Westlaw Today
powered by Reuters

Recent litigation disputes involving generative AI — considerations for technology owners developing and licensing AI innovations

By Bryan Mechell, Esq., Robins Kaplan LLP

OCTOBER 13, 2023

Technology and software license disputes involving intellectual property and contract rights carry significant risk in terms of potential business disruption and damages. The increasing implementation and rapidly evolving uses of generative AI at software and technology companies could lead to increased disputes over the use of protected data to train generative AI models as well as ownership of outputs.

For example, generative AI models can be exceptionally powerful because they can learn from very large datasets — such as the internet — but those datasets can be interspersed with copyrighted and other protected material.

Proposed class action lawsuits filed this year against Github, Stability AI, OpenAI and Meta — including recent actions filed by George R.R. Martin, John Grisham, Pulitzer Prize winner Michael Chabon, comedian and author Sarah Silverman, and various other authors against OpenAI and Meta — raise important questions about liability for unauthorized use of copyrighted materials to train generative AI models without consent, credit, or compensation, as well as questions about ownership of generative AI outputs.

For intellectual property owners protecting their generative AI innovations, as well as end users licensing generative AI tools, these lawsuits underscore the importance of closely monitoring the composition of generative AI training data sets, scope and content of outputs, and license terms regulating use of these rapidly evolving technologies.

GitHub litigation — code generation from open-source code

On November 3, 2022, a class action complaint was filed against GitHub, Microsoft, OpenAI, and related corporate groups.[1] Beginning with an overview section entitled "A brave new world of software piracy," the complaint alleges that the Defendants trained Codex and Copilot (coder-assisting generative AI programs) on public code that was protected by open-source licenses, but the AI does not provide attribution of authorship or copyright when outputting that code.

This, the complaint alleges, amounts to Digital Millennium Copyright Act (DMCA) violations, breaches of contract, unlawful competition, and privacy violations under California law.

On May 11, 2023, the court denied portions of the Defendants' motions to dismiss, leaving various remaining claims for future resolution.[2] On the issue of standing to file suit, the court's order noted that Plaintiffs did not sufficiently allege injuries to their privacy rights because they did not allege that personal information was being improperly reproduced or distributed.

It will likely be some time before various ongoing litigations provide clarity on important questions raised about the full scope of IP protections that apply to training data sets.

On the other hand, the court ruled that the Plaintiffs had standing to pursue injunctive relief because "the Court [could] reasonably infer that, should Plaintiffs' code be reproduced as output, it [would] be reproduced in a manner that violates the open-source licenses under which Plaintiffs published their code."[3]

Notably, the Court found that the Plaintiffs did not state an injury resulting from the Defendants' use of licensed code for ***training*** the generative AI — e.g., a use that allegedly could have constituted a breach of the open-source licenses at issue — so the Court did not address whether the use of licensed code for AI training would be sufficient to confer standing.[4]

Regarding the DMCA claims, the court declined to dismiss the Plaintiffs' claims under 17 U.S.C.A. § 1202(b)(1) (prohibiting removal or alteration of copyright management information) and 1202(b)(3) (prohibiting the knowing distribution of copyrighted works that have been altered or stripped of copyright management information). The court found the Plaintiffs' allegations sufficient to survive the

motion to dismiss stage that the Defendants had some knowledge that its generative AI would produce information without proper copyright management information, violating the DMCA.

Regarding the alleged breach of GitHub "open-source licenses," the court found that the Plaintiffs sufficiently alleged that the license agreements governed the public code and reproduction of that code would require proper attribution. Again, the court did not address — because the Plaintiffs did not allege — whether using the licensed code to train the AI could constitute a breach of the license agreement.

Stability AI litigations — image generation from copyrighted pictures

On January 13, 2023, three artist plaintiffs filed a class action complaint against Stability AI, Midjourney Inc., and DeviantArt, Inc.[5] The complaint alleges that DreamStudio, the Midjourney Product, and DreamUp are all image-generating AI programs built on Stability's Stable Diffusion program.

Stable Diffusion, the complaint alleges, is a 21st century collage tool responsible for training, collection, and compression of images across the internet without permission from the artists. According to research cited in the complaint, diffusion is a machine-learning technique for algorithms to copy, learn, and reconstruct images from its own training set.

The complaint asserts that these programs produce images that are exclusively derived from copyrighted images. The complaint further alleges that the library used to train Stable Diffusion — paid for by Stability — contains 5.85 billion images indiscriminately scraped from the internet.

The complaint asserts multiple counts against Stability AI and the other defendants, including: (1) direct copyright infringement, (2) vicarious copyright infringement, (3) violation of the DMCA by falsifying and removing/altering copyright management information, (4) violation of California state laws governing the right of publicity, (5) violation of the common law right of publicity, (6) unfair competition, and (7) breach of contract.

In a similar action, Getty Images filed suit in February 2023 against Stability AI alleging that Stability AI copied more than 12 million photographs from Getty Images's collection and used them without permission to train the Stability AI to generate more accurate depictions based on user prompts.[6] The Getty Images lawsuit asserts bases similar to those asserted in the pending class action against Stability AI et al., and also includes claims of trademark infringement, trademark dilution, deceptive trade practices.

OpenAI and Meta litigations — text generation from copyrighted books

Comedian and author Sarah Silverman joined two other authors in proposed class action lawsuits filed on July 7, 2023, against OpenAI and Meta platforms.[7] The complaints tell a story similar to the GitHub and Stability AI complaints — in this case that OpenAI developed generative AI models (including GPT-1, GPT-2, GPT-3, and GPT-4) that were trained on legally protected copyrighted works without consent, credit, or compensation. The complaints allege that OpenAI's AI and Meta's AI were trained from a massive dataset that contained books from a "shadow library" containing some of the Plaintiffs' copyrighted works.

The complaints assert multiple causes of action, including: (1) direct copyright infringement, (2) vicarious copyright infringement, (3) unauthorized removal of copyright management information under the DMCA, (4) unfair competition, (5) unjust enrichment, and (6) negligence.

OpenAI has moved to dismiss the bulk of the claims — the "heart" of which it argues are copyright claims — on the basis that they "misconceive the scope of copyright, failing to take into account the limitations and exceptions (including fair use) that properly leave room for innovations like the large language models now at the forefront of artificial intelligence."[8]

Other big-name authors — such as George R.R. Martin, John Grisham, and Pulitzer Prize winner Michael Chabon — have also recently filed class action complaints against OpenAI for using copyrighted works without permission to train generative AI models.

The complaint filed by George R.R. Martin, John Grisham and others against OpenAI on September 19, 2023, includes numerous examples where, when prompted, ChatGPT accurately generated summaries and outlines of the plaintiffs' copyrighted works, and asserts that ChatGPT could not have generated these results if OpenAI's LLM ("Large Language Model") had not ingested and been "trained" on the copyrighted works.[9] The complaint asserts claims of direct, vicarious, and contributory copyright infringement, and seeks damages attributable to the infringement.

The proposed class action filed by Michael Chabon and other writers against OpenAI on September 8, 2023, for example, asserts claims of copyright infringement, unauthorized removal of copyright management information, unfair competition, negligence and unjust enrichment.[10] The complaint alleges that OpenAI uses copyrighted works in its training datasets that are built by scraping the internet for text data — which necessarily leads OpenAI to capture, download and copy copyrighted written works, plays, and articles.

It also alleges that the data includes copyrighted materials from the Standardized Project Gutenberg Corpus or Project Gutenberg itself, as well as "shadow libraries" containing massive collections of pirated books (e.g., Library Genesis ("LibGen"), Z-Library, Sci-Hub, and Bibliotik).

Considerations for technology owners developing and managing generative AI

It will likely be some time before various ongoing litigations provide clarity on important questions raised about the full scope of copyright and other IP protections that apply to training data sets, as well as the input queries and generated outputs of generative AI tools. Legislation and regulations at the federal level appear likely, but are in varying early stages of development.

In the meantime, technological innovation and progress continues and intellectual property owners protecting their generative

AI innovations — as well as end users licensing generative AI tools — should consider the following key issues.

What is the training set?

Generative AI disputes frequently involve disputes about the underlying data used to train an AI model. A key feature of generative AI systems is their ability to consume an enormous amount of data, which makes up the entire universe of a system's initial knowledge. If that initial training set contains legally protected materials, a system may be unable to readily ignore or unlearn from materials already trained on. This is one reason why the integrity of training sets is critical to the real-world performance of an AI model.

Just like custodians of public libraries, owners of AI model training sets should be mindful of their obligations to the owners and authors of materials in those libraries. The GitHub and Stability AI complaints, for example, not only assert wrongful conduct by the users of protected material, but also by the managers of the protected materials. Much of the GitHub complaint centers around 11 different kinds of "open-source licenses" that allow programmers to publicly share code but also retain attribution for their work.

Similarly, the Stability AI complaint suggests that DeviantArt betrayed its artist contributors by creating its own generative AI built on DeviantArt's own image library. These pending litigations therefore suggest that owners and managers of public or private data should be cautious when using or licensing data for AI training.

To mitigate potential risks associated with leveraging public or private date for generative AI applications, software licensees and licensors involved in developing and managing generative AI products should develop a full understanding of the underlying training set. For starters, this includes maintaining a detailed record of the sources, libraries, metadata, and the compositions of each — which provides the basic materials needed to assess risks associated with an AI system trained on protected materials.

For software owners developing and managing home-grown tools that implement generative AI, this can also include a process for flagging material that is known to be legally protected, as well as tracking metadata, so that licensors and licensees have a clear accounting of the content of the training set. This might allow end-users, for example, to toggle between datasets containing a determinate number of flags.

In addition, in cases where the training data set often has significant business value — which it frequently does — technology owners and developers of generative AI tools should pay special attention to implementing provisions in master service agreements and EULAs that clearly articulate the scope of authorized uses, restrictions, warranties, and attribution for underlying content. The scope of allowed uses and access to the underlying training data set could have significant implications to available IP protection.

Despite the importance of the training set to the performance of generative AI models, courts have not yet squarely addressed whether the specific ***act of training*** an AI on protected materials is an infringing offense. That said, currently-pending proposed class actions against OpenAI and Meta may ultimately provide guidance addressing questions around the training process — at least in the context of alleged DMCA violations — where plaintiffs allege that OpenAI and Meta violated the DMCA by removing copyright management information from copyright protected works during the training process.

Importantly, the inquiry into the act of training a generative AI model on protected data may vary for different kinds of generative AI based on license terms attaching to the underlying dataset, and because what the AI model outputs after it has been trained will impact a court's analysis.

What is the output?

One of the primary values of generative AI models lies in the quality and accuracy of the system's output — which are both directly correlated to the quality of the underlying data used to train the model.

Pending litigations involving generative AI assert that the owners of these AI models — such as OpenAI, Meta, and others — reap significant benefits by training their models using quality copyright protected materials. For example, if OpenAI's ChatGPT is prompted to generate writing in a style of a specific author, it is capable of outputting a response based on patterns learned from analysis of that author's work available within the applicable training set.

The quality of the underlying material improves ChatGPT's ability to generate realistic, convincing responses. Accordingly, plaintiffs in these actions assert not only that ChatGPT itself is an infringing derivative work, but also that the text responses generated by ChatGPT for the end user are infringing derivative works.

Technology owners tasked with developing and managing generative AI tools should pay special attention to potential use of protected material in training data sets, and should consider implementing protocols to ensure proper and accurate attribution for any protected work used to train the underlying model.

While the outcome of pending litigations addressing this issue remains unclear, this could help mitigate potential risk of DMCA-related violations involving, e.g., alleged alteration or removal of copyright-management information from copyright protected material used to train an AI model, or failure to provide proper attribution of underlying open-source code such as in the GitHub litigation.

Another approach to risk management on the generative AI output side could include automated review and flagging of generative AI responses. For example, a generative AI system could incorporate a robust gatekeeper that checks for potential protected material in generated responses. This might help the system apply proper attribution to underlying protected materials, and in appropriate cases, could flag problematic material for further review.

Is fair use a defense for generative AI tools using copyrighted materials?

The "fair use" defense to copyright infringement is one battleground issue for generative AI disputes. Section 107 of the Copyright Statute

establishes four factors that a court considers when determining if the use of a protected work is fair:

(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.

OpenAI's arguments regarding fair use in the ongoing litigations underscore significant questions about the applicability and scope of fair use defense articulated in Section 107 of the Copyright Statute as it applies to generative AI. For example, OpenAI argues that courts have recognized "that use of copyrighted materials by innovators in transformative ways does not violate copyright" and that this is a key legal principle "upon which countless artificial intelligence products have been developed by a wide array of technology companies."[11]

Citing the U.S. Supreme Court's recent decision in *Google LLC v. Oracle*, OpenAI asserts that it is not an infringement to create wholesale copies of a work as a preliminary step to develop a new, non-infringing product, even if the new product competes with the original.[12]

It will likely be some time until the courts, regulatory bodies, and perhaps Congress provide guidance on the scope of Section 107 of the Copyright Statute as it applies to generative AI. In the meantime, technology owners tasked with developing and managing generative AI tools should consider the extent to which those tools transform any protected material as part of generating the output.

For starters, verbatim regurgitation of protected material may have a higher likelihood of falling outside the scope of a fair use defense. And other factors, such as how the copyrighted materials is used, the nature of the underlying work, how much of the work is used, commercial intent, and the effect of the use on the on the value of the underlying work are all important considerations.

Closing thoughts

Recent litigation disputes involving generative AI underscore important legal questions about the scope of copyright and other IP protections relating to generative AI models that could have significant implications for technology and software licensing. Technology and software companies developing and licensing generative AI innovations should closely monitor, catalogue, and assess training data used by generative AI tools, pay special attention to license terms that outline authorized uses and protect IP rights, and develop a robust review process for monitoring compliance with developing laws and regulations.

Notes

[1] *Doe 1 v. GitHub Inc.*, No. 22-cv-6823, *complaint filed*, 2022 WL 16743590 (N.D. Cal. Nov. 3, 2022).

[2] *Doe 1 v. GitHub Inc.*, No. 22-cv-6823, 2023 WL 3449131 (N. D. Cal. May 11, 2023).

[3] *Id*. at 10.

[4] *Id*. at 8, n. 7.

[5] *Andersen v. Stability AI Ltd.*, No. 3:23-cv-201, *complaint filed* (N.D. Cal. Jan. 13, 2023).

[6] *Getty Images (US), Inc. v. Stability AI Inc.*, No. 1:23-cv-135, *complaint filed* (D. Del. Feb. 3, 2023).

[7] *Silverman v. OpenAI Inc.*, No. 3:23-cv-3416, *complaint filed*, 2023 WL 4448007 (N.D. Cal. Jul. 7, 2023); *Kadrey v. Meta Platforms Inc.*, No. 23-cv-3417, *complaint filed*, 2023 WL 4463445 (N.D. Cal. Jul. 7, 2023).

[8] *Silverman v. OpenAI Inc.*, No. 3:23-cv-3416, *motion to dismiss filed* (N.D. Cal. Aug. 28, 2023).

[9] *Authors Guild et al. v. OpenAI Inc. et al.*, No. 1:23-cv-8292, *complaint filed* (S.D.N.Y. Sep. 19, 2023).

[10] *Chabon et al. v. OpenAI Inc. et al.*, No. 3:23-cv-4625, *complaint filed* (N.D. Cal. Sep. 8, 2023); *See also Chabon et al. v. Meta Platforms Inc.*, 4:23-cv-4663, *complaint filed* (N.D. Cal. Sep. 12, 2023).

[11] Motion to Dismiss at 2-3, Dkt. No. 32, *Silverman v. OpenAI Inc.*, No. 3:23-cv-3416 (N.D. Cal. Aug. 28, 2023).

[12] *Id*. at 8 (citing *Oracle*, 141 S. Ct. at 1199; *Connectix*, 203 F.3d at 603–08).

About the author

Bryan Mechell is a trial attorney in the intellectual property and technology group at **Robins Kaplan LLP** in Minneapolis, where he focuses his practice on litigating complex technology and software license disputes. He can be reached at BMechell@RobinsKaplan.com.

This article was first published on Westlaw Today on October 13, 2023.