Intellectual Property & Technology Law Journal

Edited by the Technology and Proprietary Rights Group of Weil, Gotshal & Manges LLP

VOLUME 36 • NUMBER 2 • FEBRUARY 2024

AI Copyright Contributory Infringement and the Fair Use Defense – Part I

By Jon Grossman & Scarlett L. Montenegro Ordonez

Since the introduction of various artificial intelligence (AI) tools, there has been a slew of AI intellectual property (IP) infringement cases. For the most part, these infringement claims have been directed at the AI tool developers – namely those who have had a hand in creating AI data sets and related instructions which automatically direct the scraping done by the AI program.

But what about the AI user? Does his or her use of an AI tool, such as Chat GPT, CoPilot or Stable Diffusion, constitute a form of direct or indirect copyright infringement for content that was impermissibly copied?

And, if liability attaches, is there an available fair use defense?

This two-part article looks at both of those subjects. This first part considers the situation of the AI user and copyright contributory infringement. The second part of this article, to be published in the next issue of the *Intellectual Property & Technology Law Journal*, examines the fair use defense.

THE TECHNOLOGY

At its simplest form, AI is a technology that combines computer science and robust datasets to enable problem-solving. AI is almost breath-taking in its scope encompassing the fields of machine learning and deep learning. There are many types of AI algorithms and applications that have been and continue to be routinely used. For purposes of this article, we will mainly concern ourselves with AI deep learning.

Chat GPT, for example, like a lot of new AI products, was built on the shoulders of years of AI development. ChatGPT comprises a Chatbot, which is a software or computer program that simulates human conversational like text or voice. ChatGPT also employs deep learning, which refers to a neural network of more than three layers. Chat GPT's deep-learning models process unstructured raw data (by way of example, all of the collected works of an artist), and generate statistically probable outputs.

Known as generative models, ChatGPT encodes simplified representations of its data sets by drawing new results that have a high statistically probable representation of the original data. Chat GPT also utilizes a large language model (LLM) neural network trained through data input/output sets where the text is unlabeled or uncategorized, and the model involves a self-supervised or semi-supervised

Jon Grossman, is senior counsel and Scarlett L. Montenegro Ordonez is an associate in the Washington, D.C., office of Blank Rome LLP. Jon practices in the areas of patent law, licensing work, and copyright with an emphasis on computer software issues. He may be contacted at jon.grossman@blankrome.com. Scarlett works on IP transactional and litigation matters. Scarlett can be contacted at scarlett.montenegro@blankrome.com.

learning methodology. Information is ingested, or content is entered into the LLM, and the output is what that algorithm predicts the next word will be. The input can be proprietary corporate data or, in the case of ChatGPT, publicly available information, such as data scraped directly from the internet. LLMs are controlled by parameters which help the neural network decide between different answer choices.

In the case of ChatGPT, those parameters are sizable. OpenAI's GPT-3 LLM upon which ChatGPT is built, has 175 billion parameters, and the company's latest model – GPT-4 – is purported to have in the neighborhood of 1 trillion parameters.

These new AI products have been controversial particularly since they depend heavily on massive datasets of scraped data. The scraped data sometimes involves the copyrighted works of many creators. Not surprisingly, there has been a slew of recent copyright infringement actions where the infringement allegations focus on claims that the AI product impermissibly scraped copyrighted content from the internet without permission from the content owners.

The *Trembly* lawsuit is a typical example:"OpenAI made copies of Plaintiffs' books during the training process of the OpenAI Language Models without Plaintiffs' permission. Specifically, OpenAI copied at least Plaintiff Tremblay's book *The Cabin at the End of the World*; and Plaintiff Awad's books *13 Ways of Looking at a Fat Girl* and *Bunny*."[1]

Scraping is a form of neural network training which teaches the neural network to perform a task. Neural networks learn by initially processing several large sets of labeled or unlabeled data. By using these examples, these networks can learn to ultimately process unknown inputs more accurately. AI-powered web scraping uses numerous algorithms tied to forms of AI like, machine learning, natural language processing, and computer vision processes in order to automate data extraction from various websites.

DATA SETS

Often these data sets form a relatively complex web, some created by the AI product maker and some by third parties. "[I]n its June 2018 paper introducing GPT-1 (called "Improving Language Understanding by Generative Pre-Training"), OpenAI revealed that it trained GPT-1 on BookCorpus, a collection of "over 7,000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance." OpenAI confirmed why a dataset of books was so valuable: "Crucially, it contains long stretches of contiguous text, which allows the generative model to learn to condition on long-range information."[2]

Hundreds of large language models have been trained on BookCorpus, including those made by OpenAI, Google, Amazon, and others. BookCorpus, however, is a controversial dataset. It was assembled in 2015 by a team of AI researchers for the purpose of training language models. Those researchers copied the books from a website called Smashwords.com that hosts unpublished novels that are available to readers at no cost. The copied novels, however, are largely copyrighted. They were copied into the BookCorpus dataset without consent, credit, or compensation to the authors."[3]

Accordingly, scraping creates infringement risks because the chatbot LLM programs, such as Chat GPT, allegedly may copy copyrighted data while being trained to create human-like responses. The scraped information comes from lots of sources including libraries, statistical databases, social media, Internet searches, radio, television, and podcasts.

THE COMPLAINTS

In another example, Getty Images sued Stability AI for copying millions of its photos without a license and using them to train its generative AI tool: Stable Diffusion.[4] According to the Getty complaint, the use of Getty's images enabled Stable Diffusion to generate more accurate image depictions based on user prompts. "Rather than attempt to negotiate a license with Getty Images for the use of its content, and even though the terms of use of Getty Images' websites expressly prohibit unauthorized reproduction of content for commercial purposes such as those undertaken by Stability AI, Stability AI has copied at least 12 million copyrighted images from Getty Images' websites, along with associated text and metadata, in order to train its Stable Diffusion model."[5]

Getty also asserted that it had licensed millions of suitable digital assets to other leading technology innovators for AI-related purposes, and that the Stable Diffusion competes with it unfairly. Getty indicated that its damages for willful infringement could go as high as $1.8 trillion.[6]

Other cases include a case filed against Google in July 2023.[7] As reported by CNN on July 12, 2023, Google was sued in a proposed class action for allegedly misusing large amounts of personal information and copyrighted material to train its artificial intelligence systems.[8] The complaint was filed by eight individuals seeking to represent millions of internet users and copyright holders. The complaint states that in order to train its AI products, including Google's Bard, it "has been secretly stealing everything ever created and shared on the internet by hundreds of millions of Americans" and using this data to train its AI products, such as its chatbot Bard.[9] The complaint also claims Google has taken "virtually the entirety of our digital footprint," including "creative and copywritten works" to build its AI products.[10]

Another lawsuit similar in nature to the Google's Bard lawsuit was filed in late September by the Author's Guild and 17 writers against OpenAI, the developer of ChatGPT.[11] As reported on Verge on September 20, 2023, the complaint alleges that OpenAI conducted wholesale copying of content into its model and which could produce derivative works which could impact each author's future market.[12]

Another complaint was filed in November 2022 by a small group of individual software developers alleging that GitHub, Microsoft, and OpenAI violated copyright, contract, privacy, and business laws, by using their source code culled from the GitHub's open source platform to create the OpenAI's Codex machine learning model and GitHub's Copilot programming assistant.[13] The complaint was later amended in June 2023 to reportedly only assert eight counts including accusations of violating the Digital Millennium Copyright Act, breach of contract, unjust enrichment, and unfair competition claims. It also adds allegations of intentional interference with prospective economic relations and negligent interference with prospective economic relations.[14] The software developers allege that the Codex and Copilot tools were created from their code, and sometimes reproduced it, without explicit permission or concern for the terms under which they licensed their works to Github.

The judge rejected the defense motion to dismiss the plaintiffs' claim that Codex's capacity to reproduce code represents a breach of software licensing terms.

In early September of last year, in a move to assure end users, Microsoft announced that it was indemnifying end users from copyright infringement in connection with their use of their CoPilot product.

Most recently (December 27, 2023) the New York Times sued OpenAI and Microsoft alleging that the companies were infringing its copyrights by training their AI, such as ChatGPT, on their articles.[15] In response, OpenAI published a blog post where it argued that "using publicly available internet materials is fair use."[16]

⋆ ⋆ ⋆

Editor's note: The conclusion of this article will appear in the next issue of the *Intellectual Property & Technology Law Journal*.

Notes

1. Trembley et al. v. Open AI, Inc., et al., Civ. No. 3:23-cv-03223 (N.D. Cal. 2023).
2. Compl. ¶ 28, Tremblay v. OpenAi, Inc., 3:23-cv-03223-AMO, (N.D. Cal. June 6, 2023).
3. Id. at 5-6.
4. Getty Images Inc. v. Stability AI, Inc., Case 1:23-cv-00135-UNA (D. Del. 2023).
5. Id. at 3.
6. Matt O'Brien, Photo giant Getty took a leading AI image-maker to court. Now it's also embracing the technology, AP News, Sept. 25, 2023.
7. J.L et al v. Alphabet Inc. et al. Case No. 3:23-cv-3440, (N.D. Cal. 2023).
8. Catherine Thorbecke, Google hit with lawsuit alleging it stole data from millions of users to train its AI tools, CNN, July 12, 2023.
9. Id.
10. Id.
11. Authors Guild et al. v. OpenAI et al., Civ. No. 1.23-cv. 8292 (S.D.N.Y. 2023).
12. Alex Castro, George R.R. Martin and other authors sue OpenAI for copyright infringement, Verge, Sept. 20, 2023.
13. J. Doe 1 et al. v. GitHub Inc. et al., Case c4.22-cv. 6823 JST (N.D. Cal. 2023).
14. See Thomas Claburn, GitHub accused of varying Copilot output to avoid copyright allegations, The Register, June 9, 2023.
15. Cade Metz, OpenAI Says New York Times Lawsuit Against It Is 'Without Merit', NYT (Jan. 8, 2024,

Updated Jan. 9, 2024, 1:55 PM), https://www.nytimes.com/2024/01/08/technology/openai-new-york-times-lawsuit.html.

[16] OpenAI, OpenAI and journalism, OPENAI (Jan. 8, 2024), https://openai.com/blog/openai-and-journalism#OpenAI.

Wolters Kluwer