Comments in Response to the
U.S. Copyright Office, Library of Congress,
Notice of Inquiry and Request for Comments
Regarding
Copyright and Artificial Intelligence (AI) Systems

Docket No. 2023-6, 88 FR 59942

Lee A. Hollaar

Emeritus Professor of Computer Science
University of Utah

hollaar@cs.utah.edu

October 24, 2024

Lee Hollaar is an Emeritus Professor in the School of Computing at the University of Utah. Before retiring in 2014, in addition to teaching courses in computer networking and hardware and software systems, he regularly taught computer intellectual property. He is the author of *Legal Protection of Digital Information* published by Bloomberg BNA (first edition: 2002, second edition: 2016), developed as a text for that course. He received his B.S. degree in Electrical Engineering from the Illinois Institute of Technology in 1969 and his Ph.D. in Computer Science at the University of Illinois at Urbana-Champaign in 1975. While a faculty member, he completed two years of law school at the University of Utah.

Professor Hollaar was on sabbatical leave in Washington DC during the 1996-97 academic year as a Committee Fellow in the intellectual property unit of the Committee on the Judiciary of the United States Senate, where he worked on patent reform legislation, database protection, and was the Committee's resident technologist during the development of the Digital Millennium Copyright Act.

He has been a technical consultant or expert witness in a wide range of computer-related litigation, including being principal technical expert in Caldera's antitrust suit against Microsoft and consultant to the states that joined the United States' antitrust suit. He has been a Special Master in a number of copyright and trade secret cases, assisting the courts in determining whether there was a colorable claim of infringement and in one case supervising the software discovery.

Question 5: Is new legislation warranted to address copyright or related issues with generative AI?

For special rules governing AI, you have to be able to precisely define what AI is. Until that can be done, there is little point in asking whether new legislation is warranted because it will be unclear who it might affect.

The definitions at the end of the Request may be sufficient to indicate the general area of interest, but contain far too many vague terms that may include things that are not intended. For example, it defines Artificial Intelligence as "A general classification of automated systems designed to perform tasks typically associated with human intelligence or cognitive functions." That could be stretched so far as to include a bookkeeping system if one considers a human bookkeeper as have some intelligence. Even the definition recognizes that it is problematically broad, because it goes on to say "This Notice uses the term "AI" in a more limited sense to refer to technologies that employ machine learning, a technique further defined below."

But those qualifications don't really help. Machine Learning is "A technique for building AI systems that is characterized by the ability to automatically learn and improve on the basis of data or experience, without relying on explicitly programmed rules." It appears that an AI system may be any computer program that has as least one aspect that does not rely on "explicitly programmed rules," whatever that might mean.

Decades ago, there was considerable research and a few actual systems for information retrieval that were based on "relevance feedback." They augmented their queries based on what users in the past had felt was relevant to a related query, sometimes determining that based on how long a user spent reading a retrieved document and how that document may contain different terms than one that users quickly skipped over. You see similar techniques being used by online stores to try to show users what most likely they are looking for. Is that being done "without relying on explicitly programmed rules"?

In patent law we are seeing the confusion cause by drawing ill-defined lines, particularly ones that don't have a strong technological underpinning. Patents are not available for "abstract ideas." See, for example, *Bilski* v. *Kappos,* 561 U. S. 593 (2010), *Alice Corp.* v. *CLS Bank Int'l,* 573 U. S. 208 (2014). It is fair to say that nobody knows exactly where those cases, or even follow-up cases in the Federal Circuit, draw the patentability line in any but the clearest cases. That is no way to impose legal requirements.

Not being able to draw a clear line becomes a problem for many of the following questions that suggest specific required actions to be taken by AI system developers. Question 15 suggests requiring developers of AI models to "collect, retain, and disclose records regarding the materials used to train their models." Question 16 suggests requiring AI developers go further by notifying "copyright owners that their works have been used to train an AI model." Question 28 asks about requiring "AI-generated material to be labeled or otherwise publicly identified as being generated by AI."

Technologists are clever people, who given a definition will either try to fall within it or outside of it as benefits them. Or, if they don't want to worry about failing to address any new requirements particular to AI systems, will follow those requirements even if they think they shouldn't apply in their situation, adding unnecessary deadweight to the country's economy.

During the development of the Digital Millennium Copyright Act, there was a concern of the burden it could place on "internet service providers" and there was an effort to find a way to exclude them. But that proved problematic because there didn't seem to be a definition that wasn't both under- and over-inclusive. So, instead trying to find that definition, Congress instead provided four "safe harbors" that limited monetary damages for copyright infringement when their conditions were satisfied. See 17 U.S.C. § 512.

It is far better to avoid requiring particular behavior based on a vague definition and instead legislate to specifically address particular problems.

Question 8: Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use?

As noted by Senator Orrin Hatch, the then-chairman of the Senate Committee on the Judiciary, which has responsibility for intellectual property legislation, in his floor statement when the Senate was considering the NET Act, "Congress has long recognized that it is necessary to make incidental copies of digital works in order to use them on computers." 143 Cong. Rec. at S12690.

It is best to consider the ultimate use of the copyrighted work when determining if that is fair. For example, if the "training" of the work results in an incidental copy that can later be supplied to a user as a substitute for the original work, looking at the fourth fair use in particular, that ultimate use is likely unfair. If, instead, no copy of the original work is retained after the training, as would be the case where it is being used to determine probabilities for the nodes of a neural network and there is no way to reconstitute the original work, that tends strongly toward fair use.

Courts have recognized the making of incidental copies as a fair use in the "reverse engineering" cases (*Atari Games v. Nintendo*, 975 F.2d 832 (Fed. Cir. 1992), *Sega v. Accolade*, 977 F.2d 1510 (9th Cir. 1992), and *Sony v. Connectix*, 203 F.3d 596, (9th Cir. 2000)). In finding fair use in those cases, the courts looked at the ultimate use of the reproduced copyrighted work, and finding that reverse engineering for purposes of interoperability was a fair use, held that the incidental copies that were necessary for that use were also fair uses. Congress ratified that position in the DMCA, where it stated that circumvention of a protection mechanism for purposes of reverse engineering was not a violation. 17 U.S.C. § 1201(g).

And when a copyright owner persuaded a court to consider the incidental copy made when a diagnostic program is loaded into the memory of a computer as an infringement so that independent service providers could not us it (*MAI Sys. Corp. v. Peak Computer*, 991 F.2d 511 (9th Cir. 1993)), Congress again recognized that the making of incidental copies is a fair use if the ultimate use of the copyrighted work is not an infringement.

> Title III of the bill amends section 117 of the Copyright Act (17 U.S.C. 117) to ensure that independent service organizations do not inadvertently become liable for copyright infringement merely because they have turned on a machine in order to service its hardware components.

Senate Report 105-190, at 21.

Just as with that case, the making of incidental copies of a copyrighted work during the training of an AI system should not be used to bootstrap protections that are not otherwise provided by copyright law.

For example, in the copyright infringement suit against OpenAI filed by the Authors Guild and a number of its members, the complaint contains a number of claims of copyright infringement. These for author John Grisham are representative:

> 198. For example, when prompted, ChatGPT accurately generated summaries of several of the Grisham Infringed Works, including summaries for *The Chamber*, *The Client*, and *The Firm*.
>
> 199. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The King of Torts*, one of the Grisham Infringed Works, and titled the infringing and unauthorized derivative "The Kingdom of Consequences," using the same characters from Grisham's existing book.
>
> 200. When prompted, ChaptGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Last Juror*, one of the Grisham Infringed Works, and titled the infringing and unauthorized derivative "The Juror's Dilemma," using the same characters from Grisham's existing book.
>
> 201. When prompted, ChatGPT generated an accurate summary of the final chapter of *The Litigators*, one of the Grisham Infringing Works.
>
> 202. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Grisham Infringed Works.

https://authorsguild.org/app/uploads/2023/09/Authors-Guild-OpenAI-Class-Action-Complaint-Sep-2023.pdf

Note that many of these things are likely fair uses, if they are copyright infringement at all, when done by a person. For example, Wikipedia contains “summaries” of all three of the works mentioned in 198. It also contains a summary of the work mentioned in 201, although not specifically the final chapter.

Simply because the training of an AI system makes incidental copies should not create new forms of copyright infringement, such as summarizing a work.

Question 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)?

This question seems to assume that the use of a work requires the permission of its copyright owner for it to be used to train the AI program, ignoring that it might be a fair use, as previously discussed. The portion of a work used may not even be protected if the originality of the work is in its organization and that organization is not considered during the training.

It may be an insurmountable task to obtain "affirmative consent" from a large number of copyright owners. Notice of copyright is optional, and even if there is a notice it most likely does not contain the information necessary to contact the current copyright owner (who may be different than either the author or the original copyright owner). Remember, we may be talking about thousands or even hundreds of thousands of works here. This may be a particular hardship for academic researchers who may not have the time, skills, or funding to solicit permission from every copyright owner.

Question 9.4 seems to suggest a new exclusive right to control a copyrighted work: to prevent the use of a work in the training of an AI program even if that use is not otherwise an infringement. And Question 10.3 goes on to suggest a collective licensing regime established by Congress to monetize this new right for copyright owners. But as mentioned above, without a clear definition of what "training an AI model" is, it is hard to see how such a requirement can be imposed.

Question 9.5 goes even further down the road of creating unheard of rights for a copyright owner in an AI context. It seems to suggest a new "moral right" of the original creator of information: objecting (and therefore blocking) their work for training an AI program, even though the actual owner of the copyrighted work is willing to permit it. Even if the United States fully recognized "moral rights," this suggestion goes well beyond the traditional rights of attribution, anonymity, and integrity that protect the author's honor or reputation. And even if such a thing were desirable, might be impossible to locate the original creator to ask permission.

And it is not clear whether "fair use" would trump such a right. In short, this is a really bad idea.

Question 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?

Given that it will be difficult to have a definition for "AI model" that is not substantially under-inclusive or over-inclusive, it would be hard to know if such a requirement applies to a particular activity, especially for academic researchers.

What are the consequences of not meeting such a requirement? If there aren't any, then it is just a "good practices" suggestion, which has no place in copyright law. As to the present consequences of not retaining such records in future litigation, see the discussion in Question 24 below.

Who should the information be disclosed to: Copyright owners of the information? Some government agency? Your competitors, if they ask?

Question 16 goes even further, asking if there should be an obligation of notifying every copyright owner that their works have been used to train an AI model. The same difficulty of asking for consent to use a work exists here. And if there is such an obligation, what is the effect of not fulfilling it?

Question 22: Can AI-generated outputs implicate the exclusive rights of preexisting copyrighted works, such as the right of reproduction or the derivative work right?

There doesn't seem to be anything in the copyright statutes that suggests that AI-generated outputs can't infringe the copyright in preexisting works. The question really is who you sue for infringement, which is Question 25.

Regarding derivative works, it is probably going too far to say that an AI model, and any output from it, is "based on" each and every work that went into training the model, particularly if there is no mechanism to determine which particular training works resulted in a neural network giving a result. The definition of a derivative work in 17 U.S.C. § 101 provides examples that are not at all similar to the result of training: "a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which a work may be recast, transformed, or adapted."

What these have in common is that the original work is discernible in the derivative work. A training that merely configures a neural network without retaining a copy of the original work is substantially different from that.

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

The existing Abstraction-Filtration-Comparison test commonly used in computer software cases also provides a good framework for more conventional writings, which should not be surprising because it was originally proposed by Judge Learned Hand as a way of determining whether a motion picture with a similar theme infringed the copyright on a play (*Nichols v. Universal Pictures*, 45 F.2d 119 (2nd Cir. 1930) and has stood the test of time.

There is no reason to have a different test that is applied based on an imprecise definition of when something is a generative AI system.

Question 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 standard of proof for each element of copyright infringement is simply the preponderance of the evidence. Before the filing of any suit, there should have been some indication that a copyrighted work was used in the training of the AI model. See Fed.R.Civ.P. 11(b)(3).

At that point, the defendant cannot simply state that the work wasn't used in training the AI model, but will need to present some evidence. And that will be difficult if they don't have some record showing that the work wasn't used for the training, especially if there are indications that it was.

Moving forward, the copyright owner will be able to get information from the alleged infringer either by document requests, written questions, or depositions of appropriate people, most likely under a protective order to preserve any trade secret or other rights of the defendant. And the court can impose sanctions for not being forthcoming, especially if it is shown that information from the training logging the documents used has been deleted, up to rendering a default judgment and imposing monetary sanctions.

This is not unlike the information that is needed for proving copyright or patent infringement or misappropriation of trade secrets by a computer program, and procedures for learning that information are well-established. Courts take these disclosure requirements very seriously, sometimes even appointing a technologist as special master to supervise the discovery.

From my experience as a consultant or special master in a variety of cases, I believe that the current civil discovery rules are more than sufficient and having special rules for AI systems, if such things could be defined, would only be a complication.

Question 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?

Although the copyright statutes do not have provisions for secondary liability, the Supreme Court has imported contributory infringement and inducement of infringe from the patent statutes (See *Sony Corp. of America* v. *Universal City Studios, Inc.,* 464 U. S. 417 (1984), and *Metro-Goldwyn-Mayer Studios Inc.* v. *Grokster, Ltd.,* 545 U. S. 913 (2005)). Unless experience proves otherwise, these cases and their progeny provide a good answer to the question.

Question 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?

It depends on how the work containing copyright management information is processed during the training. But in most instances, it appears to be quite a stretch to have an instance where 1202(b) is violated when there isn't conventional copyright infringement.

Congress included this provision as part of its addition of Chapter 12 to help the development of Copyright Management Information systems, particularly those developed in the future, by providing statutory protection to back up any technology protection for the system. It was written in general terms as not to dictate the future direction of such protection systems while not becoming obsolete through new technologies.

As such, it should be read narrowly and in light of its original purpose, not stretched to create new protections where Congress had not specifically intended.

Simply using a copyrighted work for training would not normally be regarded as "removing or altering" the CMI, especially if the training doesn't result in the storage for later use of the work.

And it is not at all clear how an AI program would "know" about the results of its actions, a requirement of 17 U.S.C. § 1202(b).

Question 28: Should the law require AI-generated material to be labeled or otherwise publicly identified as being generated by AI?

The problem with this suggestion is that not only isn't there a clear definition for AI-generated material, but how much of an identified document has to be generated. If I use an AI-based spelling, grammar, or cite checker, do I have to mark it? Do I even know if a tool that I used uses an AI model, however defined? Can I avoid having to label something if I change a word or two?

Because some generative AI systems seem to hallucinate court opinions that don't really exist, some courts are now requiring that AI-produced filings be identified. But that is a very special case, and is being addressed without any new statutory requirements.

This question again points out the difficulty with special rules based on whether something meets an ill-defined definition of "AI-generated material." The obvious way out of this difficultly is for everybody to label any material as "May contain AI-generated material," the same as many current "safety" warnings, in which case such a law is essentially useless.

Question 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)?

We are again back to the definitional problem of what is an "AI system," especially if a human is involved in the final creation of the output. Now we will also have to determine what the "style" of a particular artist is and whether something is in that style.

The copying of the style of an author or artist or composer is so common in a non-AI context that there are even special terms to describe it. If it pays homage to the works, it is a "pastiche." If it mocks it, it's a "parody." The Supreme Court, in *Campbell* v. *Acuff-Rose Music, Inc.,* 510 U. S. 569 (1994), held that producing a parody of a work was a fair use, even if it were done for commercial purposes.

It seems like the Artists Guild thinks that such a protection already exists. In the portion of their copyright infringement complaint included with my response to Question 8, paragraphs 199 and 200 are interesting because "the next purported installment" seems to be the outline for a sequel to the books mentioned "using the same characters." Such things are commonly called "fan fiction" and there is no reason why being produced by an AI system should change whether it in an infringement or not.

Summary

Without a precise definition of what is an AI system, it makes little sense to consider special requirements for such systems. Instead, if there are problems that develop that are not addressed by current law, Congress can enact appropriate and narrow legislation along with exceptions to retain traditional balances in protection as it has done in the past.