GitHub Copilot: an analysis from a legal and technical perspective.

Davide Moletta

University of Trento

January 11, 2023

Keywords: GitHub Copilot, Generative deep learning models, copyright, Fair use in machine learning, open source.

Abstract

In recent years, content produced by machines has become so popular that it's possible to find every type of artwork generated in such a way. The major way to create content like this is via generative deep learning models which learn from a dataset in order to create original content similar to the training one. The focus of this paper will be GitHub's newest project, GitHub Copilot, an Artificial Intelligence system capable of generating code when prompted with natural language descriptions of some functions. Three reasons have led to the recent buzz around Copilot: the first is that it's an innovative tool that could open many doors for future research and developments; the second is that the generated code for complex functions is buggy and insecure; and the third is that it is still in a grey area of the law. We will see how it works from a technical point of view and why it has been described as borderline illegal, if not illegal, and non-ethical. In addition, we will discuss what countermeasures are in place and how Copilot could affect the future of the open-source community, as well as the development of AI-based generative systems.

1 Introduction

Artificial intelligence (AI) systems have gained a lot of popularity and accessibility in the last few years, and as a result, they have become more flexible and efficient, thus they have become more useful in many situations. AI is used in a wide range of fields, such as virtual assistants, autonomous vehicles, medical imaging analysis, and so on. The aim of this technology is to assist humans by automating or simplifying certain tasks. It is incredibly helpful that artificial intelligence software can do complex and time-consuming tasks. For example, it can speed up data analysis in the research field, resulting in faster research.

Many people have expressed concern about artificial intelligence systems as we don't know their full capabilities and there are little to no regulations regarding this field. Questions such as "What should be the rules to ensure that we can always trust AI systems?", "Who is liable when an AI fails at its job?" or "Is it necessary to put an ethical limit to what an AI can and cannot do?" remain unanswered.

In this article, we will analyze GitHub Copilot[1], an AI-based pair programming software created by GitHub[2] to help developers write code. In sections 2 and 3, we will introduce the current state of the art. There will be an explanation of generative deep learning models such as Copilot and, after that, we will review all the functions and capabilities of this software as well as its problems. Chapter 4 is dedicated to an analysis of GitHub's actual situation and why they are being sued. In section 5, we will discuss the user's perspective and all the risks involved in using Copilot. Finally, in chapter 6 there will be a comparison between Copilot and its major competitors with special attention to the legislative point of view.

During the entire paper, we will also discuss the potential side effects that artificial intelligence systems and, in particular, GitHub Copilot could have on the future of both machine learning models training and the open-source community.

2 Generative deep learning models

In this section, we discuss Generative Deep Learning (GDL) models, we'll see what they are and how they work. A GDL model is a specific type of machine learning algorithm. The key components of a GDL network are the training set, the generative model and a pseudo-random noise. I'll use the example of David Foster from his book "Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play" to show how these algorithms work. "*Suppose we have a dataset containing images of horses. We may wish to build a model that can generate a new image of a horse that has never existed but still looks real because the model has learned the general rules that govern the appearance of a horse. This is the kind of problem that can be solved using generative modeling.*"[4]. In this case, the dataset containing horse images is the training set on which the model learns. The generative model then tries to produce a new image based on the observations of the training set and the noise.

The example provided above is about images but this can be done even with music, text and combinations of them. In this paper, we will examine GitHub Copilot, a GDL model capable of generating code snippets.

[1]GitHub Copilot: https://github.com/features/copilot
[2]GitHub: https://github.com

3 GitHub Copilot

As we saw in the Introduction, GitHub Copilot is an AI pair programmer, but what does it mean? Basically, it has three major functions. It is capable of describing functions in natural language as comments and it can refactor code, meaning it reshapes the code without changing its behaviour. However, the key function and the most interesting one is the third. Copilot can translate natural language requests into generated code. For example, if we ask for a function that determines if a number is even (image 1) the Copilot will generate a preview of how the code will look like (image 2). When the user accepts the offered solution, Copilot will write the code in the editor (image 3). Even with basic functions, we can see that there are cleaner and better ways to write the same code (image 4). Still, the function written by Copilot works fine and it's correct but, obviously, it will struggle with more complex functions.

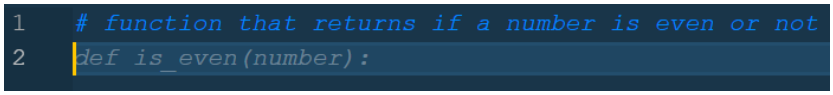


Figure 1: Natural language request given to Copilot

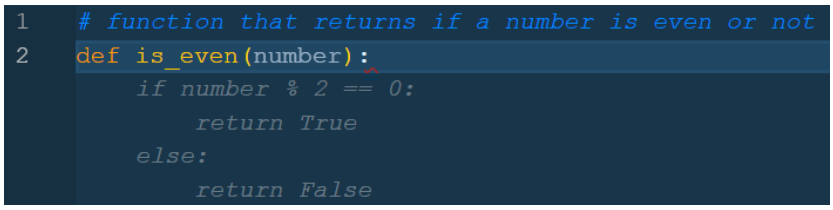


Figure 2: Preview of the proposed snippet

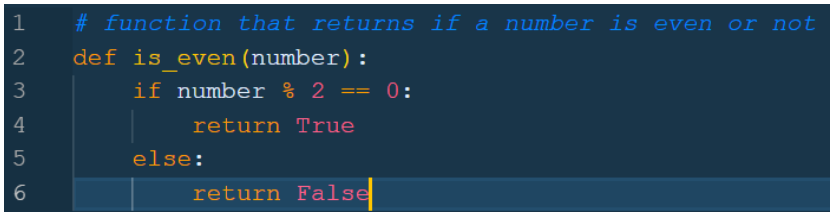


Figure 3: Code written with Copilot

3.1 Copilot Training

GitHub Copilot's core is the OpenAI Codex, a generative deep learning model created by OpenAI[3]. This model does most of the work since it's the component

[3]OpenAI: https://openai.com

```
1  # function that returns if a number is even or not
2  def is_even(number):
3      return number % 2 == 0
```

Figure 4: Better code layout for "is_even" function

that translates natural language into code and processes other functions. In section 2, we described how GDL models work. They need a training set to learn from, the bigger the better. As stated by GitHub: "*Copilot has been trained on natural language text and source code from publicly available sources, including code in public repositories on GitHub.*"[6]. This means that Copilot has been trained on billions of lines of code coming from different repositories. This explains why Copilot is somewhat effective at doing its work and why people are so concerned about it.

3.2 Copilot Problems

Copilot, like other GDL models, typically suggests code snippets by generating an output based on the original source code on which it has been trained. Unfortunately, this doesn't always happen. There has been evidence of cases [11][2] in which Copilot gave as output a code snippet copied verbatim from the source code of a user's GitHub repository or copied a wrong license for generated code. This can happen when working with a GDL model and GitHub confirmed it on the official website of Copilot saying that: "*Our latest internal research shows that about 1% of the time, a suggestion may contain some code snippets longer than ∼150 characters that matches the training set.*"[6] .

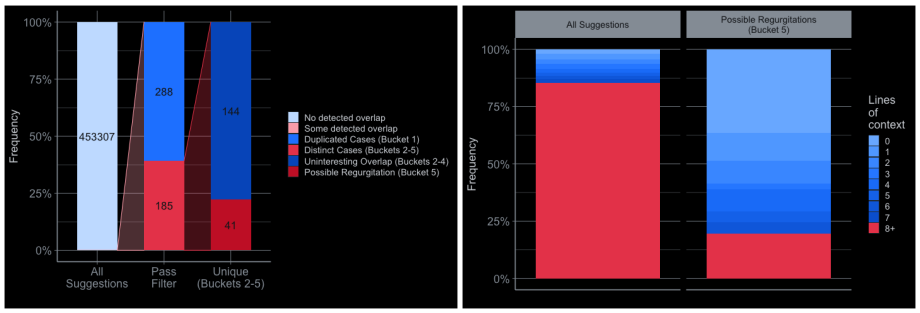
The problem is that it appears that when training Copilot on public repositories the licenses were ripped from the source code. Therefore, if Copilot copies verbatim a code snippet when doing its job, it can violate those licenses by not referencing the original author. Further, not all licenses are 100% permissive, for example, if you use a piece of code protected by a Copyleft license you must distribute your code under the same license otherwise you would violate the license conditions.

3.3 GitHub's internal research

The data discussed in the previous section come from an internal study conducted by GitHub in 2021 [8]. The purpose of this research was to find out how well Copilot works and how often it copies code verbatim. The research saw almost 300 employees working while actively using Copilot, this led to 453,780 suggestions to study. The researchers set a threshold for code snippets that had at least 60 words corresponding to the ones found in the lines of code on which Copilot was trained. It's possible to see in image 5a that after the filter was applied only 473 cases were considered for the study. After the duplicates were removed there were only 185 cases (the ones in bucket 1). The remaining cases

were divided into buckets from 2 to 5 and only the 41 in bucket 5 were considered interesting for the researchers. The study states that: "*That corresponds to one recitation event every 10 user weeks (95% confidence interval: 7 - 13 weeks).*"[8].

Moreover, they saw (image 5b) that the more context is given to Copilot the more unique its suggestions are. This means that if you ask Copilot for a code snippet in an empty file the odds of it regurgitating a snippet copied verbatim are higher.



(a) Suggestion cases of Copilot during the study

(b) Copilot's possible suggestions based on given context

Figure 5: Results of GitHub's research [8]

4 GitHub's situation

Both GitHub and the company that owns it, Microsoft[4], as well as OpenAI, are accused of violating the DMCA[5] as well as infringing huge amounts of copyright through their actions. We will now take a closer look at why these accusations were made and how Microsoft and GitHub responded to them. Moreover, we will see what is considered fair use in the United States and in the European Union. In addition, there will be a comparison between the Google Books case and the training of GitHub Copilot.

4.1 GitHub's defense

Former CEO of GitHub, Nat Friedman, defended the legality of Copilot stating two points [10]. The first is that training machine learning systems on public data is considered fair use. The second is that the output of Copilot belongs to the user and not to GitHub. While on the first point discussions are still being held, on the second point he is correct, the output belongs to the user of Copilot. Since the user accepts the suggestion it's his responsibility to make

[4]Microsoft: https://www.microsoft.com
[5]DMCA: https://www.dmca.com

proper use of it and to make sure that the code works as intended. GitHub even stated it on their official website: "*You are responsible for the code you write with GitHub Copilot's help. We recommend that you carefully test, review, and vet the code before pushing it to production, as you would with any code you write that incorporates material you did not independently originate.*"[6].

Another claim made by Microsoft is that training an AI model on public data is not only fair use but also covered by the GitHub Terms of Service. Looking at the Terms of Service, there is a section about public repositories which states: "*We need the legal right to do things like host Your Content, publish it, and share it. You grant us and our legal successors the right to store, archive, parse, and display Your Content, and make incidental copies, as necessary to provide the Service, including improving the Service over time. This license includes the right to do things like copy it to our database and make backups; show it to you and other users; parse it into a search index or otherwise analyze it on our servers; share it with other users; and perform it, in case Your Content is something like music or video. This license does not grant GitHub the right to sell Your Content. It also does not grant GitHub the right to otherwise distribute or use Your Content outside of our provision of the Service.*"[9]. This gives permission to GitHub to process the contents of users but not to rip off licenses from the original code. If Copilot is able to take the code from one repository and copy it verbatim, it must be able to reference the correct original owner if the license requires doing so.

4.2 Fair Use analysis

To see whether Copilot can be considered a fair use of the original work, we will see the regulations in effect in both the United States and the European Union.

Regarding the U.S. we will refer to Section 17 of the Copyright Act (17 U.S.C. § 107 - Limitations of exclusive rights: Fair use) [12]. It is not stated in this provision what exactly makes up fair use, but rather provides parameters upon which judges may determine whether or not a use is fair. The purpose and the economic character of the use, the nature of the work, the amount and substantiality of the portion used, as well as the impact of the use on potential markets, are all factors that determine whether the use is fair.

In the E.U., instead, directive 2001/29/EC [3] provides rules on when a reproduction of a copyrighted work can and can't be permitted. With Article 2, authors may authorize or prohibit any reproduction of their work in any form. As an exception to this right, Article 5 states that a transient or incidental act of reproduction shall be allowed if it is an essential part of a technological process and has no independent economic significance.

In their article "COPYRIGHT IN GENERATIVE DEEP LEARNING" [5] Giorgio Franceschelli and Mirco Musolesi argue that it is not impossible that GitHub Copilot could be considered fair use. They assert that: "*In this case, the use is not for research purpose, and it seems more expressive than non-expressive. In addition, its economic character is not negligible (Copilot is free to use, but companies may use it). The public availability of the works helps*

satisfying the second criterion for fair use. Then, the work is entirely used during training, but the substantiality of the use is questionable, as discussed before. Finally, the effect upon the potential market depends on the model itself. If it cannot substantially reproduce an existing source code or, if it can, it is able to identify it and refer the user to the original source, also this fourth condition is satisfied and the use could presumably be seen as a fair use."[5].

Since 2021, the year in which the article was written, things have changed. Copilot is now a paid software for both corporations and individuals, and thus the economic character of the software has changed. This could make it more difficult to understand if Copilot's training is actually fair use or not.

4.3 Google Books case

Now that we have considered the opinions of Giorgio Franceschelli and Mirco Musolesi we will see a recent case that may support the claims made by Microsoft.

This case is the "Google Books case", a class action filed by the Authors Guild[6] against Google[7]. Google was accused of a massive infringement of copyright by scanning and indexing over 20 million books for the creation of a search database without asking permission from the copyright owners nor paying any fees. In its decision, the Supreme Court determined that Google's new tool had transformed the original work, was beneficial to readers and authors and was therefore considered fair use.

Google's product has nothing to do with AI models but could be used as an example since the operation of scanning books is similar to the training operations for a GDL model.

So, whether training an AI system on public data falls under the fair use doctrine still remains in dispute and, until a Court takes an official decision, we will not have an answer to this question.

4.4 Class Action against GitHub

Matthew Butterick and the Joseph Saveri Law Firm have recently filed a lawsuit against GitHub for its apparent Software Piracy. The investigation started in June 2022 when Matthew Butterick wrote about the problems of the Copilot in the article "*THIS COPILOT IS STUPID AND WANTS TO KILL ME*"[8]. After the investigation, on November 3rd 2022, Matthew Butterick and his colleagues filed a class action against Microsoft, GitHub and OpenAI to challenge the legality of Copilot. The claim was that: "*By training their AI systems on public GitHub repositories we contend that the defendants have violated the legal rights of a vast number of creators who posted code or other work under certain open-source licenses on GitHub.*"[7]. Microsoft violated 11 open-source licenses that

[6]The Authors Guild: https://authorsguild.org

[7]Google: https://www.google.com

[8]THIS COPILOT IS STUPID AND WANTS TO KILL ME: https://matthewbutterick.com/chron/this-copilot-is-stupid-and-wants-to-kill-me.html

require attribution of the author's name and the DMCA, which prohibits the act of removing copyright management information from intellectual property, according to the plaintiffs.

This could be a life-changing event for AI systems since this is the first class action case against them. As a result of the decision of the judge, these types of systems may have to follow more strict rules regarding training and data generation in the future. So, until the end of this legal action, there will be no answer to whether the training of GitHub Copilot is considered fair use or not.

5 Users' situation

The other side of the coin in this situation is the users' side. While it's true that Copilot could simplify basic operations and "substitute" the coder in writing repetitive and straightforward code, it is also true that it could get the end user in trouble if not used properly.

5.1 Risks and problems

As already explained, GitHub Copilot can translate human language into code. It's pretty accurate at doing so thanks to the massive amount of data on which it has been trained. However, there are still problems with the output of this software. The first and most critical is the one discussed in the previous section. While using Copilot you can accidentally infringe copyright since it's possible that the output is copied verbatim from a public repository, which requires you to follow the license conditions in order to use that piece of code. Moreover, you could have your final code filled with code snippets that not only are copied from repositories protected by open-source licenses but could even be protected by different licenses, which grant different potential uses of the source code. For example, you could have a code snippet licensed with the MIT License[9] which only requires you to include a copyright notice and another code snippet licensed with the GNU GPLv3[10] which, being a copyleft license, requires you to include a copyright notice, make the source code publicly available and to apply the same license to your project. This means you can distribute some material under a license of your choice, not knowing that part of your code was originally under the GNU GPLv3 and that you should distribute your material under this license.

Another problem is the fact that the output code is sometimes buggy and full of security flaws, especially when complex functions are requested to Copilot. This means that without proper testing and revision, there are more chances that using Copilot could cause harm than helping coders. This has to be kept in mind, especially for companies that are more at risk when it comes to security. In fact, most of them are in doubt about whether to let their employees use Copilot or not. There are many variables to take into account and it's a tough

[9] MIT License: https://spdx.org/licenses/MIT.html
[10] GNU GPLv3 License: https://www.gnu.org/licenses/licenses.html

decision for CEOs and CTOs. Microsoft itself stated that every code created with the help of their assistant should be revisited and tested as you would do with code written on your own or found online.

5.2 Open Source Community

While coders and companies struggle to decide whether or not to use Copilot, the open-source community is angry at GitHub for what they've done. The idea behind this community of coders all around the globe is that code is meant to be free for anyone to use, whether they wish to use it personally or contribute something to the project. In that sense, this community is an ever-growing one that aims at achieving the best possible result with the help of everyone in order to continually improve current technology.

It's not a surprise that seeing Copilot "stealing" the source code from repositories of millions of these coders got them angry, first of all, because the licenses which oblige to reference the original author were ripped from the source code and second because this can lead to an unwanted situation in which everyone can use open source code for every purpose without ever needing to refer to an author or to keep the derived work open source. For example, Copyleft code could be implemented into a commercial project without following the license conditions. In other words, Copyleft code could be included in a company project that will be released as paid software.

This is obviously not correct, even though the chances of something like this happening are close to zero, they are not completely zero. This is a problem that needs to be addressed as soon as possible so that the original authors are properly acknowledged and users can use Copilot without risking violating copyright accidentally.

6 Comparison

Because of AI systems' growth in popularity, it's not surprising that GitHub Copilot isn't the only pair programmer available out there. In this section, we will compare Copilot to one of its competitors and we will see why Copilot is the only one receiving all these criticisms.

The alternative that we will analyse is Tabnine[11], a freemium AI pair programmer. It can auto-complete the code you are writing and translate natural language requests into generated code, basically the same functions as the ones offered by GitHub. While it is true that Copilot is considered because it is more popular and is developed by a larger company, there are two fundamental differences between Copilot and Tabnine or other similar software. One is that the code used to train Tabnine is open source and protected only by permissive licenses. In addition, while using the generative feature of the software you can get copyright notices. Those slight differences might seem insignificant, but they get rid of a lot of problems like the incompatibility between licenses

[11] Tabnine: https://www.tabnine.com

discussed above and the fact that you could accidentally include code copied verbatim without even knowing and without giving credit to the original author. Of course, this doesn't mean that Tabnine is flawless. The suggested code should be revisited and tested even in this case since it could be buggy or not secure but at least users don't have to think that much about possible copyright infringements while writing their code.

Is it correct, however, to "steal" the code created by the open source community in order to improve the system and give users generated code based on the original code? Some think that this is correct since the purpose of open source code is to be available for everyone to use in any way they want while others think that this is not what the open source community wanted since using projects like this could lead to a situation where many people use open source code to write their own without distributing the final result, and thus slowing innovation and destroying this community based on sharing and mutual help.

7 Future work

Artificial Intelligence technology continues to advance, raising serious ethical questions concerning the relationship between AI and copyright law. On one hand, AI can be used to create original artworks, literature, and music that would not have been possible without its assistance. In contrast, the use of artificial intelligence in the creation of these works raises a number of questions regarding authorship, originality, and ownership.

Of course, the fact that the actual copyright law was created before the invention of AI plays a role in this. As we have seen in this paper there are a lot of grey areas regarding the training of AI systems and the data that are used in this process as well as the ownership of the generated output. In fact, not every AI-generated content can be considered owned by anyone, especially in cases where human interaction is minimal.

Moreover, some AI systems are trained in a specific way such that the model copycat a specific author. For example, can an AI model that produces music in the style of Pink Floyd be considered derivative work? If the answer is yes, are the possible negative consequences that this could have on the author taken into account?

Unfortunately, we still don't know how things will evolve and what is the future of AI systems. One thing is sure however, we need a balance between the development and capabilities of AI models and the protection of the rights of creators and users.

In section 4.4, we have seen that GitHub got sued over its AI system. The result of this legal case could lead to new regulations protecting people's data from being collected and used without their consent by companies.

Furthermore, to address concerns about the safety and ethical implications of artificial intelligence, the E. U. Council has proposed the Artificial Intelligence Act[12], which "aim to ensure that artificial intelligence (AI) systems placed on

[12] Artificial Intelligence Act: https://eur-lex.europa.eu/legal-

the EU market and used in the Union are safe and respect existing law on fundamental rights and Union values."[1].

With this proposal, the European Council wants to sort AI systems into categories based on their risk and to put limitations on high-risk systems that can violate fundamental rights. For example, they want to prohibit the use of AI systems for social scoring. Moreover, they want to support innovation through AI regulatory sandboxes. These sandboxes allow developers to validate and test AI systems in real world conditions. This Act could create an infrastructure that put ordinary people in a safer spot, not only by limiting high-risk AI systems but even by listing those systems in order to increase transparency. All of this without limiting the innovation of this new technology.

8 Conclusion

During this article, we have seen what GitHub Copilot is, how it works, and its problems from a technical and legislative perspective. Section 4 of this report examined the actual situation of GitHub, specifically what their claims are and why they are so criticized. In Section 5, we analyzed the same situation from another perspective, the users' point of view. We examined the complications that can arise when actively using Copilot and the possible consequences of improper use. In Chapter 6, GitHub Copilot has been compared to other similar software. This led us to understand why Copilot is the only one getting buzzed.

Speaking of the class action against GitHub, we still don't have a winner or a loser, not until the judge rules. It is possible that this lawsuit will take a while due to the delicate nature of the subject. In particular, the results of this process could change the future of AI systems. For example, if training an AI model on publicly available data will be judged as a violation of the fair use doctrine developers will have to find new ways to collect data. This could lead to smaller, less specific datasets and higher costs, and thus a general slowdown in the development of upcoming AI software.

However, it is different to say that something is illegal or non-ethical, and even if we cannot establish whether or not Copilot is illegal, I still believe that it is against ethical principles. It is an artificial intelligence model that harvests billions of lines of code from public repositories and distributes them through a paid software service. Particularly, the fact that the original authors get no acknowledgement when their code is given as an output destroys the purpose of licenses. This is the fear of the open source community, the fact that one can access almost every line of open source code without even knowing where it came from it's the exact opposite of what they want. In the worst case, people could lose interest in joining the open-source community, leading to less sharing and no contribution. Obviously, this is a bit utopian since there would be no data for Copilot to improve without open-source projects. This doesn't take away that Copilot can be a threat to open-source coders and their visibility.

content/EN/TXT/?uri=celex%3A52021PC0206

References

[1] Artificial Intelligence Act: Council calls for promoting safe AI that respects fundamental rights. https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights.

[2] Armin Ronacher's regurgitation case from GitHub Copilot. https://twitter.com/mitsuhiko/status/1410886329924194309.

[3] Directive 2001/29/EC of the European Parliament. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32001L0029.

[4] Foster, D. *Generative deep learning: teaching machines to paint, write, compose, and play.* O'Reilly Media, 2019.

[5] Franceschelli, G., and Musolesi, M. Copyright in generative deep learning. *CoRR abs/2105.09266* (2021).

[6] GitHub Copilot: Your AI pair programmer. https://github.com.

[7] GitHub Litigation. https://githubcopilotlitigation.com.

[8] GitHub Copilot research recitation. https://github.blog/2021-06-30-github-copilot-research-recitation/.

[9] GitHub Terms of Service on user generated content. https://docs.github.com/en/site-policy/github-terms/github-terms-of-service#d-user-generated-content.

[10] Nat Friedman statement on fair use in machine learning. https://twitter.com/natfriedman/status/1409914420579344385.

[11] Tim Davis' regurgitation case from GitHub Copilot. https://twitter.com/DocSparse/status/1581461734665367554.

[12] U.S. Copyright Office Fair Use Index. https://www.copyright.gov/fair-use/.