Comments on Copyright Office Docket 2023-6 from Software Freedom Conservancy

These comments come from Software Freedom Conservancy and were prepared by our Policy Fellow, Bradley M. Kuhn and Director of Compliance, Denver Gingerich.

Software Freedom Conservancy ("SFC") is a nonprofit charity centered around ethical technology. Our mission is to ensure the right to repair, improve, and reinstall software. A key method and strategy in that mission is *copyleft* licensing. Copyleft is a form of reciprocal copyright licensing that does not seek financial recompense in exchange for permission to engage in activities controlled by the exclusive rights granted to copyright holders. In contrast to most copyright licensing, copyleft instead places requirements on those who wish to engage in such activities. These requirements ensure others can, in turn, "promote the Progress in Science and the useful Arts" through the reciprocal provision of those permissions. Copyleft licenses mandate that licensees receive both the permissions and methods to engage in creation and dissemination of improvements to the works of authorship. Copylefted software, licensed under well-known licenses such as the General Public License (GPL), is an essential component of key parts of our technology, including the well-known Android system for mobile devices, and the Linux Kernel used in most servers on the Internet today.

We have serious concerns that non-traditional copyright licensing, such as copyleft, has not been fully considered in the present debate and discussion regarding machine learning training and automated content generation. Incumbent, powerful, for-profit, multinational technology companies ("Big Tech") have been disingenuous in their assumptions and descriptions of these new technologies. We suspect many copyright holders in the entertainment and arts areas will agree with us on that point. However, because of the unique nature of copyleft to create a shared, protected commons with egalitarian rights for all, we understand that our views on remedies for the current situation will differ from most. We respectfully ask the Copyright Office to give careful consideration on the impacts to the copylefted commons, which includes not only Free and Open Source Software ("FOSS") (which is our area of expertise) but also third-party projects such as Wikipedia and OpenStreetMap. Most notably, any compulsory financial licensing system fundamentally fails as a solution for a copylefted commons when used as part of Training Datasets for machine learning.

Below we answer a select group of the questions posed in the Copyright Office's Document Number 2023-18624:

Question 3

> *3. ... Please identify any papers or studies that you believe are relevant to this Notice.*

Bradley Kuhn and SFC previously published a paper, entitled *If Software is My Copilot, Who Programmed My Software?*, which is relevant to many of the questions asked in this inquiry.

We also believe that Microsoft's own document regarding the creation of their Github Copilot product, entitled *GitHub Copilot research recitation*, interestingly admits to a number of situations where their product does literal copying from the Training Material into the output material — and as such clearly infringes copyrights in the Training Dataset.

Question 9 and 9.3

> 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)?*

and

> 9.3. *What legal, technical, or practical obstacles are there to establishing or using such a process? Given the volume of works used in training, is it feasible to get consent in advance from copyright owners?*

These questions speak to one of the many examples in this discussion where copyleft licensing is unique. The creation of copylefted software relies on the human collaboration and ingenuity of public sharing and review of the source code of the work — which is usually created in public with many contributors over long periods of time.

The process of sharing, mandated by copyleft licenses, has led to the most reliable software in the world today. Inspired by the successes in copylefted software creation, communities like Wikipedia have used this style of licensing in areas beyond software. Copyleft licensing, in almost any area of human creative endeavor governed by copyright, provides an egalitarian system for collaborative authorship.

With machine learning and generative systems, Microsoft, OpenAI, Meta and other "Big Tech" companies believe they have created the ultimate "copyright washing machine" — whereby an automated system takes as input terabytes of copyrighted material, much of which is licensed under copyleft licenses such as the GPL, and allows users to output material that pays no mind to the licensing terms of the Training Dataset that is the main and essential component of their system.

Big Tech's position has been quite uniform in this regard: they have argued that a machine learning system, and its output, never infringes copyright of the Training Dataset. They argue (in the alternative) that if such copyright infringement does occur, it is a forgone conclusion that the reproduction of works from the Training Dataset in their output is fair use. We believe this position is facile, as do the many who have filed lawsuits alleging copyright infringement by the machine learning systems already in production.

Copyleft licenses (such as the GPL) do put clear requirements on both the copying and reproduction, as well as the preparation of derivative works. Specifically, copyleft licenses require that all copies, reproductions, and derivative works must be licensed under *precisely the same terms* as the copyleft license itself.

In the simplest case, any reproduction or copy of a copyleft licensed work must carry with it a copy of the license.

Thus, copylefted works are already in a strange position regarding use in machine learning training. The works have already been broadly licensed for many uses and improvements, yet Big Tech is arguing — without case law or evidence — that these licenses already provide all the necessary permissions. Big Tech argues further that the copyleft requirements to contribute back to the commons do not apply to their uses. Rather than engage in a good-faith debate about questions of fair use, Big Tech is instead arguing that the rights holders in the copylefted commons already gave them permission.

It is not surprising that there are multiple lawsuits pending related to these uses. Lawsuits require significant resources, and unfortunately the cases that have been brought to date prioritize financial compensation rather than the development of materials for the public use as copyleft licenses intend.

Question 6.2

> 6.2 *To what extent are copyrighted works licensed from copyright owners for use as training*

materials?

The creators of the Copilot system (a Generative AI system that generates software source code, and which is a product of Microsoft's GitHub subsidiary) readily admit that copylefted software is present in their Training Dataset. They further admit that they have programmed their system to explicitly *remove* copies of the GPL itself that might appear in the output, but not the material that is licensed under the GPL. Ironically, if you'll forgive the anthropomorphism, even the non-human Generative AI itself seemed to "want" to include the copyleft license in its output, and its creators must explicitly hold the system back from exposing that copylefted materials are in the Training Dataset. We believe this unsurprising but rare admission is a tip of the proverbial iceberg, since the largest Training Materials available publicly *are* copylefted materials (including the body of copylefted FOSS, plus Wikipedia and OpenStreetMap — to name but a few).

In addition to creating copylefted software ourselves, we regularly consult and cooperate on collaborative projects with many authors of these works. To our knowledge, these authors are never consulted nor provided an option to license their works for use as Training Materials. Rather, the companies who use their works as Training Materials claim fair use with no evidence, and do not follow the terms of the reciprocal copyright licenses the authors have chosen to use.

Question 6.3

> *6.3 To what extent is non-copyrighted material (such as public domain works) used for AI training?*

It is indeed worth noting that, at least with regard to software, there is almost no software in the public domain at all. Furthermore, the only useful form of the software suitable for and useful for Training Datasets is its human-readable source code written in programming languages that are still in active use. The copyright would be unlikely to have expired on any such works known to exist. Most importantly, most of the works of source code publicly available for use as Training Material will be copyrighted works licensed under FOSS licenses such as a copyleft license.

Question 6.4

> *6.4. Are some or all training materials retained by developers of AI models after training is complete, and for what purpose(s)? Please describe any relevant storage and retention practices.*

We are extremely concerned that there is no ability to "trust but verify" answers to this question that might be provided by those who produce models.

We believe that only strict provenance requirements — so that developers of AI Models must indicate precisely which works are in all Training Datasets used by that model — would allow verification of any claims about these factors.

Question 7.4

> *7.4. Absent access to the underlying dataset, is it possible to identify whether an AI model was trained on a particular piece of training material?"*

We are quite certain the answer, today, is unequivocally "No". SFC convened a Committee On AI-Assisted Programming and Copyleft, which included some notable and accomplished individuals in both industry and academia, and we further carried out listening sessions with many other researchers. We have consistently found that the current state of research in attribution for AI Models in its infancy. Today, it is impossible —

given access only to a "prompt" (or other end user interface) — to determine if that model was trained on a particular Training Material.

Given the current state of research in this area, the only way to know what Training Material might have been used to produce outputs of an AI model is for that provenance information to be created at the time the Training Dataset is determined, and for that provenance information to be delivered to users of the system.

Failure to provide such information leads to many bad outcomes. For example, in software source code based training, we have seen clear examples where you can easily prompt Microsoft's Github Copilot to produce someone's name who is a known contributor to FOSS. However, in cases where someone is a prolific contributor, it's logistically impossible to trace back and determine *how* their name ended up appearing in output. This has serious privacy concerns that are not easily addressed holistically, since the Training Material that caused someone's name to appear in output is not readily apparent.

Question 22

> *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?*

We have seen many examples where individuals using Microsoft's Github Copilot serendipitously generated what was clearly a work governed by the terms of the GPL. Almost as quickly as those results are made publicly known, Microsoft's GitHub adjusts their back-end system so that it will no longer reproduce that particular work from any prompt. Not only does this create a bizarre moving target game for those who seek to show their rights are infringed, it also completely disproves the universal position by Big Tech companies that these systems by their fundamental nature do not generate infringing works. If the output works were indeed not infringing, why would Microsoft and Github need to repeatedly modify Copilot's back-end so that it does not generate those works?

Question 24

> *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?"*

Copyleft rights-holders face an extremely labor-intensive process to demonstrate (or even discover) infringement. Even when infringement is suspected, only manual, subjective means (and sometimes even reverse engineering) are necessary to prove that infringement.

This is among the reasons why we call for accurate provenance information regarding Training Datasets to be provided to the users of all machine learning systems.

Question 27

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

Given this "arms race" that Big Tech has already engaged in, we believe that it will become increasingly difficult to detect and adjudicate infringement claims, not only because of technical constraints, but also due to resource constraints.

Even by human authors, we regularly see violations of copyleft licenses, and our organization works arduously to resolve as many as we can and protect the rights that copyleft licenses grant to consumers. Since copyleft is not financially motivated, we fund this work through donations and grants. We are concerned that if there is not substantial regulation on the use of copyrighted works (and, in particular, copylefted works) in Training Datasets, we are likely to see explosive growth in infringement.

Furthermore, the commons of copylefted materials (such as software and the large Wikipedia encyclopedia) continue to receive contributions and improvements because of the contractual requirements that the copyleft licenses place on contributors: namely, they must make the work available under the same license and grant their users the same rights. We are gravely concerned that infringing AI-generated output may eviscerate copyleft's unique aspects.

If Big Tech's prevailing incorrect "it's all fair use" interpretation of AI-generated output is even partially adopted, this key tool of copyleft — designed to encourage egalitarian collaboration and assure advancement of science and art by requiring publication of means and methods, with permission to remix and reuse through contractual obligations in a copyright license — will evaporate into the past. With it, ironically, the copyleft incentive system that *created* the very software that made Machine Learning systems possible will, itself, also evaporate! After all, nearly all these machine learning systems run atop the copylefted kernel, Linux — which would not exist if not for those very licensing terms that required everyone — from the individual hobbyist to these same Big Tech companies — share their changes and improvements with everyone.

Big Tech is moving quickly toward their policy goals. Copyleft authors already face an incredibly well-resourced industry that has sent us the clear message that our licensing has been circumvented and we have little choice but to adjudicate our claims in Courts. We can simply say that we are flabbergasted at the brazenness of their position, and we hope the Copyright Office and Congress will rein in their overstep rather than leave the less powerful folks with no choice but expensive litigation to defend the copylefted commons.

Question 12

> *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.*

In an automated and disciplined fashion, no. Attribution — traced from Training Material to Generative AI output — is not currently possible, according to academic researchers and research literature. Currently the only way to determine this "degree of contribution" is with subjective, *human* analysis after the fact. Such analysis would also require unfettered access to the Training Dataset.

With access to the Training Dataset, one can at least determine which works *possibly* contributed to a given AI Model's output, and can engage in subjective analysis to determine similarities between those works in the Training Dataset and the output.

Question 15

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

Yes. We believe that all such actors should have a full chain of custody information and entire records of the Training Dataset, and removals and additions to it in subsequent training.

Question 15.1

> 15.1. *What level of specificity should be required?*

At the very minimum, a URL or other unique identifier for each and every work in the Training Dataset should be provided to all users. If the work is not available at any URL, a complete title and author list that uniquely identifies the work should be provided.

Question 15.2

> 15.2. *To whom should disclosures be made?*

Many people will have reason to access the Training Dataset — users, copyright holders, and people impacted by the use of the models widely and in surprising circumstances. The use of these Training Datasets is so widespread that the materials should be made as widely available as possible. The public deserves the right to know how key infrastructure works.

Question 5

> 5. *Is new legislation warranted to address copyright or related issues with generative AI?*

Past copyright legislation has favored the views of for-profit rights holders. We believe, for many reasons (both related to machine learning and many other issues), that recent changes in the software industry warrant a substantial overhaul of copyright rules. We, and other publicly focused charitable organizations, would gladly participate in that process to bring the unique perspective of software rights activists to balance the likely heavy lobbying effort that Big Tech has already brought to bear on Congress on copyright.

Question 10.1 and 10.3

> 10.1. *Is direct voluntary licensing feasible in some or all creative sectors?*

and

> 10.3. *Should Congress consider establishing a compulsory licensing regime?*

We are extremely concerned about the possibility of mandatory compulsory licensing regimes based on financial compensation — be they implemented voluntarily by industry or by Congress. Copyleft is unique in its design to place controls on copyright-governed activities that seek the betterment of society rather than financial compensation. The remuneration that copyleft authors seek is not pecuniary, but rather preservation of the reciprocal copyright license terms in any future uses of the work. Given the Constitutional motivations of copyright, we believe the copyleft system fits well in creating incentives and requirements other than financial ones for advancement and improvement to key works of authorship that benefit society. In other words, clear evidence shows that copyleft, at least in the scientific and artistic realms of software and encyclopedia creation, has done a better job to promote Progress than financially-based systems. We fear compulsory licensing regimes will dismantle these copylefted ecosystems.

For example, we are aware of the lawsuit filed by Joseph Saveri Law Firm, LLP and Matthew Butterick related (in part) to the inclusion of copylefted works as part of the Training Dataset of the Microsoft's Github Copilot. We have both publicly and privately raised concerns with the Plaintiffs' lawyers that they have, indeed, prioritized the wrong issues with regard to the public software commons. Ultimately, this litigation seeks financial compensation for the Plaintiffs (and their lawyers) and not actual enforcement of the copyleft provisions of the licensing of the Training Dataset. As such, we feel that the litigation may erroneously push the Courts to dissolve copyleft requirements into financial recompense. In our view, the proper compliance with copyleft licenses has benefits that simply cannot be compensated or adjudicated merely with financial payments. We fear the outcome of this class action lawsuit could yield a de-facto financial compulsory licensing scheme for copylefted software, which would eviscerate copyleft entirely.

While compulsory financial licensing succeeds for many authors, it yields an unwanted outcome for authors of copylefted works. Many of these authors were compensated already for their time in creating the copylefted work. Having already been financially compensated once, they chose copyleft licensing to receive, back into the public commons, any improvements and innovations on their work — to facilitate and accelerate their own future innovations. The latter contains value for their careers and society that cannot be quantified financially. Furthermore, this type of non-monetary licensing requirement directly fits with the original motivations for the creation of the copyright system. We ask that such forms of licensing be considered carefully by the Copyright Office and Congress when changing copyright rules and regulations.

Finally, we realize that compulsory licensing schemes are sometimes agreed to by industry. As such, we want to draw particular attention the software industry and the dangers of self-regulation on this issue. In many ways, copylefted software is a counter-culture movement to create software via facilitated collaboration by hobbyists, small businesses, and Big Tech alike — all placed on equal footing through the copylefted terms. We fear that any discussion of compulsory licensing will fail to take into account the views of the smaller operators in this large community, and we already know that Big Tech has sought to eviscerate copyleft for some time.
As such, we cannot trust traditional software industry trade associations and companies to properly look out for the concerns of authors of copylefted works.