

Comments on Artificial Intelligence and Copyright
[Docket No. 2023-6]

Item: Background

I am a vanity-published author with a book properly registered with the copyright office. It has not been a source of significant income to me; total royalties have amounted to less than the price of registering the copyright.

The views presented here are mine, resulting from some 75 years of often voracious reading of varied printed materials and, more recently, of discussions of intellectual property and its assorted legal structures in various online fora.

Item: First Issue

"Training" of AI programs is indistinguishable from reading of published materials. As a result, "training" of AI should be considered not only fair use, but proper use, with the proviso that said materials shall have been obtained by legal means.

When an author - I shall limit my discussions to written works here -- puts his work into the marketplace, he generally does so with the expectation that it will be read. Motivations may be for recognition or recompense, but the expectation nonetheless is that the work will be read, and, it is to be hoped, discussed. It seems to me that "training" and AI is not conceptually different from reading or indexing, or otherwise making normal use of the material.

Item: Second issue

I am sympathetic to the doctrine that "copyright protection in the United States is limited to works of human authorship," and that should remain the guiding principle. A computer program is not a person and should not have the rights of a person. The program should also not have the responsibilities of a person, though the person or persons who use the program should bear all of the responsibilities -- and liabilities -- of authorship. The person(s) who wrote and originally trained the program should bear responsibility for any inbuilt errors, biases, etc., but those issues would not generally be in the purview of copyright oversight.

Item: Third question

"Copyright liability principles" could hardly apply to the output of an AI program unless or until said output is "published," i.e., disseminated, by whatever means, into the public arena, in which case the person or persons responsible for such publication shall be exposed to the same liabilities as though they were the originators of the (offending) material.

Item: Mimicking

As noted in the Inquiry document, this area is probably not the proper purview of the Copyright Office. It is an area that demands legislative action.

General Questions

1.

Generative AI is a tool. As with any tool, it can be a valuable adjunct, or it can be misused and abused.

Used carefully, it can be a boon to writers, and possibly artists, in broadening viewpoints and getting one "off dead center" when the creative urges go slack. It could be a tool functioning, as apprentices once did, to offload some of the drudgery associated with creation.

Probably the greatest risks are in the area of fraud. Stories fraudulently attributed to a known author, erroneous information -- the issue of AI-generated mushroom-hunting guides leaps to mind -- presented as being from a reliable or authoritative source, active disinformation about a public issue or public figure, presented as being from authentic source -- this last a real danger because of the ability to mimic -- all are detrimental uses being practiced by unscrupulous operators. And it is very early in the use of generative AI programs at this time.

Real creators will probably find it a useful tool at times. Copyright owners will probably see an uptick on infringement because there is a much lesser requirement of effort on the part of the infringer.

Technology developers will probably see little effect from the current crop of programs, except as they may make some development easier and less remunerative.

The public will see even more BS pumped into the airwaves, over the internet, and onto bookshelves than they do now, mostly because the BS generators can get more done with less effort in a given time. As for quality and reliability of content, well those are pretty low already. But quality of content has not historically been, nor probably should be, the concern of the copyright office.

2.

I am no longer in industry, having retired nearly 15 years ago. Creators will likely see more competition, or at least apparent competition, because of the ease with which content can be generated with these programs. And creators can expect to see more fraud around their names, story lines, fictional characters, etc., and even more poor-to-bad code being promulgated.

3.

I have no knowledge in this area.

4.

I am not aware of any potentially useful regulation or legislation being proposed anywhere to deal with this subject matter.

International consistency would be nice, but it is difficult to imagine the USA, Great Britain, France, Germany, Russia, China, the Arab nations and the Third World getting together on any agreement that would impose restrictions that I personally would find acceptable or desirable.

5.

Yes, new legislation is warranted. And not just in regard to generative AI; the whole body of IP regulation is a shambles.

Specific elements I would consider crucial would be that AI is a tool, not an actor, and that the user should bear all onus for any abuse or misuse of the tool.

6.

These questions deal specifically with the training of AI models, and I have no specific knowledge thereof. That said, I do not see any logical or philosophical difference between an author reading a number of copyrighted works to establish a knowledge base and background for his work and an AI reading a variety of works to establish a base for its work.

As an author, I feel that when a creator presents his work in the marketplace, he gives up the absolute authority he had when the work was held privately. The Law does, and should, delineate certain authorities over the disposition of his work once it has entered the agora, but those authorities should no longer be absolute, no matter what well-funded rights holders claim.

7.

I have no direct knowledge of how AI models are trained.

8.

Any and all use of copyrighted materials in training is fair use if the materials have been obtained legitimately.

The "purpose and character" of the works are irrelevant in training.

The inputs to the AI program should have the same protections, and no more, than the inputs to the education of students. It is the output from the programs that needs to be regulated. An AI that has been trained on, say, a body of copyrighted fiction, to generate organizational information on the body of fiction for the private use of the program user should be completely protected as fair use. Similarly, an AI-generated business model used internally in a corporation should be protected, so long as the training data were acquired legitimately, and so long as the output of the model is not disseminated.

9.

So long as the trained model makes no reference to the copyright holder and does not attribute compliance, connivance, or participation by the copyright holder, use of such materials should be fair use, with the ongoing caveat of legitimate acquisition of the copyrighted materials.

If I read books by Peter, Parkinson, Deming and Brooks, then restructure my business -- or my career plans, for that matter -- do I owe anything to those authors, apart from not having stolen the books in the first place? And if not, why should establish a basis of concept in a computer program be any different?

10.

Copyright holders' consent should not be required in the training of AI models.

11.

The question highlights the futility of the issue. Copyright maximalists would say that consent should be obtained by each of these players, and any others who might be involved in the development and use of the programs. There is a strong tendency for such to be followers of Gordon Gekko. I am not one such, and I am not sympathetic to those who are.

12.

Given the size and complexity of the training-data sets used to train LLM AI models -- the variety that seem to be the focus here -- it is highly

unlikely to be able to discriminate between a particular source and several other that present similar or parallel arguments or propositions. Edge cases where such isolation of cause-to-effect can probably be handled by plagiarism rules.

13.

Given the monopolistic pricing practices of the typical corporate rights holder, the effect would be to stifle most academic and small-player development. Only the large, well-funded developers would be able to participate.

A licensing model that gave the rights holder an interest in the product -- the sort of thing that might appeal to a small developer whose pockets were not deep, could quickly raise anti-trust and conflict-of-interest issues if the rights holder were approached by another, possible competing, developer.

14.

I do not believe that there are, or should be, any copyright liability for training AI models, apart from the requirement that the data sets be acquired legitimately.

15.

1. Good practice would require the collection and retention of comprehensive lists of sources used in training AI models. With modern storage being relatively inexpensive, retention of the entire training dataset would be a good thing, but might reasonably be done by reference, minimizing the need for multiple copies of large datasets. The inputs to the operating program -- I think they refer to them as hints; in my old field they would have been constraints -- should definitely be retained. Putting such files in escrow upon release of the product should be adequate, with, perhaps, updates in the case of revisions. Such files should be considered proprietary information and protected accordingly.

2. Disclosure should be protected under Fourth Amendment terms.

3. Developers importing third-party models should, as a minimum, receive the comprehensive listing of inputs.

4. Costs are likely to be moderate, as the necessary infrastructure of ancillary programs and hardware should already be in place for the development of the AI code. Handling the records as archival, rather than ongoing, should cut down on costs, but it won't be free.

Escrowing files would not be free, either, but it would protect against loss of information in the not unlikely case of a company going out of business or being acquired by another company that did not have direct interest in the history of the AI model. The National Archive or private groups like Iron Mountain should be able to give fairly accurate estimates of the costs involved. Sensitive files should be encrypted, with encryption keys held separately.

16.

There should be no obligation to inform copyright owners of what should be considered fair use, any more than there should be an obligation to write a fan letter to an author whose book you read.

17.

This is not an area of my expertise, but I have no recollection of having seen mention of pertinent or applicable laws.

18.

If someone using an AI tool present themselves as the author of the work, then they should be recognized as the author, with the understanding that they thereby assume all liability for any errors or misrepresentations in the work.

The institution of the ghost writer is well established in literature. In this case, the ghost is in the machine.

19.

I don't see where AI-generated material should have special status vis-a-vis any other-sourced material. AI-generated material should only be copyrightable when presented by a human "author."

20.

See 19 supra

21.

Only an attorney could construe "authors and inventors" to include computer programs. One might as well argue the rights of saw and chisel in the use of furniture.

22.

An AI is a tool. AI-generated works do not exist in the marketplace in the absence of human involvement in placing them there. The outputs should come under precisely the same rules as those for conventional human-generated works.

23.

Rules for determining infringement may need to be changed, but not for AI-generated material.

24.

If by "copying" you mean "plagiarism," there already rules in place. Demonstrating access is fatuous in the presence of public libraries. If records were kept in accordance with 15 supra, then existing rules of discovery should be adequate.

25.

Infringement is the responsibility, and liability, of the person (or corporate entity) that introduced the work to the agora.

26.

??

27.

Whosoever shall introduce AI-generated materials to the agora -- the marketplace, the public forum, including internet, media, streaming services, etc. -- shall assume all liability for the content of such materials; use of AI shall not be a defense. Releasing AI-generated output to the agora anonymously or using a fraudulent identity should be made a crime -- probably a felony.

28.

I would prefer labeling, either prefatory or postscript, as is routinely done by Project Gutenberg or on code that has been released under GPL or similar license. Compulsory labeling may become problematic because of the potential difficulty of demonstrating AI authorship. If the practice were to become common, prepending a logotype or emoji-like symbol should be adequate.

1. First responsibility should lie with the entity which introduces the output to the agora. Further responsibility would lie with anyone quoting, "liking," redistributing, or otherwise promulgating the content.

2. I don't see any technical barriers. My focus has been on the written word, but images could have title-block-like notifications; audio works could have a pepended tone sequence.

3. Failure to label is a tough one. Given that I propose that the nominal author assumes all responsibilities and liabilities, failure to label should only be an issue when there is apparent intent to defraud. Removal of label, after the original author has decided that the work needs to be labeled, is more clearcut as a violation and might be considered as intent to defraud

29.

At this stage I have seen no announcements of even reasonably reliable tools to identify AI-generated content.

30.

"Imitation is the sincerest flattery." Representation as being authentic content from the person whose likeness, etc., is presented should constitute fraud. Presentation as AI-generated replica of the original, being factual, probably should not be, per se, a violation, unless there is something in the content that would insult the original person, bring them into disrepute, or otherwise cause detriment to their image. With the caveat that certain classes of satire or lampoon may be protected speech.

31.

If I had faith in their ability to craft reasonable rules, I would advocate for Congress to pass such legislation. I would say that it should set a ceiling for rights and a floor for responsibilities. both on the part of the AI user and on that of the holder(s) of NIL rights in question.

32.

So long as there is no attempt to pass off the outputs as being created by the human creator in question, or their associate, apprentice, employee, etc., such output should not constitute a criminal matter. If rule were to be put in place, they should apply to any author, artist, performer, etc., who has achieved name recognition in the marketplace.

33.

I have no expertise or compelling opinion in this area.

34.

I would reiterate that the Office needs constantly to keep in mind that the AI models in question are tools, not actors or agents, and that any motivations or judgments lie with the user, or, at one (or more) remove the programmer(s) of the model.

Generative AI has great potential in the hands of ethical users to enhance

the arts experience of the consumer. As potential examples, consider portraits of fictional characters done in the style of famous story-contemporary artists. Or consider the art of the pastiche -- a story set in the milieu of a particular fictional character or series, composed (ideally) in the style of the milieu's originator; the current crop of Sherlock Holmes pastiches comes to mind. An AI could be used to generate a first draft of a story which the author would then rewrite, emend and edit as though it were self-composed, as indeed it could be considered to be, given the necessity of describing the outline, premise, characters, etc., to seed the AI's generative efforts.

Generative AI also represents a great potential hazard in terms of fraudulent art and of disinformation presented as being from a reliable or authoritative source. Some of the latter has already happened, as described in accounts of bogus travel guides and reference works being offered on Amazon a similar platforms. Things like misinformation on identifying wild mushrooms can constitute a real physical danger to the naive consumer. There is also the everpresent danger of diluting the income potential of authors and artists by flooding the market with works that purport to be from them, competing for consumer dollars.

As the quality of the output from these programs improves, it will become more and more difficult for the consumer to judge what is reliable and what is not. Measures put in place now to limit the abuses of the technology could go a long way toward protecting the consumer in the future.

I would suggest that the Office actively pursue input from the Electronic Freedom Foundation and the authors/editors of Techdirt, as they have had much exposure to the issues at question and are broadly and well informed on the subject matter. My personal prejudices are not always precisely in line with theirs, but I respect their opinions.