THE PLUS IP FIRM
Patents.Trademarks.Copyrights.

HOT TAKE: WHY OPENAI'S CHATGPT IS NOT INFRINGING ON COPYRIGHTS

DEREK FAHEY, ESQUIRE
Derek is a registered patent attorney and one of managing partners of The Plus IP Firm, PLLC. Derek's practice is focused on creating, monetizing, and protecting intellectual property assets.
Derek@plusfirm.com
www.plusfirm.com
Youtube Channel
800.768.9399 ext. 001

Currently, one of the biggest questions for copyright stakeholders and the intellectual property legal community is if artificial intelligence large language models ("LLMs"), like OpenAI's ChatGPT, is infringing on the copyright of others because of their use of copyrighted material to train the LLMs. This article will look at how LLMs use copyrighted data, OpenAI's arguments as to why its use of copyrighted material is not copyright infringement, and the cases that may support OpenAI's arguments.

I.CHATGPT'S PROCESS EXPLAINED

My understanding of language models like ChatGPT is that ChatGPT does not directly copy and store individual copyrighted works into its databases. Instead, the process involves training these models on large, diverse datasets sourced from the internet and other publicly available texts. Here's how it generally works:

Data Collection: OpenAI collects vast amounts of text data from a wide variety of sources available on the internet. This includes books, websites, articles, and other types of written content.

Training the Model: The collected data is used to train the AI model. During this training process, the model learns patterns, structures, and nuances of language by analyzing the data. It's important to note that this training involves statistical analysis and pattern recognition across the entire dataset, rather than copying specific texts or works into a database. This is a key factor to OpenAI's arguments as to why its use is "fair use", which I will further explain below.

Generative Process: Once trained, the model can generate text based **on the patterns it has learned**. The output is not a reproduction of specific texts from the training data but is generated anew each time based on the input prompt and the model's training.

No Retrieval of Specific Texts: The model does not have the capability to retrieve and display specific articles, books, or other copyrighted works from its training data. It generates responses based on its understanding and synthesis of the information it has been trained on.

Therefore, the process is more about teaching the AI how language is used and structured rather than storing or replicating specific copyrighted works. The design focuses on understanding and generating language in a broad sense, rather than on retaining or recalling individual texts.

II.OPENAI'S FAIR USE ARGUMENT

OpenAI argues that its use of copyrighted content is fair use under United States copyright law. 17 U.S.C. § 107 establishes the fair use defense to copyright infringement. The statute instructs courts to consider the following factors:

> (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.

Regarding transformative Use, OpenAI argues that its models transform the copyrighted material in a significant way. The training process involves analyzing and learning patterns from large datasets, but the output (like the text generated by ChatGPT) is not a direct copy of any specific source. Instead, it's a new, unique piece of content that reflects a synthesis of information and patterns learned from numerous sources.

Regarding the nature of the copyrighted work, OpenAI argues that it uses a wide range of texts, including both factual and creative works. The use of factual works (like scientific papers or news articles) is often seen as more favorable under fair use than the use of highly creative works (like novels or songs).

Regarding the amount and substantiality of the Portion Used, OpenAI argues that while the training process involves large datasets, no single work is predominantly or exclusively used. The model does not rely on substantial portions of individual copyrighted works, but rather on broad patterns learned from massive, diverse corpora.

Regarding the effect on the market, OpenAI argues that its models do not replace or diminish the market for the original works. The outputs from models like ChatGPT are not substitutes for reading the original texts. Instead, they are often used for purposes like education, research, or generating new content, which can be different from the purposes of the original works.

III.OPENAI CAN ARGUE THE INTERMEDIATE COPYING PRINCIPLE

Existing case law from the Ninth Circuit Court of Appeals and the Supreme Court support a fair use defense for OpenAI's use of copyrighted material. In these cases, courts recognized the concept of intermediate copying as potentially qualifying for fair use, particularly when the copying is a necessary

step in creating a new product that does not infringe upon the copyright owner's rights.

This principle was evident in cases where defendants reverse-engineered software to understand its functional requirements for compatibility purposes or to create a new product, and where the final product did not contain the copyrighted material.

In *Sega Enterprises Ltd. v. Accolade, Inc*., the case revolved around Accolade's method of reverse engineering Sega's game console to develop compatible games without Sega's licensing. 977 F.2d 1510, 1521 (9th Cir. 1992). Accolade purchased Sega's game cartridges, copied them to understand the console's operational requirements, and then created their own games. Although this process involved copying Sega's copyrighted material, the Ninth Circuit found that such intermediate copying was a necessary step for developing non-infringing, independent software, thus falling under fair use due to its transformative purpose and the absence of direct market competition with Sega's games.

In *Sony Computer Entertainment v. Connectix Corp*., the case centered on Connectix's creation of the Virtual Game Station ("VGS"), a PlayStation emulator that allowed PlayStation games to be played on Macintosh computers. 203 F.3d 596 (9th Cir. 2000). Connectix used intermediate copying by reverse engineering Sony's PlayStation console to understand its functionality without accessing the proprietary BIOS code directly. Although initial versions of VGS used a copy of Sony's hardware and firmware (BIOS), later versions included a "clean room" reverse-engineered version. The court acknowledged this intermediate copying as part of a fair use analysis, recognizing the transformative nature of creating a new product that allowed games to be played on a different platform.

In *Google LLC v. Oracle Am., Inc*., the Supreme Court ruled in favor of Google, stating that its use of Java API declaring code constituted fair use. 141 S. Ct. 1183 (2021). The court considered the transformative nature of Google's use, emphasizing the role of APIs in enabling interoperability and the development of new programs. This decision underscored the importance of balancing copyright protection with fostering innovation and compatibility in the software industry, recognizing the need for legal frameworks to adapt to technological advancements.

For OpenAI, the intermediate copying argument could support a fair use defense by demonstrating that any copyrighted material used during training is analogous to reverse engineering for compatibility or innovation purposes. Specifically, OpenAI could argue that:

- The use of copyrighted material is used to establish patterns, which is not protected by copyright law.
- The use of copyrighted material is transformative, serving a new purpose distinct from the original works.
- The use is necessary to develop and train AI models, which are innovative products contributing to technological advancement and public knowledge.
- The final outputs of OpenAI's models do not reproduce the copyrighted material but generate new, original content based on learned patterns and information.
- These arguments rely on demonstrating that OpenAI's use of copyrighted materials is for the purpose of creating new, transformative works that contribute to the advancement of knowledge and technology, rather than for the purpose of replacing or competing with the original copyrighted works.

There is no doubt that OpenAI's ChatGPT and other LLMs are powerful and useful technologies. At the same time, Copyright stakeholders are concerned that these LLMs are using their copyrighted material without permission; and thus, not being properly compensated. Several cases are currently pending that are likely to address OpenAI's fair use defense. It will be interesting to see how this will play out. For more information about Derek Fahey, this article's author, click HERE.