Westlaw Today
powered by Reuters

Training AI/ML on copyrighted information

By Aaron M. Levine, Esq., Polsinelli

SEPTEMBER 21, 2023

Among the most interesting and controversial legal issues surrounding the use of Artificial Intelligence and Machine Learning ("AI/ML") is whether it is permissible to train AI/ML systems using copyrighted material.[1] AI/ML systems *require* training and generally the more training over high-quality data will result in better performance. This need to train AI/ML systems creates an incentive for AI/ML providers to use every source of relevant data they can access, which has the potential to bring them into conflict with copyright holders.

The Copyright Act grants copyright holders several express, exclusive rights, including the rights to: (1) make copies, (2) create derivative works, (3) distribute copies, (4) publicly perform, (5) display publicly, and (6) digitally perform.[2] One right not *expressly* granted is the right to permit or forbid the use of copyrighted works in AI/ML training. Therefore, unless that training process necessarily involves creating copies or derivative works, etc. it may not be copyright infringement to use copyrighted works in training.

Where questions of copyright infringement are in play, so are questions of "fair use."

Human creators of copyrighted works are understandably unhappy with the idea that their works could be used to train their competitors (or replacements). As discussed in more detail below, creators and owners of copyrighted works have initiated litigation against several AI/ML systems for training their systems on copyrighted material based on a mixture of copyright claims, copyright adjacent claims (such as the DMCA), unfair competition, breach of contract theories and even the right to publicity.

Today, concerns over AI/ML systems training on copyrighted material have been heightened further by OpenAI's announcement that, in preparation for the release of GPT-5, a new web crawling bot, GPTBot, will expand its scraping of the internet to create a massive trove of training data.[3] According to the Decrypt reporting, OpenAI will simply assume that you grant permission for them to scrape your website unless you add a disallow rule. Of course, legally speaking silence is consent only when one has a duty to speak.

In the wake of these developments, a number of potentially ground-breaking lawsuits that have been filed in the U.S. and in the UK, seeking to put an end to these practices before they become too widespread (or too late) to stop. Ultimately, however, it seems likely that Congress (or some other legislative body in the U.S. or elsewhere[4]) will need to address the question of appropriate use of training data, so that human creators can be compensated while permitting AI/ML training to progress.

Alternatively, a market-based solution could arise such as exists with ASCAP and BMG, large aggregators of rights that offer non-discriminatory licenses for copyrighted works for display or performance. A similar aggregation of rights to music, text and image data could be created such that AI/ML systems can be trained on quality information available on a non-discriminatory basis and human creators can be compensated.

Why training data is critical

A fair initial question is why any of this is happening? Before we get to the answer, it is important to provide at least a high-level understanding of how AI/ML systems work.

At the heart of an AI/ML system is an algorithm — a collection of complicated mathematical and logical procedures. Think of the AI/MI system's algorithm as a black box that has numerous dials and levers.[5] A training routine then can be thought of as a worker that goes into the black box and makes a series of tiny (but numerous) adjustments to the dials and levers. The goal of the trainer is to optimize these settings so that the "error," the difference between the black box's output from the desired output represented by the data set, is as small as possible.

With enough inputs/outputs and enough time, our imaginary trainer will search the space of dial and lever settings for their optimum settings. Once trained, the black box (AI/ML algorithm) will quickly provide pretty good outputs when given inputs it has never seen before. With higher quality and quantity of training data, pretty good can become very good. While garbage-in results in garbage-out, the reverse can also be true, in theory.[6]

Sources of training data

Given the crucial nature of training data, providers of AI/ML systems need to create their own training data or use existing data to train. The creation of training data can be done experimentally, or by relying on data created through customer interactions with a system.

Tesla's vehicles, for example, include as standard equipment the sensors that a fully autonomous vehicle needs. Tesla owners collect training data as they drive. This data is then used to develop

THOMSON REUTERS®

autonomous features, such as self-parking, adaptive cruise control and collision avoidance as Tesla works towards fully autonomous driving. The use of customer-created training data is typically governed by contract terms, although these can certainly become controversial.[7]

Likewise, licensing training data is also not problematic, as the owner or controller of that data can be compensated. As an example, Alphabet and Universal Music are in talks to license their artists' voices and melodies for AI-generated songs.[8]

Of course, some high-quality data sources are in the public domain or freely licensed. For example, both the U.S.[9] and UK[10] governments compile a vast array of data on nearly innumerable topics and the public is free to use this data for any lawful purpose. Private entities, such as Project Gutenberg can perform a similar function and can have very permissive licenses that may facilitate using their data for training purposes.[11]

However, other seemingly obvious sources of training data, such as Wikipedia, may not be available on the same "free" terms and conditions as their generally permissive licenses require that attributions and copyright notices be maintained — a license feature that forms the basis of several claims against AI/ML service providers.

Some very large databases of purported training data, such as LAION specifically warn users that "LAION datasets are simply indexes to the internet, *i.e.*, lists of URLs to the original images together with the ALT texts found linked to those images."[12]

Copyright, license and DMCA issues

As mentioned above, copyright owners hold several exclusive rights over their works — including the right to make copies and derivative works.[13] The classic example of creating derivative works would be to adapt a novel or play into a motion picture. Section 101 of the Copyright Act defines a derivative work:

> as a work based upon one or more preexisting works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, condensation, or any other form in which a work may be recast, transformed, or adapted. A work consisting of editorial revisions, annotations, or other modifications which, as a whole, represent an original work of authorship, is a 'derivative work'[14]

If a derivative work is created, the preexisting material contained within the derivative work becomes part of the copyrighted derivative work as a whole, but the copyright owner of the derivative-work copyright does not obtain superior or exclusive rights in the preexisting material.[15] Where questions of copyright infringement are in play, so are questions of "fair use," a commonly employed defense to copyright infringement allegations.

Fair use considerations include whether the second work adds something new to the original, whether the copyrighted work is expressive as opposed to factual, whether the entire work has been used or only a portion and finally, whether the market for the copyrighted original has been impacted by the alleged infringement.[16]

While most of these fair use factors are self-explanatory, transformative uses require some additional explanation. Transformative uses occur if the result and the original serve different market functions or if the transformative use offers something new or different from the original or expands its utility.

For example, using a thumbnail of a book cover in a search engine is a "transformative," even if commercial, where it functions in a completely different manner as the original work. That being said, commercial uses often result in findings against fair use. Whether output of a trained AI/ML system is "transformative" or not, is going to depend on the purpose of the AI/ML system's task.

Beyond copyright infringement, the Digital Millennium Copyright Act includes several copyright adjacent causes of action that copyright holders can make use of. For example, § 1202(b)(1), (2) and (3) makes it unlawful to tamper with "copyright management information" or "CMI" from copyrighted works.[17]

CMI includes information such as the copyright notice, the title and other identifying information. Since the outputs of AI/ML services do not present any copyright notices, it seems clear that the CMI was removed at some point in the process. This makes sense given the black box nature of the AI/ML systems themselves. The AI/ML system does not know — and it may not be possible for anyone to know — where a particular selection text, code or musical measure came from, unless it represents a true one-for-one copy.

Software development creates a particularly tricky problem. Open-Source Software (OSS) licenses typically grant licensees wide latitude to make copies, distribute and make derivative works. However, even the most permissive OSS licenses, such as the MIT and Apache 2.0 licenses require that the licensee maintain attributions, copyright notices and disclaimers of warranties.

For several OSS licenses, failing to do this is potentially the *only* realistic way one can violate the license. Of course, this is not the only way that a licensee can violate a license agreement by using licensed information in training.

Absent some clear guidance from lawmakers, many training data questions revolve around technically specific questions as to whether a particular AI/ML system's training process actually create copies or derivative works. Unfortunately for purposes of certainty, training routines *could* create copies, they do not *necessarily* create copies and the ultimate answer is likely to be technically dependent on the inner workings of a specific AI/ML system.

To return to the lever and dial analogy, your self-published book might be devoured by an AI/ML system and its content may be represented in the settings of the dials and levers, but it may be impossible to identify an actual "copy" of your book inside of the black box and it may be impossible to even identify which lever and dial settings represent the impact of your work on the AI/ML system. It is even possible that your particular work had no impact on the lever and dial settings at all.

Likewise, while generative AI/ML systems *could* create as output one-for-one copies or clearly derivative works, the output need not be so. By way of example, a seemingly obvious derivative work would be, AI "Frank Sinatra" covering "Where is My Mind" by the Pixies.[18] Less obvious derivative works would be prompting a

generative AI/ML system to create a picture of dogs playing poker in the style of Picasso. Picasso was famously a cat person and Cassius Coolidge was not a cubist.

AI/ML system providers argue that these are not truly derivative works and prefer the more positive sounding term of "influence." They point out that human creators work this way freely and often unconsciously. They have at least a point. I once took a painting class and unconsciously painted my dog Harry in a manner clearly influenced by the Blue Dogs of George Rodrigue.[19] Until a fellow student pointed out the similarities, I had not realized what I had done:

In the end, I might as well have named it "Sepia Dog," instead of Harry.[20] Yet, if it is permissible for a human to be influenced without infringing a copyright by creating a derivative work, should an AI/ML system not be allowed the same latitude? The New Orleans Museum of Art (NOMA) cannot keep me from training my own brain by staring at Blue Dog, in fact, arguably NOMA exists for this very purpose.[21]

On the other hand, a human could create music or art or literature absent their influences. The results would be different and maybe not as good, but something would still exist. "Sepia Dog" would have been very different if I had not been influenced by George Rodrigue, but I still could have painted Harry. Without training data created by human creators, it is not clear that generative AI could generate anything usable at all.[22]

The AI/ML lawsuits have begun

Artists, musicians, authors and even software developers are unsurprisingly unhappy with the idea of their creations being used to train the AI/ML systems that will compete or potentially replace them.

Potentially more ominously for the AI/ML system providers, large content aggregators such as the infamous Getty Images have also begun taking action.

- *J. Doe 1 et al. v. Github, Inc.*[23] While the chatbots and image generating AI/ML systems have received much of the press attention, the ability of AI/ML system to generate usable source code from text prompts is potentially magnitudes greater in economic impact than the ability to create videos of Joe Biden eating cement. The largest single public repository of source code is probably Github. Github has been beloved of OSS developers and a vast wealth of source code exists for free under OSS licenses of various flavors. CoPilot is a new tool that has been made available, for a fee, that will utilize the power of ChatGPT to create source code "suggestions" from text prompts. The training data for CoPilot came from the public repositories of GitHub. OSS licenses, as mentioned above, typically allow licensees to make copies and even derivative works free of charge. However, these very same licenses require licensees to carry forward the copyright notice, authorship, and warranty disclaimers. Moreover, these licenses "run with the code," so to speak and if one, for example incorporates code into a larger project that larger project becomes licensed under that same license. So-called "copy left" licenses impose other, more restrictive conditions, but even the most permissive popular OSS licenses require these minimum attributions. The Plaintiffs in the *Doe 1* lawsuit allege, among other things, that CoPilot's output strips all of this CMI and licensing information from the code it produces.
- *Andersen et al. v. Stability AI Ltd. et al.*[24] was filed as a class action against Stable AI Ltd., a UK company behind the "Stable Diffusion" tool that powers several image generating AI/ML systems, including Midjourney. The *Andersen* case alleges direct copyright infringement, vicarious copyright infringement, a DMCA claim for altering CMI, as well as several state law claims including violation of the right to publicity, unfair

competition and breach of contract (for violating terms of use). The *Andersen* case is an outlier to a certain extent from some of the other AI/ML system training cases in that in *Anderson*, the Plaintiffs allege direct copying as a potential violation.

- *Kadrey et al. v. Meta Platforms, Inc.*[25] was filed against Meta's AI offering LLaMA and includes the comedienne Sarah Silverman among its plaintiffs. Silverman and her fellow Plaintiffs allege that Meta used, among other sources, a training set that included the contents of "shadow library" websites, which include the Plaintiff's books. The counts recited in the complaint include direct copyright infringement, vicarious copyright infringement, DMCA allegations (again for removing CMI) and unfair competition. The *Meta* complaint also includes allegations of negligence and unjust enrichment.
- *Getty Images, v. Stability AI Ltd.*[26] Getty Images, the notoriously litigious aggregator of image copyrighted works has filed lawsuits against Stability AI, Ltd. in the District of Delaware and in the United Kingdom. The United Kingdom filing has not (as of writing) been made public. However, the Delaware complaint will likely share many similarities with the UK filing. Specifically, Getty Images alleges copyright infringement, trademark infringement, trademark dilution, a DMCA claim based on false CMI, a DMCA claim based on the removal of CMI, and two state law unfair competition claims. Of all the cases, the *Getty Images* cases are the most serious. Getty was able to provide in its complaint numerous examples of Stable Diffusion outputs with noticeable, if deformed, Getty Images watermarks, such as this particularly horrifying sample:[27]

It is difficult to argue against this being a derivative work of an authentic Getty Images offering. Normally, one would expect a settlement where such evidence exists, particularly where it involves an experienced litigant, such as Getty Images.

Artists, musicians, authors and even software developers are unsurprisingly unhappy with the idea of their creations being used to train the AI/ML systems.

In this instance, however, if generative AI is not brought in line, from Getty Images' point of view, Getty Images has little reason to exist going forward. Indeed, my personal observation is that many banners, thumbnails and other locations where one would have seen a stock image (likely previously sourced from Getty itself), one now sees clearly AI generated images in its place.

Conclusion

Some reasonable system is needed that balances the interests of creators and copyright holders with the need of AI systems to train on high quality data sets. Until these lawsuits percolate through the court system of the U.S. or until Congress creates a global solution thereby creating legal certainty, training AI systems is going to be a contentious issue and potentially a dangerous issue from a liability point of view. Creators and providers of AI systems will have to exercise care to use truly public domain data sources and ensure that their training data is aggregated and collected legitimately.

Notes

[1] The AI/ML systems referred to herein are primarily generative AI/ML tools also known as large language models and/or "chatbots."

[2] *See* 17 U.S.C.A. § 106.

[3] *See* https://bit.ly/3PCgbUl.

[4] "In the 2023 legislative session, at least 25 states, Puerto Rico and the District of Columbia introduced artificial intelligence bills, and 14 states and Puerto Rico adopted resolutions or enacted legislation." (Source: https://bit.ly/3r5ql6z).

[5] Mathematically, these levers and dials are represented by large matrix multiplications with a leavening of non-linear functions.

[6] Importantly, these kinds of systems are very different from scientific models that use known physical laws to make predictions or conduct simulations. Modern AI/ML systems are not necessarily built on any actual understanding or underlying phenomena. Expert Systems are a type of AI/ML system that attempts to directly emulate the decision-making trees of human experts and rather than acting as black boxes are typically designed as hyper-complex networks of in-than-else rules.

[7] *See e.g.*, https://bit.ly/45Px1EU.

[8] *See* https://bit.ly/3ZiX0SP.

[9] https://bit.ly/3Pfuz3y

[10] https://bit.ly/45UDYED

[11] *See* https://bit.ly/3Rlac7p

[12] *See* https://bit.ly/46etEY1. In other words, the LAION data sets are not "clean" from a copyright perspective.

[13] *See* 17 U.S.C.A. § 106.

[14] *See* 17 U.S.C.A. § 101.

[15] Under § 106(2), the owner of the original has the "exclusive rights to do and to authorize ... [the] prepar[ation] of derivative works based upon the copyrighted work." This begs the question as to what happens if that authorization is revoked. Section 203(b) addresses this, and in essence, it means that once permission to make derivative works is revoked, no further derivative works can be created; however, it does not hinder the copyright owner's exploitation of an authorized derivative work. The movie The Shining was based on a Stephen King novel, that Stanley Kubrick adapted, and Stephen King allegedly did not like. As much as King might like to have the movie version disappeared, it was at the time it was created, an authorized derivative work and Warner Bros. was free to release it, and sell VHS, DVD's, BluRay's and stream it on Amazon Prime.

[16] *See e.g.*, 17 U.S.C.A. § 107.

[17] *See* 17 U.S.C.A. § 1202(b).

[18] *See* https://bit.ly/44WBi81.

[19] *See* https://bit.ly/3PAEP7Q

[20] I will be sticking with my day job.

[21] *See* https://bit.ly/3rbFl2I

[22] As many suspect, the proportion of AI generated content versus human content will skyrocket in the next few years. See https://bit.ly/4695UUX. This will inevitably result in AI/ML systems being trained on the content of AI/ML systems, absent improvements in AI content detection. Particularly, because the Copyright Office has ruled that AI/ML generated works are not entitled to robust copyright protection. *See* https://bit.ly/44P0FJ1. In turn, this may lead to ultimately to cultural stagnation as everything is a "diffused" copy of a copy of a copy of a copy in the style of Harry Styles.

[23] Case No. 4:22-CV-6823 (N.D. Cal.).

[24] Case No. 3:23-CV-201 (N.D. Cal.).

[25] Case No. 3:23-CV-3417 (N.D. Cal.).

[26] Case No. 1:23-CV-135 (D. Del.).

[27] *See Getty Images (US), Inc. v. Stability AI, Inc.*, complaint at ¶ 58, *retrieved from* https://tmsnrt.rs/3LmOOen.

About the author

Aaron M. Levine is a shareholder in **Polsinelli**'s intellectual property practice and specializes in assisting clients with protecting and monetizing their IP assets. He has significant experience crafting and negotiating IP agreements, including AI/ML service and training data licenses, software development agreements, privacy policies, terms of service and transfers. He is based in Houston and can be reached at alevine@polsinelli.com.

This article was first published on Westlaw Today on September 21, 2023.