Reply Comment on Copyright Office Docket 2023-6 from Software Freedom Conservancy (SFC)

This reply comment was prepared by SFC's policy fellow, Bradley M. Kuhn.

We see that the Copyright Office has received an overwhelming response to 2023-6, and we are very sympathetic to the staff's ability to review and consider so many comments. As such, we do not wish to belabor with an extended reply comment. We focus narrowly in our reply on the comments submitted by Microsoft (including arguments on behalf of their GitHub subsidiary) — found in comment ID COLC-2023-0006-8750.

We would like to draw the Office's attention to the self-contradictory position from Microsoft regarding expansion of copyright law. Microsoft has historically supported the most maximalist interpretation of copyright law. Yet, in this narrow area, Microsoft jumps to an unsupported conclusion of "fair use" for its novel uses of copyrighted works in Training Sets. In essence, Microsoft asks us all to simply ignore the existing rights of copyright holders who are already bringing forward infringement claims. The proper place for Microsoft's "fair use" claims is as an affirmative defense before a Court. To concede Microsoft's "fair use" claims would be the first step in eviscerating the copyleft licenses that protect the primary commons of software source code, which, in turn, comprise much of the software in Training Sets already in use for these Generative AI systems.

Our community designed "copyleft" licensing as a counter-measure against ever increasingly expansive copyright law for software. Namely, as explained in our original comment (ID COLC-2023-0006-9052), copyleft utilizes existing copyright law to ensure rights of consumers. Thus, every additional restriction and power granted to authors under copyright is utilized by copyleft to increase users' rights. Some say this in a pithy way: "rather than All Rights Reserved, copyleft assures that All Rights are Reversed [in favor of the consumer]."

This is, of course, inconvenient to Microsoft with regard to their GitHub Copilot product (in particular). Microsoft admits that they incorporate publicly available materials into their Training Sets, but fails to note that these public works have licenses with very specific compliance requirements that Microsoft has completely ignored. Indeed, Microsoft's own research recitation admits that large amounts of their Training Set included *copylefted* source code.

Meanwhile, we ourselves (the SFC) are a copyright holder in some of the copylefted works believed to be in Copilot's Training Set. Despite having reached out to them, Microsoft has simply refused to substantively discuss our concerns of copyright infringement of our works. Others in the Free

and Open Source Software ("FOSS") community report the same experience. Microsoft ignored the concerns and inquiries of rights holders, and now promulgates a dubious interpretation that would give them blanket permission to ignore rights holders forever. We remain concerned that many different aspects of Microsoft's GitHub Copilot system may violate copyleft licenses and infringe our copyrights.

Furthermore, we find Microsoft's argument that "any requirement to obtain consent for accessible works to be used for training would chill AI innovation" disingenuous. When the FOSS community began our work to create the large body of copylefted FOSS (that now even Microsoft relies on!), we could have easily said that our ability to innovate with FOSS was "chilled" since Microsoft withheld its source code from us for its operating systems and applications. Microsoft would not (and still does not) give us consent to license those works under our copyleft licenses, or concede any "fair use" right for us to do so[1]. We could have argued that Microsoft withholding that consent was chilling our innovation (because it *was* — the work to create the large body of copyleft FOSS would have gone much faster if we'd had permission to incorporate Microsoft's source code). However, no one in the FOSS community made this argument to the Copyright Office (or to anyone else) because we expect that the Office would have found that argument as nonsensical then as Microsoft's similar argument is today.

While we can all debate whether the existing copyright system is effective in "promot[ing] the Progress of Science and useful Arts", Microsoft (and similar Big Tech companies) should not get to have their cake and eat ours instead. The copyright rules and requirements that Microsoft claims will "chill innovation" are the same rules that they argue, with regard to their own source code for products such as Microsoft Windows and Microsoft Office, *allow* Microsoft to innovate.

Microsoft meanwhile asserts that copyright licenses chosen by authors *using* Generative AI systems should be respected, but not so for the copylefted works in the Training Sets of its licenses and protections. In our initial comments, we expressed our fear that companies like Microsoft sought to 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 — and instead chose their own proprietary licenses for that output. We are actually surprised that Microsoft so blatantly proposes just that in their comments. Ironically, Microsoft says, without the extension of copyright protection on output (and the removal of copyright protections on input) that "the commercial viability of the works made using AI tools is undermined". In fact, Microsoft's proposal undermines the licensing of the public commons — including software such as Linux and other software at the heart of today's technology.

---

[1]Microsoft touts that its "work in this space includes:. . . Tens of thousands of software titles, including some of the world's most popular productivity tools, developer tools, operating systems, and apps . . . [and] Thousands of video games, representing some of the most popular entertainment titles in the world." To our knowledge, the source code for these works has *not* been included in the GitHub Copilot Training Set, while many of our own (and other FOSS authors) works of software source have been. We have asked Microsoft many times to explain, if the Training Set is best served by inclusion of many works of source code, why Microsoft remains skittish to include its own works of software source code in its Training Sets. Microsoft has not answered.

Microsoft proposal seems like the opening position for negotiation, not a serious proposal for a fair compromise between individual rights holders and Big Tech. In that context, we reiterate a key point from our original comment: a compulsory licensing regime for Training Set inclusion, which appears on the surface to be a reasonable compromise between Microsoft's "all uses in the Training Set are fair use" position, and the position of most copyright holders alleging infringement, remains highly problematic regarding most works available in the public commons. Given that Microsoft admits that the public commons (much of which is under copyleft licenses) is the most important resource for Training Sets, we have even further trepidation regarding that compromise, as it would eviscerate copyleft's key rights-extending provisions. In essence, a compulsory licensing regime would also be a unilateral "free pass" to ignore the copyleft terms chosen for works by individuals, hobbyists, and small companies — just like the fait accompli "fair use" that Microsoft is already promulgating.

While we and other FOSS activists might support a full reconsideration of copyright rules for software from the ground-up, we do not think a piecemeal reworking of some rules in some contexts, particularly to merely serve the interests of large corporations, is in the interest of authors who do not have Big Tech's resources. Such changes would be particularly toxic to those of us who have chosen to license our copyrights under copyleft licenses, which were specifically designed to assure full transparency and the complete sharing of source code.

Finally, we would like to draw the Office's attention to additional comments submitted by the Federal Trade Commission in comment ID COLC-2023-0006-8630. The FTC's comment includes a transcript of a panel that SFC participated in earlier this year. Our comments on that panel provide a supplement to our main comment submitted in COLC-2023-0006-9052. We thank the FTC for passing along our additional comments to you in that form, and thank the Copyright Office for your careful consideration of these important issues.