Supreme Court's recent decision in *Andy Warhol Foundation v. Goldsmith* have on that position?

The Supreme Court's 2023 decision in *Andy Warhol Foundation v. Goldsmith* ("*AWF*") emphasizes that the question of "whether an allegedly infringing use has a further purpose or different character … *is a matter of degree*, and the degree of difference must be weighed against other considerations, like commercialism."[7]

AWF reaffirms the importance of transformative use and implicitly rejects lower court rulings that had found uses to be transformative where there was no significant difference in purpose. Simply adding a layer of new expression or a new aesthetic over-the-top of someone else's expressive work and communicating both the old and new expression to the public in a commercial context, without further justification, is not fair use. The Second Circuit was wrong to suggest in *Cariou v. Prince*, 714 F.3d 694 (2d Cir. 2013) merely imposing a "new aesthetic" on an existing work was enough to be transformative. It was correct to retreat from that position in *Andy Warhol Foundation v. Goldsmith* 11 F.4th 26 (2021). The Supreme Court's decision in *AWF* simply reinforces the position that the Second Circuit had already taken. It is not a major change in the law of fair use, even if it did puncture some wishful thinking about fair use.

AWF helpfully clarifies the reason why a transformative use has featured so prominently in the case law: the more transformative a use is, the less likely it is to substitute for the copyright owner's original expression. Consider classic fair uses such as parody, commentary, or criticism may include substantial portions of the author's original expression, but these uses are so intrinsically different that they do not usually pose any risk of expressive substitution. In contrast, merely adding an overlay of new expression provides no such comfort. Deriving uncopyrightable abstractions from training data and using those obstructions to generate novel images/music/text is highly transformative.[8] Nothing in the *AWF* indicates to the contrary.

[7] Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith (2023), Slip Opinion at 12. (emphasis added).

[8] A.V. v. iParadigms Liab. Co., 544 F. Supp. 2d 473, 482 (E.D. Va. 2008): "This Court finds the "purpose and character" of iParadigms' use of Plaintiffs' written works to be *highly transformative*. Plaintiffs originally created and produced their works for the purpose of education and creative expression. iParadigms, through Turnitin, uses the papers for an entirely different purpose, namely, to prevent plagiarism and protect the students' written works from plagiarism. iParadigms achieves this by archiving the students' works as digital code and makes no use of any work's particular expressive or creative content beyond the limited use of comparison with other works." AV Ex Rel. Vanderhye v. iParadigms, LLC, 562 F. 3d 630, 640 (4th Cir, 2009): "The district court, in our view, correctly determined that the archiving of plaintiffs' papers was transformative and favored a finding of "fair use." *iParadigms' use of these works was completely unrelated to expressive content* and was instead aimed at detecting and discouraging plagiarism." Authors Guild, Inc. v. HathiTrust, 755 F. 3d 87, 97 (2nd Cir. 2014): "… we conclude that the creation of a full-text searchable database is a *quintessentially transformative* use.");

Question 16. Scraping the Internet for data – text, images, audio, video, etc. – for use in training AI models has all the current focus. However, once this has been done the focus may shift to sources of data that are not as readily accessible, such as private user data. Do you foresee companies using cloud-based file storage systems – such as Microsoft OneDrive, Google Drive, Dropbox etc. – as a potential source of data to be scraped? What are your thoughts on this?

(1) Scraping training data from closed sources, such as private cloud storage, without express permission would almost certainly violate federal and state laws targeted at computer hacking.

Most obviously, it would amount to access without authorization in contravention of the Computer Fraud and Abuse Act. Furthermore, web scraping that involved defeating or circumventing technological protection measures would also likely violate the DMCA's anti-circumvention provisions, see Section 1201. Web scraping can also implicate people's privacy rights. As of 2023, there is no broadly applicable federal privacy law and many of the relevant state statutes are quite recent.

Furthermore, in the right circumstances, web scraping could also give rise to a claim sounding in unfair competition, unfair and deceptive trade practices, trespass to chattels, conversion, trade secret claims, tortious interference with a contract, tortious interference with a prospective economic advantage, unjust enrichment, and misappropriation.

(2) Of course, none of this would prevent a company like Microsoft, Amazon, or Google from using their customers private cloud storage data as training data if that activity is permitted under their terms of service. Cloud storage companies tend to give themselves an incredibly broad latitude in this regard and Congress may wish to consider whether additional consumer and business protections are required in this context.

(3) Web scraping is often a violation of the terms of service of the relevant website. Whether terms of service are enforceable contracts is a question of state law. The Solicitor General recently expressed doubts as to the enforceability of terms that are merely posted on a website without requiring some express affirmation or agreement in an amicus brief in *ML Genius Holdings LLC v. Google LLC*. In that case, the Second Circuit held that the browse-wrap terms of service that placed limits on the reproduction of music lyrics posted to the ML Genius website were preempted under the Copyright Act. The Supreme Court denied

Authors Guild, Inc. v. Google, Inc., 804 F.3d 202, 216-7 (2d Cir. 2015): "We have no difficulty concluding that Google's making of a digital copy of Plaintiffs' books for the purpose of enabling a search for identification of books containing a term of interest to the searcher involves a *highly transformative* purpose, in the sense intended by Campbell." Authors Guild, Inc. v. Google, Inc., 804 F.3d 202, 217 (2d Cir. 2015): "… through the ngrams tool, Google allows readers to learn the frequency of usage of selected words in the aggregate corpus of published books in different historical periods. *We have no doubt that the purpose of this copying is the sort of transformative purpose described in Campbell* as strongly favoring satisfaction of the first factor."

ML Genius' petition for certiorari. Congress should consider whether legislation is needed to clarify the scope of Copyright Act preemption in relation to contracts.

(4) Even in cases where training and AI model on copyrighted works amounts to a non-expressive use, the particular facts of a given case could tip the balance against fair use. I will say more on this in my answer to question 18, but for the moment it is worth noting that a court could consider that obtaining training data by violating the CFAA, Section 1201 of the Copyright Act, privacy laws, or binding contractual restrictions, is a significant factor militating against fair use.

Question 17. To what extent should the use of AI impact whether a human creator receives a copyright? In other words, if AI is being used as a "tool," should a human still be able to receive a copyright if they have independently contributed creative content?

Please refer to Appendix A of my written testimony, "When Should A Human Be Credited With Authorship Of Something Created Using Generative AI?"

Question 18. Let's assume that under *Andy Warhol Foundation v. Goldsmith* the use of copyrighted works for training AI is not considered transformative. Do you believe the use of these works would still qualify as fair use looking at the four factors? Which particular factors support your position?

The assumption is far-fetched. For the reasons explained in my answer to question 15, there is no good reason to think that the recent Supreme Court decision undermines the fair use status of non-expressive uses.

However, even without making the assumption, I can offer some thoughts on how the individual fair use factors should be applied in relation to generative AI.

> *(i) Assuming that machine learning training amounts to a non-expressive use, its "purpose and character" will favor a finding of fair use under the first statutory factor.*

The first fair use factor calls for an evaluation of "the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes."[9] Since the Supreme Court's 1994 decision in *Campbell v Acuff Rose*, whether the defendant's use was 'transformative"—meaning that the use added "something new, with a further purpose or different character"—has been the central question under the first factor.[10] The Court's 2023 decision in *Andy Warhol Foundation v. Goldsmith* ("*AWF*") emphasizes that the question of "whether an allegedly infringing use has a further purpose or

[9] 17 USC 107.

[10] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994). For an empirical analysis, see Clark D. Asay, Arielle Sloan, and Dean Sobczak. *Is transformative use eating the world,* 61 B.C. L. REV. 905 (2020).

different character … *is a matter of degree*, and the degree of difference must be weighed against other considerations, like commercialism."[11]

If a machine learning model was trained in such a way that its outputs were substantially similar to its inputs, the copying that produced the training data would not amount to a non-expressive use, it would not be considered transformative, and the remaining fair use factors would go against fair use. On the other hand, assuming that the outputs are not substantially similar to the inputs, the copying that produced the training data would be a non-expressive use, it would be highly transformative, and it would thus be of a "purpose and character" that was consistent with fair use — regardless of whether it was undertaken by a commercial or nonprofit entity. In this scenario, which should be the more common one, the remaining fair use factors would be addressed as follows…

> *(ii) The second factor, the nature of the copyrighted work, has no independent relevance, it is not a factor that goes either for or against fair use, it is the context in which the other three factors must be evaluated.*

Factor two, the nature of the copyrighted work, simply reminds courts to take context into account when addressing the substantive considerations of purpose and character (factor one), amount and substantiality (factor three), and effect (factor four).

Some authorities suggest that the nature of the work, whether it is creative/informational, or published/unpublished is a stand-alone consideration such that some works merit greater copyright protection than others. This approach is ill-conceived. The nature of the work is not an independent factor that weighs in favor or against a finding of fair use, it is simply the context in which courts must apply the substantive considerations of purpose, proportion, and effect set out in factors one, three, and four respectively.

To elaborate, images are not less worthy of copyright protection than text, but it is much harder to selectively comment on an image or use just part of an image as evidence than it is with purely textual works. Accordingly, full quotations of an image might be reasonable and proportional in circumstances where partial quotation of the text would be.

The statute is not wrong to direct courts to think about the nature of the work; indeed, it would be quite impossible to analyze the purpose, proportion, and effect of the defendant's use without taking into account the nature of the work. Moreover, works like computer

[11] Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith (2023), Slip Opinion at 12. (emphasis added).

software may have special characteristics that inform a fair use analysis.[12] Note that the majority's fair use analysis in *Google v Oracle* begins with the second factor.[13]

> *(iii) If a use is non-expressive, then the third statutory factor which considers "the amount and substantiality of the portion used" will also favor finding of fair use.*

The ultimate question under the third fair use factor is whether the amount of copying was reasonable in relation to a purpose favored by fair use.[14] Although non-expressive uses typically involve making complete literal copies, that copying has been found to be reasonable because it is an intermediate technical step in an analytical process that does not lead to the communication of the underlying original expression to a new audience. Accordingly, courts in in non-expressive use cases have found the third factor weighs in favor of the defendant.[15]

> *(iv) If a use is non-expressive, the fourth statutory factor which considers the effect on the "potential market for or value of the copyrighted work" will also favor a finding of fair use in many cases.*

The "market" and "value" referred to in the fourth fair use factor are not simply any benefit the copyright owner might choose to nominate; these terms mean something much more specific. A critical book review that quotes from a novel does not have an adverse market effect if it persuades people to buy different book instead;[16] a report from a plagiarism

[12] Google LLC v. Oracle America, Inc., 141 S. Ct. 1183, 1198 (2021) ("… fair use can play an important role in determining the lawful scope of a computer program copyright … . It can help to distinguish among technologies. It can distinguish between expressive and functional features of computer code where those features are mixed. It can focus on the legitimate need to provide incentives to produce copyrighted material while examining the extent to which yet further protection creates unrelated or illegitimate harms in other markets or to the development of other products.")

[13] Id. at 1201.

[14] Campbell v. Acuff-Rose, 510 U.S. 569, 586-87 (1994). ("[T]he extent of permissible copying varies with the purpose and character of the use.") In *Campbell*, the Court characterized the relevant questions as whether "the amount and substantiality of the portion used ... are reasonable in relation to the purpose of the copying," and noting that the answer to that question depends on "the degree to which the [copying work] may serve as a market substitute for the original or potentially licensed derivatives[.]" *Id.* at 586-588.

[15] AV Ex Rel. Vanderhye v. iParadigms, LLC, 562 F. 3d 630, 642 (4th Cir, 2009); Authors Guild, Inc. v. HathiTrust, 755 F. 3d 87, 98 (2nd Cir. 2014) "In order to enable the full-text search function, the Libraries, as we have seen, created digital copies of all the books in their collections. Because it was reasonably necessary for the HDL to make use of the entirety of the works in order to enable the full-text search function, we do not believe the copying was excessive." Authors Guild, Inc. v. Google, Inc., 804 F.3d 202, 221 (2d Cir. 2015) "Complete unchanged copying has repeatedly been found justified as fair use when the copying was reasonably appropriate to achieve the copier's transformative purpose and was done in such a manner that it did not offer a competing substitute for the original." Authors Guild, Inc. v. Google, Inc., 804 F.3d 202, 221-222 (2d Cir. 2015) "As with *HathiTrust*, not only is the copying of the totality of the original reasonably appropriate to Google's transformative purpose, it is literally necessary to achieve that purpose. … While Google makes an unauthorized digital copy of the entire book, it does not reveal that digital copy to the public. The copy is made to enable the search functions to reveal limited, important information about the books."

[16] Campbell v. Acuff-Rose, 510 U.S. 569, 591-592 (1994) "We do not, of course, suggest that a parody may not harm the market at all, but when a lethal parody, like a scathing theater review, kills demand for the original, it does not produce a harm cognizable under the Copyright Act. Because parody may quite legitimately aim at garroting the original, destroying it commercially as well as artistically, the role of the courts is to distinguish between biting criticism that merely suppresses demand and copyright infringement, which usurps it."

detection service might depress the market for helping students cheat on their homework, but that is hardly a cognizable injury under copyright law. [17] More generally, in *Campbell* and subsequent cases, the courts have recognized that the copyright owner has no protectable interest in preventing criticism, parody,[18] or simply locking up unprotectable ideas and expression.[19]

By definition, if a use is non-expressive then it poses no direct threat of expressive substitution and thus should generally be preferred under the first factor (purpose and character) and considered harmless under the fourth factor (market effect).

The argument that copyright owners have an inherent right to charge for non-expressive uses, and thus suffer an adverse market effect under the fourth factor, is transparently circular. In theory, every defendant in every fair use case could pay the plaintiff for the right to engage in the challenged use, but if the use is fair there is no obligation to pay. To avoid such circular arguments, courts have limited market effect under the fourth factor to those that represent a cognizable copyright interest.[20] Accordingly, in *HathiTrust*, the second Circuit rejected the plaintiff's argument that not being paid for text mining was a cognizable harm, noting that "[l]ost licensing revenue counts under Factor Four only when the use *serves as a substitute for the original* and the full-text-search use does not."[21] Likewise, in *Google Books*, the court insisted on focusing "on whether the copy brings to the marketplace *a competing substitute for the original*, or its derivative, so as to deprive the rights holder of significant revenues because of the likelihood that potential purchasers may opt to acquire the copy in preference to the original."[22]

The substitution the courts are referring to here is expressive substitution, not simply the threat of a more competitive marketplace. A non-expressive use can be harmless under the

[17] AV ex rel. Vanderhye v. iParadigms, LLC, 562 F.3d 630, 464 (4th Cir. 2009) ("Clearly no market substitute was created by iParadigms, whose archived student works do not supplant the plaintiffs' works in the 'paper mill' market so much as merely suppress demand for them, by keeping record of the fact that such works had been previously submitted .... In our view, then, any harm here is not of the kind protected against by copyright law.")

[18] Campbell v. Acuff-Rose, 510 U.S. 569, 577-79 (1994); NXIVM Corp. v. Ross Inst., 364 F.3d 471, 482 (2d Cir. 2004) ("[C]riticisms of a seminar or organization cannot substitute for the seminar or organization itself or hijack its market."); Bill Graham Archives v. Dorling Kindersley, Ltd., 448 F.3d 605 (2d Cir. 2006) ("A copyright holder cannot prevent others from entering fair use markets merely by developing or licensing a market for parody ... or other uses of its own creative work.") (internal quotations omitted).

[19] Sega Enters., Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992); Sony Computer Entm't, Inc. v. Connectix Corp., 203 F.3d 596 (9th Cir. 2000).

[20] Campbell v. Acuff-Rose Music, 510 U.S. 569, 591-92 (1994) (no cognizable market effect where parody or criticism depress demand for the original work); see also Sony Computer Entm't, Inc. v. Connectix Corp., 203 F.3d 596, 607 (9th Cir. 2000) (noting that a videogame manufacturer's desire to foreclose competition in complementary products was understandable, but that "copyright law ... does not confer such a monopoly."); Bill Graham Archives v. Dorling Kindersley, Ltd., 448 F.3d 605, 615 (2d Cir. 2006) ("[A] copyright holder cannot prevent others from entering fair use markets merely by developing or licensing a market for parody, news reporting, educational or other transformative uses of its own creative work.") (citations and quotations omitted).

[21] Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 100 (2d Cir. 2014) (emphasis added).

[22] Authors Guild v. Google, Inc., 804 F.3d 202, 223 (2d Cir. 2015) (emphasis added).

fourth factor even if it results in the creation of a competing product—as long as the competing product does not contain an infringing level of original expression taken from the plaintiff's work. In *Sega v. Accolade* and again in *Sony Computer Entertainment v. Connectix Corp.,*[23] the Ninth Circuit found that reverse engineering a gaming console in order to produce interoperable games (*Sega*), and a rival gaming platform (*Sony)*, was fair use. In both cases the Ninth Circuit found that there was no cognizable market effect because the rival products did not contain any protectable expression derived from the plaintiffs' consoles. The defendants were entitled to use uncopyrightable elements from those consoles to make new independent creative expression possible.[24]

(v) However, other considerations may nonetheless tilt the fourth factor against fair use.

Non-expressive uses that substantially undermine copyright incentives could be considered unfair. To recap on my written testimony,

> (1) A court in some future case may well consider whether a defendant had lawful access to the works used as training data under the fourth factor.
>
> (2) Likewise, a future court might extend the fourth factor to consider whether, in scraping material from the Internet, the defendant ignored robot.txt files indicating a desire to opt out of search engine indexing and similar activities. Likewise, a court might conclude that scraping material from a website in violation of its terms of use was relevant to the fourth factor, if the inability to rely on such exclusions substantially undermined copyright incentives.
>
> (3) A plaintiff might argue that it is unfair to systematically extract valuable uncopyrightable material from a website or other information source and then use that material as a substitute for the functionality of the website. This argument would be strongest where the systematic extraction was likely to significantly undermine the website's incentives for original content production.
>
> This argument is hard to reconcile with the view that the idea-expression distinction is meant to encourage competition where the competing product does not include too much of the plaintiff's original expression. But it is not foreclosed by existing precedent.

These may be valid considerations under the fourth factor, but I would not elevate them to independent factors or prerequisites. They may have different salience in different cases and will generally be more relevant in commercial fair use cases than non-commercial ones.

[23] Sega Enterprises Ltd. v. Accolade, Inc., 977 F. 2d 1510, 1523. Sony Computer Entertainment v. Connectix Corp., 203 F. 3d 596, 608.

[24] *Id.*

Question 19. One concern about generative AI that has been raised by creators is that unauthorized copies of their works are being made during the process of collecting data and training a respective model. Could you please explain how copies and how many copies of such data are made and when within the lifecycle of creating and executing an AI system – from start to end?

The answer to this question may vary significantly depending on the model being trained. However, I can answer the question with reference to a generic example:

Model development begins by identifying and obtaining access to the relevant training data. It is hard to imagine that any large model could be trained without at least one locally stored copy of the training data. To avoid overfitting (and thus hopefully minimize the risk of copyright infringement and other analogous harms), it is important to deduplicate the training data. Practically speaking, this is hard to do without creating a semi-permanent local copy. To address questions of bias and filter out toxic materials, the potential training data needs to be analyzed carefully before training begins. Again, this is much more practical with access to a semi-permanent local copy. Storing a semi-permanent local copy also makes sense if the developer anticipates the need to retrain the model from time to time. Continued access to the training data in its original form may also be necessary to evaluate the performance of the model, and to take additional steps to mitigate the potential for copyright infringement, or other undesirable outcomes.

The training process itself does not involve copying or storing documents in their original format. However, segmenting the training data into tokens and converting those tokens into a numerical representation is, technically, another form of copying.

To elaborate, the data used to train models like GPT-3 and other text-based large language models do not consist of words or symbols that are meaningful to, or intelligible by, humans. At the beginning of the training process the raw text data is broken down into smaller pieces, known as tokens. These tokens could be as short as one character or as long as one word (in English and similar languages). For example, the text "U.S. Senate Hearing" is probably broken down into the tokens "U" "." "S" "Senate" "Hear" and "ing". Each token is then mapped to a unique numerical value.

This is a one-to-one mapping, so you could reverse engineer the original human-readable text from these numerical representations, and thus this still qualifies as a "copy" for the purposes of copyright law's reproduction right.

The actual training process for a model like GPT-3 involves feeding the numerical representations of tokens into the model and having it make predictions about the next token. Through this process, the model learns the structures, rules, and patterns in the language. The model doesn't (or shouldn't) retain any specific copyrighted works from the training data.

The model itself is an absurdly large statistical model that can be used to predict the next token given a set of input tokens. This model is not a copy of the training data.

When the model is deployed it generates entirely new content based on the statistical patterns it learned during training.

In summary, using copyrighted works as training data for generative AI necessarily involves at least two steps that would qualify as creating a copy under the reproduction right in Section 106(1). In practice, there may be some additional technical copies created, but none with any independent economic significance.

Question 20. Some have suggested different licensing structures for compensating copyright owners for the use of their works in AI training. What licensing structures have you seen or used that have worked to the mutual benefit of both AI companies and copyright owners?

As long as commercial AI developers respect machine-readable opt-outs, refrain from sourcing training data from sites of known infringement, respect paywalls, and other technological exclusions, we can expect a vibrant voluntary licensing system to emerge. Already, AI developers and negotiating access deals with media companies and stock photography agencies.

A compulsory license in relation to AI training would be difficult to administer and would interfere with voluntary licensing. Existing statutory licenses compensate right holders on a per-play or per-use basis and thus avoid the need to assess the merit or contribution of a given work. There is no easy way to assess how much a single work contributes to a machine learning model. If every work used to train a model is valued equally, then the remuneration an author or artist received would not be calibrated to the importance or value of her work; it would also tend to approach zero as the number of works in the training data increased.

***

Thank you again for the opportunity to assist the Senate in this hearing.

Matthew Sag

Matthew Sag