GitHub Copilot: Copyright, Fair Use, Creativity, Transformativity, and Algorithms *

Gavin D. Howard †

October 27, 2021

Abstract

The recent announcement of GitHub's Copilot, a GPT-3-based machine learning (ML) model trained on source code from Free and Open Source Software (FOSS), caused controversy with its assumed violation of FOSS licenses. In this paper, I attempt to cut through the controversy and demonstrate that ML models do not necessarily violate those licenses, but that GitHub's use of it does and why.

1 Introduction

On June 29, 2021, GitHub announced [8] Copilot [10], a GPT-3-based model [3] trained on all public source code hosted on GitHub. [9]

Nat Friedman, GitHub's CEO, claims [7] that:

> In general: (1) training ML systems on public data is fair use (2) the output belongs to the operator, just like with a compiler.

This paper is an examination of the veracity, applicability, and consequences of Mr. Friedman's claims.

†The author is *not* a lawyer and is not giving legal advice.

2 GitHub's Claims

There are two parts to GitHub's claims:

- Training an ML model is fair use.
- The output of an algorithm is owned by the operator.

2.1 Fair Use

First, we must consider Mr. Friedman's claim of fair use.

He began with "In general" [7], which is accurate because ML training on public data is not *always* fair use, and the case law has not been settled yet. [9] [14]

Nevertheless, I will consider his position from his perspective, laid out by OpenAI. [17]

OpenAI has a good case that not allowing fair use for training of machine learning models would create a heavy burden on ML. I do not dispute that.

However, the document was written about works "for human consumption for their standalone entertainment value." (Page 5) Such is not the case for source code; the function of source code is for *consumption by machines* and execution *by machines* to do *useful work*.

Nevertheless, accepting their position that *training* an ML model on publically-available data is fair use does not mean that *distributing the output* of that model is, though OpenAI does make several arguments that distributing the output *is* fair use.

17 U.S.C. § 107 establishes the following factors for fair use:

(1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;

(2) the nature of the copyrighted work;

(3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and

(4) the effect of the use upon the potential market for or value of the copyrighted work.

2.1.1 Factor 1

GitHub Copilot has a commercial character because GitHub intends to sell access to it.

However, could source code-trained ML models such as Copilot have non-commercial or non-profit uses?

I argue that they cannot. Besides personal and academic uses, there are not many non-commercial or non-profit uses of generated source code because the inherent function of source code is to *do something* rather than to *entertain*.

Thus, I argue that Factor 1 does not substantially affect the legal situation of using copyrighted source code as training for ML models, especially GitHub Copilot, and that it is not likely that it will ever substantially affect the legal situation of such models.

2.1.2 Factor 2

I have already addressed that Factor 2 is different for source code. However, I will go further.

Because the nature of source code is to be executed by machines to do useful work, an ML model that uses copyrighted source code as input and still outputs source code that does the same useful work does not change the nature of the work.

To a human, two sentences that say the same thing but use different wording can have different effects. One can cause us to be apathetic, and the other can cause a range of emotions.

The types of works OpenAI are talking about have this property: a small change can trigger far different reactions in the intended audience for consumption.

Source code is not like that. It can change form, and yet, the machine can still do the *same useful work*.

This is an example of how OpenAI's argument cannot be applied to source code; the nature of the works under question is vastly different.

I must also address a point in "Fair Learning" [14] that says,

> If it's a computer program, you can take the functional aspects of that program, even if it means copying some of the code directly.

That seems like a direct refutation of my point, until you read the footnote, which says, "Well, except in the Federal Circuit." That is precisely where copyright would be litigated on appeal!

Thus, I argue that Factor 2 does not substantially affect the legal situation of using copyrighted source code as training for an ML model.

2.1.3 Factor 3

In OpenAI's argument for Factor 3, they quote "Copyright for Literate Robots" [13]:

> Verbatim copying of a complete work will be protected as fair use if the copy is used solely as input to a process that does not itself use the works expressively. Or, to put it a little more provocatively, nonexpressive uses do not count as reading.

That is true for works that are meant for human consumption, which does not apply to source code.

The main "expressive" use of source code is to have a machine execute it. Such obviously still applies to the source code output by Copilot, which means that OpenAI's argument about Factor 3, that the expressive use of the output of an ML model does not have the same purpose as the original work, is void with regards to source code.

OpenAI quotes Authors Guild v. Google [1]:

> What matters in such cases is not so much "the amount and substantiality of the portion used" in *making a copy*, but rather the amount and substantiality of *what is thereby made accessible* to a public for which it may serve as a competing substitute.

Authors Guild v. Google was about Google's scanning of books to make them searchable in Google Books. The Court of Appeals for the Second Circuit held that Google's use fell under fair use and did not infringe.

However, OpenAI neglected to mention a crucial fact: Google does *not* reproduce the content of copyrighted works publically. [2] Instead, it only reproduces "snippets," which are not even a full page, even though it will produce full pages for books under the "Partner Program" and full works for those that are in the public domain.

This means that Google was careful to not make a large amount of copyrighted works publically accessible.

Such is not the case for GitHub Copilot in particular Armin Ronacher's tweet [19], and I would argue that it will not be the case for ML models in general because all ML models like Copilot will keep suggesting output as long as you ask for it. There is no limit to how much output someone can request.

In other words, it is trivial to make such models output a substantial portion of the source code they were trained on. This is exactly the opposite of Google Books.

Thus, I argue that Factor 3 does not substantially affect the legal situation of using copyrighted source code as training for an ML model.

2.1.4 Factor 4

For factor 4, let's go back to "Copyright for Literate Robots" [13] (emphasis added):

> Verbatim copying of a complete work will be protected as fair use if the copy is used solely as input to a process that does not itself use the works expressively. Or, to put it a little more provocatively, nonexpressive uses do not count as reading. *They are not part of the market that copyright cares about, because the author's market consists only of readers.*

Generated source code has the same market as the original work, and can "serve as a competing substitute" as mentioned in *Authors Guild v. Google.* [1]

With ways to generate source code from copyrighted works, the market for such original works *is* affected. If Copilot generates source code that accomplishes certain useful work, any copyrighted source code that also accomplishes that work would see its market shrink and its value decline.

Thus, I argue that Factor 4 does not substantially affect the legal situation of using copyrighted source code as training for an ML model.

2.2 Compilers and Algorithms

Let's look at the second part of Mr. Friedman's assertion:

> ...the output belongs to the operator, just like with a compiler.

In essence, Mr. Friedman is claiming that the output of a compiler belongs to the *user* of the compiler, not the author of the source code.

That is wrong, and there is case law saying otherwise.

The GNU General Public License [6] is the seminal example of FOSS licenses, and it has been litigated in court. [20]

This matters because the GPL is meant to be a license for software whose source code is publically available. This means that any person can copy GPL-licensed source code and compile it themselves.

Does that person then own the copyright on the compiler's output? Courts have said no.

This makes sense; if copyright disappeared with the act of compiling software, then copyright would not apply to software at all.

But the comparison of Copilot to compilers *is* accurate, just not in the way Mr. Friedman intended.

Compilers take source code as input, or the output of other compilers as input, and produce software that can either be incorporated into other software or executed. ML models take source code as input and produce source code.

Both are algorithms. Mirriam-Webster defines an "algorithm" [15] as:

> **algorithm**
>
> a step-by-step procedure for solving a problem or accomplishing some end

No new creativity is applied when executing a step-by-step procedure. If putting copyrighted works through an algorithm creates a situation where copyright does not apply to the output of the algorithm, then legal absurdities such as Monolith [16] do, in fact, work as intended: they erase copyright and distribute works without infringing.

GitHub may claim that its use of copyrighted code in training Copilot and distributing its output is different from running compilers on source code because it is more transformative than compilers are, but doing so would also be unreasonable.

The reason is because compilers can, in fact, transform the code they are given in ways that copyright law would consider transformative. An example is Clang turning an linear algorithm into a constant-time algorithm. [11]

Another example is link-time optimization combined with an optimization called "inlining." Link-time optimization is done at a phase of the compilation process where every piece of code in the software, including pieces from disparate sources and under various copyrights and licenses, are combined together in one piece. Inlining is an optimization that takes that combination and effectively mixes all of the pieces together.

That mixing is almost *exactly* like what ML models have done to their training data, and that is before taking into account the use of other optimizations that may mix code even further.

As such, it is my opinion that the output of ML models should be treated no differently than the output of compilers, as claimed by Mr. Friedman, and I argue that GitHub Copilot *is* infringing on copyright, which means that FOSS licenses should apply.

2.3 Copyright on the Model

If fair use does not cover what GitHub is doing, does copyright apply to the model?

To answer that, I need to answer this: is the model is a derivative work? The US Copyright Office says [18]:

> A "derivative work," that is, a work that is based on (or derived from) one or more already existing works, is copyrightable if it includes what the copyright law calls an "original work of authorship."

The model is a derivative work because it is "based on" the original source code which it was trained on and in my opinion, does not include "original work[s] of authorship" because, as demonstrated above, the ML model is simply an algorithm, and training an ML model is *also* an algorithm, with one exception: human operators set parameters before and during training.

A claim could be made that setting parameters is the "original work of authorship" and sufficiently creative to make the model a transformative work, but I argue that it is not. It is a question the courts will have to decide.

Thus, the question of what copyright applies to the model depends on how transformative and creative it is.

3 Who Infringes?

If distributing the output is infringement, who is infringing? GitHub or the "operator," as claimed by GitHub?

The key to the answer is: who is distributing the code? If GitHub is distributing the code, *even if the operator is*, then GitHub is infringing because they are before the operator in the chain of distribution.

To answer that question, let's look at what *17 U.S. Code § 106 - Exclusive rights in copyrighted works* [4] says about distribution:

> Subject to sections 107 through 122, the owner of copyright under this title has the exclusive rights to do and to authorize any of the following:...
>
> (2) to prepare derivative works based upon the copyrighted work;
>
> (3) to distribute copies or phonorecords of the copyrighted work to the public by sale or other transfer of ownership, or by rental, lease, or lending;

Since Copilot is a commercial product, GitHub is definitively distributing copies to the public by sale. GitHub is also preparing derivative works by *training* the model.

Thus, GitHub is infringing on copyright.

4 Preventing Infringement

The fact that GitHub is infringing on copyrighted source code means either GitHub or authors of FOSS need to prevent infringement.

4.1 Challenges for GitHub

The only thing that GitHub can do to prevent infringement on FOSS works is to comply with the licenses that the authors of FOSS have applied to their works.

4.1.1 Credit to Authors

They can start with giving credit. Many FOSS licenses mandate that credit be given to the original authors.

As an example, Google Books prominently displays the author's name, as well as its title, ISBN, and many other important pieces of data about a book. [12]

This means that Google is ensuring that the public knows where to find the rest of such copyrighted works and how to *legally* obtain copies.

Copilot does no such thing, as demonstrated by Armin Ronacher's tweet. [19] If Copilot did, then it would automatically not be infringing on copyright for a lot of source code in its model because that is the only requirement in many FOSS licenses.

4.1.2 Copyleft

Beyond credit, if the output is still under copyright, all FOSS licenses still apply, especially copyleft licenses like the GPL and the AGPL.

This means that GitHub has an obligation to ensure that their customers know the terms they must follow. Without such information, developers using Copilot have no way of complying with the licenses for the code they receive without undue burden.

4.1.3 How GitHub Can Implement Compliance

To properly provide such information to their customers, GitHub needs to make Copilot do two things:

- Track the licenses and provenance of each copyrighted work used to train Copilot. This includes source code that Copilot copies from a copy.
- Whenever Copilot outputs code, it must figure out what copyrighted source code was used to influence the output.

GitHub will probably claim that the burdens of tracking the licenses and provenance of each bit of source code used to train Copilot is an undue burden. This is false; GitHub is profitable without Copilot, so it does not matter even if Copilot is completely illegal.

4.2 Challenges for Authors of FOSS

The main challenge for authors of FOSS is that there is no way to tell when their works have been copied and distributed because use of Copilot will

often happen in private spaces, and even in public, the operator of Copilot will not receive any information about the provenance of the generated code.

It may happen by chance when a company publishes open source code, and similarities to preexisting code are discovered, but relying on chance to catch infringement means most will never come to light. This means that for authors of FOSS, it may be necessary to initiate lawsuits against GitHub preemptively to guard against infringement.

Such lawsuits could have another purpose: to establish whether or not ML models are under the copyright of the training data, at least for models trained on source code.

5 Consequences of GitHub's Position

If GitHub is correct, the consequences will be dire.

First, many companies that have opened their software's source code will close the source, preventing humans from learning and building on that software.

Second, companies that wish to use code under FOSS licenses without compliance will simply use Copilot to generate something close to the code they want and modify it to make it work, something that might be called "code laundering." [5] This will put FOSS projects at a disadvantage.

Third, as a consequence of the above, people who work on FOSS in their spare time may reduce the amount they work on it, decreasing the amount of innovation and competition against proprietary solutions.

Fourth, bug fixes for FOSS may be kept hidden, decreasing the security of computer systems.

6 Conclusion

As I see it, the current legal situation regarding GitHub Copilot is as follows:

- Training an ML model on source code *may* be fair use.
- But the output of the model is still copyrighted because the four fair use factors do not substantially affect the copyright(s).
- The output of an algorithm run on copyrighted material is still copyrighted, and that fact has been established in the courts.

- GitHub is infringing because Copilot is distributing copyrighted output and because the model itself could still be under copyright.
- There would be an undue burden on copyright holders to find and litigate copyright infringements.
- The consequences of GitHub's position would be disasterous for FOSS and the technology industry as a whole.

It is my opinion that GitHub's Copilot is illegally infringing, and if the infringements are not fixed, the result will cause copyright to disappear from software, rendering copyright ineffective for protecting software.

References

[1] United States Court of Appeals for the Second Circuit. *Authors Guild v. Google*. https://h2o.law.harvard.edu/collages/37267. Accessed on 2021-08-18. Oct. 2015.

[2] Jonathan Band. *The Google Library Project: Both Sides of the Story*. https://quod.lib.umich.edu/p/plag/5240451.0001.002/--google-library-project-both-sides-of-the-story?rgn=main;view=fulltext. Accessed on 2021-08-18. 2006.

[3] Scott Carey. *Developers react to GitHub Copilot*. https://www.infoworld.com/article/3624688/developers-react-to-github-copilot.html. Accessed on 2021-08-18. July 2021.

[4] United States Congress. *17 U.S. Code § 106 - Exclusive rights in copyrighted works*. https://www.law.cornell.edu/uscode/text/17/106. Accessed on 2021-08-19.

[5] eevee. *Code Laundering*. https://twitter.com/eevee/status/1410037309848752128. Accessed on 2021-08-19. June 2021.

[6] Free Software Foundation. *GNU General Public License*. https://www.gnu.org/licenses/#GPL. Accessed on 2021-08-18.

[7] Nat Friedman. *In general...* https://news.ycombinator.com/item?id=27678354. Accessed on 2021-08-18. June 2021.

[8] Nat Friedman. *Introducing GitHub Copilot: your AI pair programmer*. https://github.blog/2021-06-29-introducing-github-copilot-ai-pair-programmer/. Accessed on 2021-08-18. June 2021.

[9] Dave Gershgorn. *GitHub's automatic coding tool rests on untested legal ground*. https://www.theverge.com/2021/7/7/22561180/github-copilot-legal-copyright-fair-use-public-code. Accessed on 2021-08-18. July 2021.

[10] GitHub. *Your AI pair programmer*. https://copilot.github.com/. Accessed on 2021-08-18. June 2021.

[11] Matt Godbolt. "Optimizations in C++ Compilers". In: *Communications of the ACM* 63.2 (Feb. 2020). Accessed on 2021-08-18, pp. 41–49.

[12] Google. *Google Books*. https://www.google.com/books/edition/1984/kotPYEqx7kMC?hl=en&gbpv=0&bsq=1984. Accessed on 2021-08-18. Aug. 2021.

[13] James Grimmelmann. "Copyright for Literate Robots". In: *Iowa Law Review* 101 (Dec. 2015). Accessed on 2021-08-18, pp. 657–664.

[14] Mark A. Lemley and Bryan Casey. "Fair Learning". In: *Texas Law Review* 99.4 (Mar. 2021). Accessed on 2021-08-18.

[15] Mirriam-Webster. *Algorithm at Mirriam-Webster*. https://www.merriam-webster.com/dictionary/algorithm. Accessed on 2021-08-18.

[16] Monolith. *Monolith*. http://monolith.sourceforge.net/. Accessed on 2021-08-18.

[17] Cullen O'Keefe et al. *Before the United States Patent and Trademark Office*
Department of Commerce
Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation. https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf. Accessed on 2021-08-18. June 2021.

[18] United States Copyright Office. *Copyright Registration for Derivative Works*. https://web.archive.org/web/20051228180915/http://www.copyright.gov/circs/circ14.html. Accessed on 2021-08-19.

[19] Armin Ronacher. *That's not the right license Mr Copilot*. https://twitter.com/mitsuhiko/status/1410886329924194309. Accessed on 2021-08-18. July 2021.

[20] FSFE Wiki. *GPL Enforcement Cases*. https://wiki.fsfe.org/Migrated/GPL%20Enforcement%20Cases. Accessed on 2021-08-18.