THE BATTLE OVER AI TRAINING DATA: COPYRIGHT, FAIR USE, AND THE FUTURE OF GENAI

By Diego F. Freire

The future of Generative Artificial Intelligence (GenAI) is under scrutiny following a copyright lawsuit The New York Times filed against OpenAI. The case involves using copyrighted works to train GenAI models and their ability to generate near copies of their training data. GenAI models are trained on vast amounts of data, including the use of copyrighted works. Several cases, such as Getty Images v. Stability AI and Silverman v. OpenAI, have previously raised questions about using copyrighted works in training GenAI models and how the generated content may resemble copyrighted works. Previously, GenAI companies have relied on the fair use doctrine to shield them from copyright claims. However, plaintiffs like the Times argue that using copyright works to train and the memorized content generated (content that may be very similar to the original work based on model memorization) does not fall under fair use. These copyright cases become even more critical as the fair use doctrine was recently revised by the United States Supreme Court in their Andy Warhol v. Goldsmith decision, placing the future of GenAI at risk.

GenAI companies utilize and train models like Large Language Models (LLMs) to generate text (used by OpenAI in ChatGPT), and Diffusion Models to generate images (used by Stability AI). These LLMs are trained on vast text datasets to master language patterns and generate coherent text by predicting word sequences. In contrast, Diffusion Models generate images by learning to transform noise added to human generated images into detailed images. Both models rely on training data, statistical probabilities, and iterative processes to generate contextually appropriate text or images. However, because of the need to use large amounts of data and the importance of using human-generated data to train such models (for example, ChatGPT-3 was reportedly trained on the equivalent of 45 million books), GenAI companies use copyrighted works to train their models. Not only that but because of the complexity of the models (for example, ChatGPT-3 reportedly has 175 billion parameters), the models may "memorize" portions of their training data and generate content that may resemble copyrighted works. Thus, companies like the Times and Getty Images, which create large amounts of copyrighted works, allege that the unauthorized use and reproduction of their works by these models is copyright infringement.

In the case of Getty Images, the company alleges Stability AI used more than 12 million of its copyrighted images without permission to train its models. Their models allegedly generated images very similar to their copyrighted works. For example, Stability AI's models allegedly generated images that included a watermark very similar to the watermark Getty Images' uses on their photographs.

Thus, Getty Image argues that by using its images to train their AI, Stability AI not only infringed upon the copyrights of those individual images but also used them to create a competing business.

In the Times case, the Times has cited instances where ChatGPT has provided users with near-verbatim excerpts from its articles. The Times contends that such use of its content diminishes the need for readers to visit its website directly, potentially impacting its traffic and revenue. OpenAI, in its defense, has argued that the examples of alleged infringement cited by The Times are caused by deceptive prompts that violate ChatGPT's terms of use. GenAI companies, like OpenAI, argue that the use of copyrighted works to train GenAI models falls under fair use as being transformative use for public enrichment and, therefore, not copyright infringement.

The fair use doctrine, therefore, becomes critical for GenAI. The fair use doctrine allows for the limited use of copyrighted material without permission from the copyright holder under certain conditions. Courts weigh factors such as the purpose and character of the use, the nature of the copyrighted work, the amount and significance of the portion used, and the effect of the use upon the potential market for or value of the copyrighted work. In the past, tech companies relied on their use of data as transformative and, therefore, fair use. However, this may no longer be true after the Warhol decision. Writing on behalf of a seven-member majority, Justice Sonia Sotomayor determined that Warhol's use of an image adapted from Goldsmith's photograph of Prince did not constitute fair use of the copyrighted material. Although the Court acknowledged Warhol's artwork as transformative, a characteristic previously sufficient for a fair use defense, it did not qualify as fair use because both the original photograph and the painting served substantially the same commercial purpose.

GenAI, heralded for its potential to mirror the revolutionary impact of the printing press, now stands at a crucial juncture. The future of GenAI companies, specifically leading GenAI models, depends on access to copyrighted works, and lawsuits like The New York Times v. OpenAI, along with similar litigations, threaten that access. This is particularly true in light of the recent Supreme Court decision in Andy Warhol v. Goldsmith, which emphasized that transformative use alone does not qualify as fair use if the original work and use both serve similar commercial purposes. Thus, this critical juncture highlights the imperative for stakeholders from all sectors—developers, copyright holders, and Congress —to unite in crafting a collaborative way forward. This partnership is crucial not just for the survival of GenAI but also for the development of Artificial General Intelligence (AGI).

Diego F. Freire *is an associate in Dykema's IP Litigation Group, specializing in IP law, encompassing patent litigation, prosecution, and transactional matters. Diego has experience helping clients protect, monetize, and enforce IP rights across various technical fields, such as AI, blockchain, quantum computing, and other electrical and software technologies.*