search ...

Navigation

Navigation

- Home (https://juliareda.eu/en/)
- My Vision for Europe: Borderless (https://juliareda.eu/my-vision-for-europe-borderless/)
- EU copyright reform (https://juliareda.eu/eu-copyright-reform/)
  - Overview (https://juliareda.eu/eu-copyright-reform/)
  - Extra copyright for news sites (“Link tax”) (https://juliareda.eu/eu-copyright-reform/extra-copyright-for-news-sites/)
  - Upload filters (https://juliareda.eu/eu-copyright-reform/censorship-machines/)
  - Text and data mining limited (https://juliareda.eu/eu-copyright-reform/text-and-data-mining/)
  - Remixing & Freedom of Panorama (https://juliareda.eu/eu-copyright-reform/parliament-additions/)
- Projects (/)
  - The FOSSA project: improving Free Software security in the EU (https://juliareda.eu/fossa/)
  - End Geoblocking (http://endgeoblocking.eu)
  - Dieselgate investigation (http://dieselgate.juliareda.eu/)
  - The Reda Report (https://juliareda.eu/copyright-evaluation-report/)
    - Overview (https://juliareda.eu/copyright-evaluation-report/)
    - My draft report explained (https://juliareda.eu/copyright-evaluation-report-explained/)
    - Final report as adopted (https://juliareda.eu/copyright-evaluation-report/full/)
- Publications (https://juliareda.eu/publications/)
- Transparency (https://juliareda.eu/lobby-meetings/)
  - Lobby meetings (https://juliareda.eu/lobby-meetings/)
  - General Expenditure (https://juliareda.eu/general-expenditure/)
- Press (https://juliareda.eu/press/)
- Contact (https://juliareda.eu/contact/)

GitHub Copilot is not infringing your copyright

This is a slightly modified version of my original German-language

article first published on heise.de (https://www.heise.de/meinung/Edit-Policy-GitHubs-KI-Copilot-ist-keine-Urheberrechtsverletzung-6128502.html?seite=all) under a CC-by 4.0 (https://creativecommons.org/licenses/by/4.0/) license.

GitHub is currently causing a lot of commotion in the Free Software scene with its release of Copilot (https://copilot.github.com/). Copilot is an artificial intelligence trained on publicly available source code and texts. It produces code suggestions to programmers in real time. Since Copilot also uses the numerous GitHub repositories under copyleft licences such as the GPL as training material, some (https://twitter.com/eevee/status/1410037309848752128) commentators (https://twitter.com/MalwareJake/status/1411351168643706886) accuse GitHub of copyright infringement, because Copilot itself is not released under a copyleft licence, but is to be offered as a paid service after a test phase. The controversy touches on several thorny copyright issues at once. What is astonishing about the current debate is that the calls for the broadest possible interpretation of copyright are now coming from within the Free Software community.

Copyleft does not benefit from tighter copyright laws

Copyleft licences are an ingenious invention with which the Free Software scene has used copyright, the sharp sword for the content industry, to promote the free exchange of culture and innovation. Works licensed under copyleft may be copied, modified and distributed by all, as long as any copies or derivative works may in turn be re-used under the same license conditions. This creates a virtuous circle, thanks to which more and more innovations are open to the general public. Copyright, which was designed to guarantee exclusivity over creations, is used here to prevent access to derivative works from being restricted.

However, it is also clear that there would be no need for copyleft licences to govern the exercise of copyright in software code by third-party developers at all if copyright did not guarantee rightsholders such a high degree of exclusive control over intellectual creations in the first place. If it were not possible to prohibit the use and modification of software code by means of copyright, then there would be no need for licences that prevent developers from making use of those prohibition rights (of course, free software licenses would still fulfil the important function of contractually requiring the publication of modified source code). That is why it is so absurd when copyleft enthusiasts argue for an extension of copyright. Any extension of prohibition rights not only strengthens the enforcement of copyleft licences, but also the much more widespread copyright licences, which aim to achieve exactly the opposite results.

But this is exactly what is happening in the current debate about GitHub’s Copilot. Because a large company – namely GitHub’s parent company Microsoft – profits from analyzing free software and builds a commercial service on it, the idea of using copyright law to prohibit Microsoft from doing say may seem obvious to copyleft enthusiasts. However, by doing so, the copyleft scene is essentially demanding an extension of copyright to actions that have for good reason not been covered by copyright. These extensions would have fatal consequences for the very open culture which copyleft licences seek to promote.

There are two main versions of the criticism levelled at GitHub for starting Copilot. Some are criticising the very use of free software as source material for a commercial AI application. Others focus on Copilot’s ability to generate outputs based on the training data. One may find both ethically reprehensible, but copyright is not violated in the process.

Text & data mining is not copyright infringement

To the extent that merely the scraping of code (https://twitter.com/bphogan/status/1411097686854488067) without the permission of the authors is criticised, it is worth noting that simply reading and processing information is not a copyright-relevant act that requires permission: If I go to a bookshop, take a book off the shelf and start reading it, I am not infringing any copyright. The fact that scraping content to train an artificial intelligence enters the realm of copyright at all is because digital technology requires making copies of content in order to process it. Copying is fundamentally a copyright-relevant act. Many of the conflicts between copyright and digital technology result from this fact. Fortunately, policymakers and courts have long recognised that digital technology would be completely unusable if every technical copy required permission. Otherwise, people who listen to music with digital hearing aids would first have to acquire a licence for it. Internet providers would have to license every conceivable copyright-protected work that their customers exchange with each other.

As early as 2001, the EU allowed such temporary, ephemeral acts of copying, which are part of a technical process, without restriction – despite the protests of the entertainment industry at the time. Unfortunately, this copyright exception of 2001 initially only allowed temporary, i.e. transient, copying of copyright-protected content. However, many technical processes first require the creation of a reference corpus in which content is permanently stored for further processing. This necessity has long been used by academic publishers to prevent researchers from downloading large quantities of copyrighted articles for automated analysis. Although these scholars had legal access to the content, for example through a subscription from their university, the publishers tried to contractually or technically exclude the creation of reference corpora. According to the publishers, researchers were only supposed to read the articles with their own eyes, not with technical aids. Machine-based research methods such as the digital humanities suffered enormously from this practice.

Under the slogan “The Right to Read is the Right to Mine”, EU-based research associations therefore demanded explicit permission in European copyright law for so-called text & data mining, that is the permanent storage of copyrighted works for the purpose of automated analysis. The campaign was successful, to the chagrin of academic publishers. Since the EU Copyright Directive of 2019, text & data mining is permitted. Even where commercial uses are concerned, rightsholders who do not want their copyright-protected works to be scraped for data mining must opt-out in machine-readable form such as robots.txt. Under European copyright law, scraping GPL-licensed code, or any other copyrighted work, is legal, regardless of the licence used. In the US, scraping falls under fair use, this has been clear at least since the Google Books case (https://www.smithsonianmag.com/smart-news/supreme-court-declines-hear-copyright-challenge-google-books-180958818/).

Machine-generated code is not a derivative work

Some commentators see GitHub Copilot as a copyright infringement because the programme not only uses copyright-protected software code, a lot of which is published under GPL, as training material, but also generates software code as output. According to critics (https://twitter.com/eevee/status/1410037309848752128), this output code is a derivative work of the training data sets because the AI would not be able to generate the code without the training data. In a few cases, Copilot also reproduces short snippets from the training datasets, according to GitHub's FAQ.

This line of reasoning is dangerous in two respects: On the one hand, it suggests that even reproducing the smallest excerpts of protected works constitutes copyright infringement. This is not the case. Such use is only relevant under copyright law if the excerpt used is in turn original and unique enough to reach the threshold of originality. Otherwise, copyright conflicts would constantly arise when two authors use the same trivial statement independently of each other, such as "Bucks beats Hawks and advance to the NBA finals", or "i = i+1". The short code snippets that Copilot reproduces from training data are unlikely to reach the threshold of originality. Precisely because copyright only protects original excerpts, press publishers in the EU have successfully lobbied for their own ancillary copyright that does not require originality as a precondition for protection. Their aim is to prohibit the display of individual sentences from press articles by search engines. It is precisely this problematic demand that the Free Software community endorses when it demands absolute control over the smallest excerpts of software code.

On the other hand, the argument that the outputs of GitHub Copilot are derivative works of the training data is based on the assumption that a machine can produce works. This assumption is wrong and counterproductive. Copyright law has only ever applied to intellectual creations – where there is no creator, there is no work. This means that machine-generated code like that of GitHub Copilot is not a work under copyright law at all, so it is not a derivative work either. The output of a machine simply does not qualify for copyright protection – it is in the public domain. That is good news for the open movement and not something that needs fixing.

Those who argue that Copilot's output is a derivative work of the training data may do so because they hope it will place those outputs under the licensing terms of the GPL. But the unpleasant side effect of such an extension of copyright would be that all other AI-generated content would henceforth also be protected by copyright. What would then stop a music label from training an AI with its music catalogue to automatically generate every tune imaginable (https://www.musictech.net/news/programmers-generate-every-possible-melody-in-midi-to-prevent-lawsuits/) and prohibit its use by third parties? What would stop publishers from generating millions of sentences and privatising language in the process?

At the World Intellectual Property Organization (WIPO), companies are already lobbying for an extension of copyright to machine-generated works. According to WIPO (https://www.wipo.int/about-ip/en/frontier_technologies/): “The main focus of those questions is whether the existing IP system needs to be modified to provide balanced protection for machine created works and inventions”, the main beneficiaries of such an extension of copyright would be the major technology corporations that are best placed to develop and scale AI applications. Such as Microsoft. Critics of GitHub’s business practices would do well not to play into their hands.

This work is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Facebook (https://www.facebook.com/sharer.php?u=https%3A%2F%2Fjuliareda.eu%2F2021%2F07%2Fgithub-copilot-is-not-infringing-your-copyright%2F)

Twitter (https://twitter.com/intent/tweet?text=GitHub%20Copilot%20is%20not%20infringing%20your%20copyright&url=https://juliareda.eu/?p=12791&via=Senficon)

(/me-for-you-in-eu

My name is Julia, I'm the Pirate in the European Parliament.

I'm fighting to **make copyright in the EU unified, progressive and fit for the future**. **Will you join me?**

E-Mail

Anmelden!

@Senficon (https://twitter.com/Senficon) JuliaRedaMEP (https://facebook.com/JuliaRedaMEP)

Date: 5.07.21 (https://juliareda.eu/en/2021/07/) Category: General (https://juliareda.eu/category/general/) Comments: 0 (https://juliareda.eu/2021/07/github-copilot-is-not-infringing-your-copyright/#respond) Author: Julia Reda (https://juliareda.eu/en/author/julia_reda/)

- Privacy Policy (https://juliareda.eu/privacy/)