Comment submitted by Professor Daniel Gervais, Vanderbilt University

Professor Daniel Gervais
Vanderbilt Law School

Comment submitted by Professor Daniel Gervais, Vanderbilt University

I am grateful for this opportunity to provide input into the Copyright Office's work on the interface between copyright and Generative AI (GenAI). The Notice of Inquiry (NoI) demonstrates the quality of the inputs received prior to NoI but also of the serious and comprehensive analysis already performed by the Office.

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

The risks associated with this technology are significant, on several levels. I will mention two. First, there will be a shift from human creators to GenAI in many cases because GenAI will ultimately be cheaper, and, unlike authors, AI systems will not have rights that those who commercialize the content thus created need to negotiate with authors, from licensing and assignment of copyright rights to reversion and attribution and integrity rights. Though this is now mostly true for so-called mundane or low-quality content, this type of content has often been the path for authors to learn their craft. Very few authors have produced the Great American Novel or wrote an anthemic song on their first go around. Mozart, who started composing at age 6, did not write much if anything before the age of 21 that we still listen to today. Second, the progress of humans is linked directly—some might say synonymous with—the progress of ideas, and ideas progress when humans communicate with one another, including through literary and artistic works. As I wrote in an Essay published a few years ago (see answer to question 3 below):

> If machines can produce [...] literary and artistic works cheaper and faster than human creators, it a highly likely that industry will favor them over their human counterparts. In the copyright sphere, delegating to machines the task of helping us understand and interpret our world has profound consequences. It is through this interpretation that humans can become true agents in the world and ultimately change it. Delegating this very task to machines is thus pregnant with implications for the future for it changes its arc. It will not be complete obliteration of course. There will always be humans who write, pick up a paintbrush, or try to make a movie or sculpture, but if most of what we are given to read, watch or listen to comes from machines, much will be lost. If copyright protection is granted on outputs without a human cause, and assuming that the cost of machine productions will be lower (and machines will not ask for ongoing royalty payments or have reversion rights) then market forces will inescapably push for a replacement of human authors whenever it is commercially feasible. D. Gervais, The Human Cause

I would urge the Office to resist the commonly held view that any and all disruption caused by AI companies and especially Big Tech is per se positive and must be allowed by law, and instead consider that a diminution of works created by human is commerce, from journalism to essays to novels, is not a clear positive.

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? (Question shortened)

The news media is probably the most affected at present. Replacing human journalists with machines has serious implications for democracy. AI machines do not have a mission to produce "quality journalism." Moreover, studies have shown that the reduction in advertising revenue flow to major platforms combined with the ability to replace human journalists is putting certain types of journalism at risk, especially investigate reporting. See the studies mentioned under the next question.

3. Please identify any papers or studies that you believe are relevant to this Notice. (Question shortened)

D. Gervais, *The Human Cause,* in RESEARCH HANDBOOK ON INTELLECTUAL PROPERTY AND ARTIFICIAL INTELLIGENCE (R. Abbott, ed), (Edward Edgar, 2022) pp 21-38
D. Gervais, *The Machine as Author*, 105 IOWA L. REV. 2053-2106 (2020)

Jane C. Ginsburg and Luke Ali Budiardjo, *Authors and Machines* 34 Berk. Tech. L. J. 343 (2019).

On journalism specifically:

Sally Young & Andrea Carson 'What is a Journalist?', (2018) 19:3 *Journalism Studies* 452-472; online: https://www.tandfonline.com/doi/abs/10.1080/1461670X.2016.1190665?journalCode=rjos20.

The following book (no online access could be found) is relevant: Robert W. McChesney, and Victor Pickard, eds., *Will the Last Reporter Please Turn Out the Lights? The Collapse of Journalism and What Can Be Done to Fix It* (New York: The New Press, 2011).

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 legislative changes in the European Union, Japan, Singapore, and Switzerland, to name just those, are all very different. For one thing, they have different limits. It is not clear that all of them are compatible with the TRIPS Agreement's dual three-step test that applies to exceptions and limitations to the right of reproduction (namely article 9(2) of the Berne Convention, incorporated into TRIPS) and article 13 of TRIPS. It thus does not seem warranted to follow any one of those approaches at this juncture.

Consistency across all jurisdictions or even just the main ones is extremely unlikely. Some degree of harmonization, over many years, may yet happen. However, the issue of cross-border uses of datasets and the infringing or non-infringing nature of GenAI outputs that cross borders as physical objects or online is subject to a judicial determination of its legality in each country in which the output is exploited or available. A cross-border solution would imply a licensing arrangement with the rightsholders concerned. Licensing would compensate authors and other rightsholders for the use of their copyrighted works, but it would not directly address the issues identified in answers to questions 1 and 2 above. I return to this below.

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.

No change recommended at this stage.

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

Many AI models and especially LLMs are trained on available online material. The status of online material is often misunderstood. A copyrighted work that is available online is not "copyright-free" unless it is licensed under those terms, or in the public domain. That said, some amateur content uploaded to various platforms and services is licensed under broad Terms of Service that allows reuse for several purposes. Some of those may allow reproduction for training. This analysis must be conducted case by case.

There are troubling reports like this one of datasets of books available electronically that were never authorized by rightsholders. I have not been able to verify this information, but it may be relevant in litigation for example, as part of a fair use determination.

Although my understanding is that keeping a copy of the training data is not absolutely necessary, it is almost always the case that a copy is kept in case retraining is required and also because training from a local copy is more efficient. With or without curation of the data (for example to remove objectionable material). Those copies are most certainly not "transient" if they are retained for days, weeks or even months.

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

Some AI training is supervised by humans, but a lot of it is not. I explain this process in the *Machine as Author* piece referred to in answer to question 3, above. For a machine to explain what it was trained on (or what it trained itself with) is not always easy or even possible. It will be interesting to see how the AI companies will comply with the EU's AI Act obligation (if and when it is adopted) to "document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law." (Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206), art 28b(4)(c)).

8. Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use? Please discuss any case law you believe relevant to this question.

The two cases identified in the NoI, namely *Warhol* and *Google Books*, are certainly relevant to the inquiry. Though the latter case is from the Second Circuit, it was cited with apparent approval in *Warhol*.

Google Books is on point, but only to a certain extent. Like with AI training, copyrighted material was used to "feed" a computer to produce a new type of content: snippets made possible by the text searchability created by the scanning of millions of books. In the case of LLMs, the output is different: new literary and artistic content. AI companies will no doubt argue that both are equally "transformative." Yet in the case of GenAI, the use by the machine is not mere character recognition; it is semantic in nature. The machines process the expression of ideas in the works to create new expression. Whether that type of use, namely to create content that may compete with the material it was trained on, can be considered fair under *Warhol*, (of

course the case was limited to an analysis under the first factor) is an open question The substitutive potential of the material created by GenAI is also relevant under the fourth factor—perhaps even more so than under the first. Whether the use is commercial or not is certainly a relevant factor, but it is not dispositive.

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)?

As a matter of principle, an opt in is more in line with letting copyright holders decide how to administer the rights in their works. That said, international law, specifically art 5(2) of the Berne Convention, is interpreted as allowing an opt out as a "formality" to enjoy and/or exercise copyright rights. Given that (a) the EU has adopted this approach; and (b) that the opt out is a limit to an exception, the opt out would likely survive under a challenge in the unlikely event that a challenge is brought to the WTO (due to the incorporation of that provision of the Berne Convention into the TRIPS Agreement). This view is reinforced by published studies on the legality of Extended Collective Licensing, as it is now used in several countries around the world.

Perhaps an opt out could be added to digital files, as the robot.txt example demonstrates. This raises the question of whether a rightsholder opts out *for a work or for each digital object*, that is, is it necessary for the right holder to locate and opt out of every digital copy of the work? There is a point at which the burden may become insurmountable. Conversely, if a "work" is mentioned on an opt out list, how does a user identify all relevant copies (or phonorecords). In other words, the opt out mechanism should be well and clearly designed. If the obligation were imposed on every digital object, then a third party who knowingly puts a digital object without such code or other opt out signal could potentially be liable under an equivalent of 17 U.S.C. §1202.

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

Given that (a) users will be both large and small entities increasingly scattered all over the world, and that (b) right holders, also big and small, will similarly be located in dozens of countries around the world, it seems highly unlikely that only bilateral transactions (one rightsholder <-> one user) would be able to solve this dilemma within an appropriate time frame and with reasonable transaction costs. It is simply not reasonable to expect a user, especially a smaller one, to identify every right holder in every copyrighted work they want to use (even assuming they can determine what is and is not a protected work) and then locate and contact those rightsholders one by one. Nor does it make business sense for even large rightsholders to have an army of licensing agents dealing with potentially thousands of small-scale users around the world, not to mention currency and linguistic barriers. Put differently, the licensing market as it now stands means that deals can happen between large right holders and large users, with everyone else potentially left hanging.

This calls for a collective licensing model to work, on a nonexclusive basis and working hand in hand with direct licensing so that bilateral deals, especially between bigger players, can still be negotiated. This would require obtaining both rights to license and usage data for distribution purposes.

Existing systems (for example in the area of reprography, including for digital reproduction, distribution and storage of copyrighted material within organizations) show that market-based solutions are possible. Reprography is a good example in that it covers most printed material

that universities and businesses use, which would overlap to a large extent with major datasets now used to train AI.

I have long been a proponent of extended collective licensing in cases where it is necessary. See for example this report from 2003. I am much less convinced that a compulsory license is the best way forward. At the very least, it is too early to call a market failure, which would be the typical reason to consider a compulsory license.

11. What legal, technical or practical issues might there be with respect to obtaining appropriate licenses for training? (Question shortened)

The license should be secured by the organization making and storing copies of works. It should also cover the making of derivative works, or some of them, as I explain in more detail below (Q22).

Some AI companies may engage third party contractors, and ideally a license should allow companies to provide the data to such contractors, with appropriate safeguards.

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.

From a copyright perspective, the question is whether the protected elements of one or more works included in the training data were reproduced in substantial part in the output. This is a traditional copyright infringement analysis. The fact that the potential infringement occurred via a machine does not exculpate the infringement, though it is relevant for the derivative work analysis, as discussed in my answer to question 22, below.

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

Such a system would compensate authors and other rights holders of the use of their works. It would not protect future generations of authors from being replaced or "preempted" by generative AI. Preventing this replacement is not possible under copyright law (and possibly under any law), but given the societal risks of eradicating journalists, essayists , songwriters, screenwriters, and novelist, we can and should slow the replacement process by continuing to reject the idea that a machine can be an author. Protecting machine outputs without human cause/originality would put the full force of the market behind the replacement and accelerate to a point where we would have no time to devise policy or otherwise find appropriate ways to adopt to GenAI.

14. Please describe any other factors you believe are relevant with respect to potential copyright liability for training AI models.

All factors may be relevant, including where the data was taken from, and the jurisdiction(s) in which the data is copied, stored, and processed/tokenized.

Transparency & Recordkeeping

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?

See the answer to question 7, above.

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

I am not sure that "notification" is the term I would use in this context. If a licensing arrangement were in place, then reporting requirements would be negotiated as part of that arrangement.

17. Outside of copyright law, are there existing U.S. laws that could require developers of AI models or systems to retain or disclose records about the materials they used for training?

Unknown.

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? (Question shortened)

A machine is not an author. It cannot be. As I explain in detail in The Machine As Author, this would upend both copyright policy goals and long-standing doctrines. US law was set in the *Feist* case, and it should continue to be applied in the same way. A machine cannot make the type of creative choices identified in *Feist*. It takes a lot of time and money to train an AI system. Yet, it also took time and money to produce thousands of telephone books (getting the data from every subscriber, arranging it and printing and distributing the books), but that "sweat of the brow" is not the basis for copyright protection. An LLM can mimic human creativity by creating outputs that may look like they could have been produced by a human author, occasionally very well, but mimicry is not a sound policy basis for a claim to an exclusive right.

Moreover, copyright is an incentive, and I am not aware of convincing data to show a major crisis of underinvestment in GenAI. See also the answer to Question 20, below.

That being said, the authorship issue is a sliding scale. A work can be produced by an author *with the assistance* of AI tools. The fact that there are machine inputs in a work that otherwise has sufficient human authorship should not prevent the copyrightability of the work, though if the machine's contribution is separable, then a question can be raised about the copyrightability of those machine-produced portions.

The question of the copyrightability of the prompt is interesting. A detailed prompt long enough and with sufficient originality might be considered a protected text, though an argument might be raised about its functional nature. The more interesting question is whether authoring (or "engineering") a prompt means that the prompt "engineer" is the author of the resultant work. In almost every case, the answer should be negative. Oen can imagine situations, however, in which a detailed prompt—or a series of consecutive detailed prompts—could contain expressions of specific ideas that reflect human creative choices directly perceptible in the machine's output, in which case the argument that the prompts' originality may have "transferred" to the output could at least be made. I see those situations as exceptional, however.

Finally, the Office has dealt with claims of copyright from selection from many machine outputs. I believe it is entirely correct to deny registration if that is the sole basis for a claim of authorship. If I walk into a gallery or shop that specializes in African savanna paintings or pictures because I am looking for a specific idea (say, an elephant at sunset, with trees in the distance), I may find a painting or picture that fits my idea. That in no way makes me an author. The fact that a machine generated the options changes nothing to that analysis.

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?

Courts and the Office have so far applied the law correctly.

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?

As noted above, I am not aware of convincing data to show a major crisis of underinvestment in GenAI. The main issue is possible copyright and section 1202 infringement during the training process, not copyrightability of outputs.

Granting rights to machine productions would be a major doctrinal jump and a normative error. It would be the first type of *second degree intellectual property*--exclusive rights not to something humans have made, but to *what was made by what they have made*. I see no reason to change copyright law so fundamentally without a compelling reason. As noted above (Question 1), the risks of accelerating the replacement of human authors should be the dominant consideration and clearly militates against recognizing machines as authors.

21. Does the Copyright Clause in the U.S. Constitution permit copyright protection for AI-generated material?

The promotion of Useful Arts is not as simple as many people argue. They say machines produce new things and that's progress, period. The societal cost of replacing human authors (songwriters, journalists, film makers, novelists etc.) has yet to be fully understood, let alone precisely measured. The risk of what could be lost in less than a generation if the process of replacement is accelerated by granting copyright to machine productions is such that "progress" must be measured not just in the availability of new (cheap) "stuff," but also in terms of broader human progress. I, for one, would defend the view that the term "Progress" in Article I, Section 8, cl. 8, means human progress.

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?

This issue is more complex than meets the eye. There is Circuit split on the issue of whether to infringe the right to make derivative works in 17 U.S.C. §106(2), the defendant's production must itself be a work, i.e. something that would be protected by copyright if it weren't infringing. If one takes the view that the answer is yes, then because machines cannot produce original works, they cannot produce derivative works, and would thus be immune from infringement of the derivative work right. I deal with this in detail in my article AI Derivatives.

That said, as Professor Nimmer explains, when the derivative work right is infringed, there is almost necessarily an infringement of the reproduction right, the right of public performance, or both (2 Nimmer § 8.09[A][1]). For example, even if a court were to follow the cases that require originality to infringe the derivative work right, I am connived that it would not give a GenAI system a "free pass" to translate Stephen King's latest novel and sell copies. It would likely consider the matter under the reproduction right.

Yet there may be less obvious forms of derivative works created during the training process and use of AI systems, including abridgments and "condensations," two terms used in the definition of the term "derivative work" in 17 U.S.C. §101. There may be other ways in which the dataset would contain derivative works. Relevant uses could be included in a licensing arrangement, as discussed above.

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?

Whether a machine or human output infringes the right of reproduction should be subject to the same analysis because the effect on the copyright holder is similar.

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?

The current doubts about the legality of copying to train LLMs and removing Copyright Management Information (CMI) and the possible award statutory damages associated with both types of violation are incentives for AI companies and copyright holders to play cat and mouse. Whether discovery rules are sufficient will be demonstrated in the numerous pending lawsuits. Clearly, however, a system that would allow AI companies to use copyrighted material openly—indeed also reporting in an appropriate way on material used—would be better for everyone. A licensing arrangement could achieve this purpose.

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?

The answer to this question probably varies case by case. The person who made copies of the material would be a primary infringer. Who might then be secondary or vicariously liable requires a factual determination.

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?

My understanding is that the tokenization process used for GenAI strips digital copies of CMI, either by removing it after the copy of by copying the work without it. This could be a violation of 17 U.S.C. § 1202, which is subject to statutory damages and does not require registration. However, right holders have the burden of showing that the removal, alteration or distribution was done "knowing, or, with respect to civil remedies under section 1203, having reasonable grounds to know, that it will induce, enable, facilitate, or conceal an infringement of any right

under this title." The uncertainty of the outcome is another reason that militates in favor of a market-based solution.

27. Please describe any other issues that you believe policymakers should consider with respect to potential copyright liability based on AI generated output.

Nothing else on this point.

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 would support such a transparency requirement, as is done for example in the EU AI Act (not yet adopted). This would allow authors and citizens more generally to understand the impact of GenAI on the creation of literary and artistic content, and reduce harm associated with deep fakes. Absent a clear industry standard enforced by major platforms, it is difficult to see how this requirement could be applied uniformly without legislation.

29. What tools exist or are in development to identify AI-generated material, including by standard-setting bodies? How accurate are these tools?

There are numerous useful recitals about this issue in the EU AI Act.

Related to Copyright
30. What legal rights, if any, currently apply to AI-generated material that features the name or likeness, including vocal likeness, of a particular person?

The scope of publicity rights varies state by state, as does the term of protection and descendibility. There is no unform international norm, complicating international release of AI-generated material that features the name or likeness, including vocal likeness, of a particular person. Consent/licensing could solve cross-border issues, but there needs to be consideration, i.e. in this case a right to license.

31. Should Congress establish a new federal right, similar to state law rights of publicity, that would apply to AI generated material? If so, should it preempt state laws or set a ceiling or floor for state law protections? What should be the contours of such a right?

The uniformity that federal law would bring would seem desirable.

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?

Style is not protected by copyright. The contours of such a right are difficult to define. Perhaps a form of unfair competition would apply to specific cases?

THE END