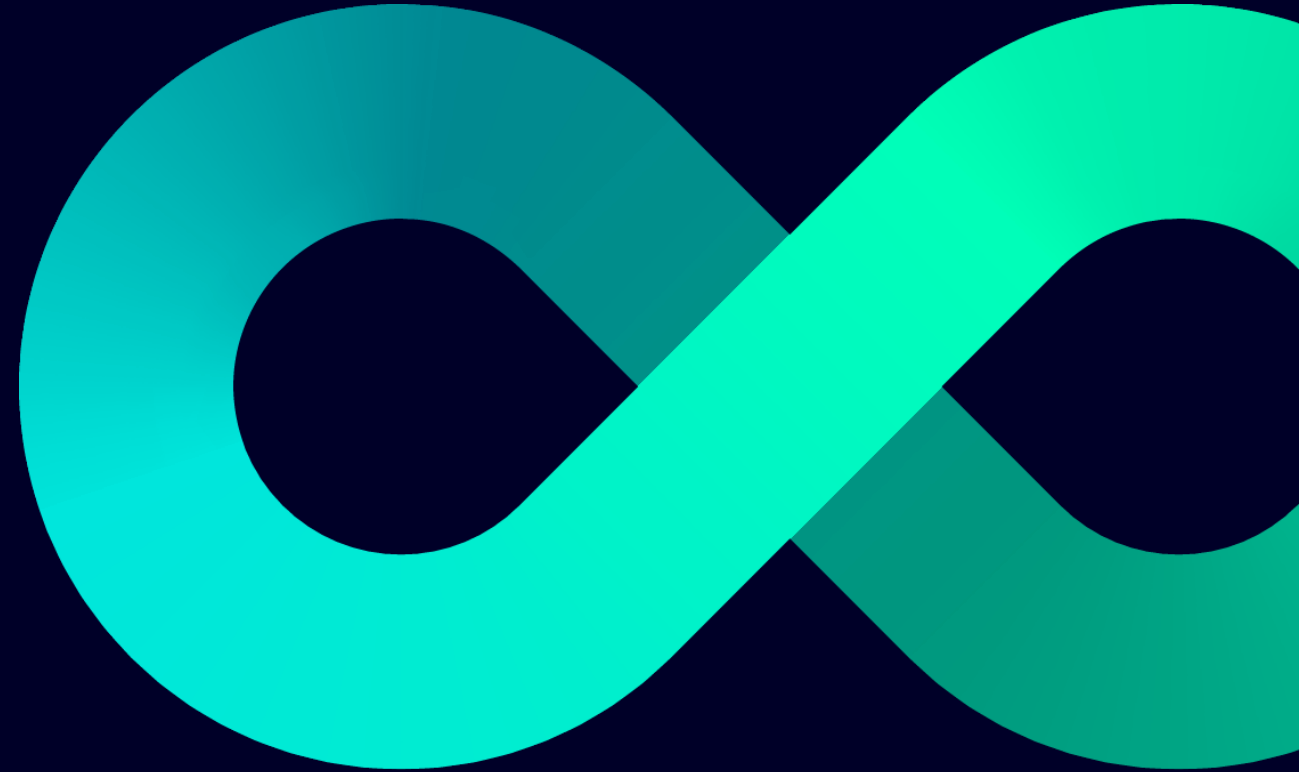


# Legal Implications of GitHub Copilot&Co.

Felix Mannewitz, OSS@Siemens

May 23, 2023



# \$ whoami \_

## **Felix Mannewitz**

Head of Legal - IT & Software Licensing

LC TEC IT&SL

Werner-von-Siemens-Str. 1

80333 München

Germany

Mobile +49 (172) 161 85 66

E-mail [felix.mannewitz@siemens.com](mailto:felix.mannewitz@siemens.com)

## Disclaimer

This presentation is not legal advice. I am a lawyer (IAAL), but I am not your lawyer. If you need legal advice, please get it from your own counsel. The content of this presentation and these slides reflect my own opinion and not (necessarily) that of my employer.

## Legal implications – The input side -

What is the legal basis for the use of training data and the creation of CODEX?

What does it mean for OSS?

### CODEX...

... is the **name of the AI model** that OpenAI created using natural language text and source code from publicly available sources, including vast amounts of source code in millions of public repositories on GitHub (most of them subject to different OSS licenses).

### In the EU:

An **explicit regulation for text and data mining** applies and GitHub has their T&Cs as an additional argument (for the GitHub-sourced data).

It is possible to **opt out** of text and data mining by using a machine-readable refusal. But this option was/is not used by the OSS-licensed projects on GitHub. Also, potential conflict with OSS principles.

### In the US:

GitHub relies on the **fair use argument** but has also the GitHub T&Cs as an additional argument for the code on public GitHub repos.

For the other data, only fair use argument available. It is an **open question** whether “fair use” will be accepted in this case, but GitHub at least has a strong argument regarding the required **"transformative"** nature of Copilot.

## Legal Implications – The output side 1/3 - Copilot could output infringing code, esp. verbatim copies of original code

- GitHub **claims** that all code is generated by the AI (=not copied) based on the CODEX model and the underlying “task” it was given.
- **But** GitHub states that in about 1% of the time, a suggestion may contain code snippets longer than ~150 characters that match the training set (=the original code used for training CODEX).
- That number went up from 0.1% claimed by GitHub in 2022.
- **OpenAI** clearly stated:  
*“Output generated by code generation features of our Services, including OpenAI Codex, may be subject to third party licenses, including, without limitation, open source licenses.”*




## Legal Implications – The output side 2/3 – Are the original licenses attached to the training data still applicable? (current assumptions)


### For verbatim copies or very similar:

If the code is copyrightable, then original **license obligations fully apply!**

This means, for example, that the respective attribution notices must be kept and/or provided for OSS portions.

### For other cases in which code is generated:

 In the EU: High likelihood that no further obligations for the generated code apply, if based on “lawful” text and data mining process.

 In the US: No obligations for the generated code, **if fair use applies.**

### Please note:

- The ongoing US lawsuits strongly dispute that “fair use” is applicable to Copilot.
- There are also claims that results are derivative works of the training material (e.g “Copyleft”). If such claim is successful, this would make the output in its current form infringing.

## **Legal Implications – The output side 3/3 – The code generated by GitHub Copilot is not protected by copyright in most jurisdictions\***

Except for few jurisdictions (UK, India, Ireland, Hong Kong, South Africa, New Zealand) copyright protection is **not** granted to machine-generated content, like output generated by Copilot.

### **Reason:**

Copyright generally requires an intellectual creation of a human. It does not apply to works that have been machine-generated. Currently the user has not enough capability to influence the output generation or insight into the process in a way that would justify the assumption that the generated work is protected by copyright.

### **Note:**

If a developer further refines/extends the machine-generated code then these changes themselves might enjoy copyright protection, if they reach the threshold of originality.

\* If it is not a copy of a protected work

## There are already lawsuits against GitHub Copilot in the US that challenge its legal foundation

- Federal class action complaints filed in California on Nov. 3 and Nov. 10, 2022.
- Lawsuits dispute that GitHub has a “fair use” argument to use public repositories on GitHub as training data.
- Claims that the original licenses of the Open Source Software used as training data do apply.
- Alleges copyright infringement because OSS attribution requirements are not fulfilled by Copilot (e.g., missing OSS license texts and copyright information).
- Various other claims (e.g., against DMCA, illegal removal of copyright information, etc.).

The lawsuits are in an early stage, outcome is open.

Status as of May 11, 2023:

- Some claims dismissed
- Most relevant claims to move forward (or plaintiffs have chance to amend their filing so that they can still move forward)
- **Breach of license claim to move forward**



## Upcoming regulatory topics and potential impact on OSS

Draft of EU AI Act (as of May 9, 2023):

- Regulation “...*shall not apply to AI components provided under free and open source licences except to the extent they are placed on the market or put into service by a provider as part of a high-risk AI system or of an AI system that falls under Title II or IV. **This exemption shall not apply to foundation models as defined in Art 3.***”
- Article 28b: Providers of “Foundation Models” are subject to regulation and must make sure that they comply with all requirements before making them available or putting them into service. Means provider must fulfill catalog of obligations as defined in Sec. 2 of Art. 28b.
- Applies also for models subject to OSS licenses.