The Computer & Internet Lawyer

Volume 41 ▲ Number 1 ▲ January 2024

Ronald L. Johnston, Arnold & Porter, Editor-in-Chief

Guarding the Digital Fortress: How to Protect Your Data from Generative AI "Scrapers"

By Jennifer A. Kenedy and Jorden Rutledge

Generative AI tools create art, music, code and other new content by learning from others' work. Open AI used three hundred billion words, or five hundred seventy gigabytes, to build its GPT-3 model, upon which the ubiquitous ChatGPT is based. This "scraping" occurred unbeknownst to many (read: all) of the various creators and owners of these 570 gigabytes. These stakeholders are understandably concerned with unauthorized use of their works and the damage such use is causing. Creators of original works should be entitled to the fruits of their labor, and there is something intrinsically inequitable that, for example, a pianist can spend tens of thousands of hours perfecting her form, timing, and style and a computer program can, after ingesting the artist's works, reproduce works in this "style" within ten seconds.[1] But inequitable does not necessarily mean unlawful, and, as with the introduction of any new technology, the legal landscape is murky.

The authors, attorneys with Locke Lord LLP, may be contacted at jkenedy@lockelord.com and jorden.rutledge@lockelord.com, respectively.

Regardless, the current legal regime does provide avenues to safeguard one's protected work from being scraped, used, and replicated to train generative AI. This article addresses a few such ways.

USING DATA

As a brief explainer, ChatGPT and similar generative AI models operate by ingesting and analyzing vast amounts of data. This data is collected by programs that scrape all the text, images, and content they come across, largely from the Internet. Generative AI models use these vast data sets to help analyze patterns, understand language, and ultimately generate new and expressive content. While companies use these data sets to train their models, the outputs created by generative AI are, generally, unique, i.e., the AI will borrow from, but not generate exact copies of, the content the models were trained on.

Owners of this "scraped" content have a few legal avenues they can use to prevent, or potentially profit from, this "scraping" of their protected data to train generative AI tools.

If the content owner also owns the website where the scraping occurs, the most straight-forward approach

is to include a licensing agreement or terms of service to which visitors of the site must agree. Many websites use these tools to enumerate permitted uses of the site's content and impose restrictions. These intellectual property licenses may explicitly prohibit "scraping" or require a fee or specific authorization, if scraping of protected content is desired.

TWO CASES

Two current cases are drawing the battle lines and deciding the winners. In *Getty Images (US), Inc. v. Stability AI, Inc.*,[2] Getty Images sued Stability AI, the company behind the Stable Diffusion AI image generator. Getty alleges that over 12 million of its copyrighted images, along with their descriptions and metadata, were scraped and then used to train Stable Diffusion. The lawsuit seeks an eye-popping $1.8 trillion.

In *Doe v. GitHub, Inc.*,[3] plaintiffs allege defendants trained Codex and Copilot – two "assistive AI-based systems" – by exposing them to large quantities of data gathered from open source code repositories on GitHub, and then used Copilot and Codex to distribute similar code to users.

On May 11, 2023 the *GitHub* court held that the plaintiff stated a claim for breach of a licensing agreement when the plaintiff's content was allegedly scraped by the defendants to train generative AI tools. There, the site's licensing agreement prohibited copies or derivative works unless the new content "include attribution, a copyright notice, and the license terms." The court held that scraping and then redistributing the code constituted a prima facie breach of this license.[4]

LICENSING AGREEMENTS

With an increased understanding of this technology, lawyers should advise clients to consider modifying or creating terms of service or licensing agreements that adapt to this changing landscape.

A licensing approach in a similar vein has been operating for years in the streaming music space. Under the Musical Works Modernization Act, streaming services must pay a fixed licensing fee for works registered under the Act.[5] Artists are incentivized to join this collective because it vastly simplifies the royalty process and increases the artists' reach. One could imagine the central stakeholders – those who own IP that is most susceptible to reproduction and copying – forming a collective and demanding fees for the use of their work. This could certainly apply to the scraping of data to train generative AI tools, but a similar licensing approach could be envisioned for outputs of generative AI models, e.g., content owners demanding a fee every time a user requests an output in "the style of X."

COPYRIGHT CLAIMS

Another avenue open to content owners, regardless of whether they also own the website, is asserting copyright claims against the scrapers. Content owners who have registered their works and are eligible for statutory damages under the Copyright Act will have the greatest incentive to bring an action. It is easy to envision a sharp decrease in demand for an artist's/creator's authentic work when anyone can request a painting/song/book/poem in a particular person's style and receive a near replica.[6] And this shuddering of demand and lost sales would be one of the more straight-forward examples of damages in this scenario.

Furthermore, a party could assert a copyright claim for direct or secondary infringement against a company that is responsible for the scraping, as even if a party only saves the content for use in training, they could still have violated an owner's copyright.[7] However, courts have not addressed many of the central questions in this area, including whether a fair use defense is applicable; whether the sheer amount of training data makes one specific (and miniscule) subset de minimis; and the transformative nature of AI learning and outputs in the copyright context.

Finally, it should be noted that there are some[8] technological solutions that can prevent scraping. However, there will undoubtedly be new workarounds to get past new technological solutions, which will require advanced and evolving protections to stave off rapidly advancing generative AI technology.

With AI partnerships and products involving two of the main search engines Google (Bard) and Bing (OpenAI and Microsoft), these "scraping bot blocking" solutions also run the risk of blocking not just the scraping bot but also preventing the website from being indexed in search results on these main search engines. While creators of original works should strongly consider technological "anti-scraping" solutions, that is not a substitute for understanding and asserting their legal rights.

CONCLUSION

Generative AI and the legal implications of its use are both in their infancy. Courts can – and should – adjudicate cases while considering the primary stakeholders and the fundamental purposes of intellectual property rights.

Similarly, stakeholders should consider novel ways to protect their property rights from users of this new technology.

Notes

1. Currently, ChatGPT and other generative tools have an impressive ability to mimic individuals who are in the public sphere. This ability will only increase over time.
2. Getty Images (US), Inc. v. Stability AI, Inc., 23-cv-00135 (D. Del.).
3. Doe v. GitHub, Inc., 22-cv-06823 (N.D. Cal.).
4. The court stated: "Plaintiffs advance claims for breach of the eleven suggested licenses GitHub presents to users that require (1) attribution to the owner, (2) inclusion of a copyright notice, and (3) inclusion of the license terms. Compl. ¶ 34 n.4. Plaintiffs attach each of these licenses to the complaint. Plaintiffs allege that use of licensed code "is allowed only pursuant to the terms of the applicable Suggested License," and that each such license requires that any derivative work or copy include attribution, a copyright notice, and the license terms. Id. ¶¶ 173, 34 n.4. Plaintiffs further allege that Codex and Copilot reproduce licensed code as output without attribution, copyright notice, or license terms, thereby violating the relevant provisions of each license. While Plaintiffs do not identify the specific subsections of each suggested license that correspond to each of these requirements, the Court finds that Plaintiffs have sufficiently identified "the contractual obligations allegedly breached," as required to plead a breach of contract claim. Williams, 449 F. Supp. 3d at 908.
5. 17 U.S.C. § 101, et seq.
6. Indeed, the Authors Guild is clearly concerned, as the influential organization recently announced guidance for publishing agreements that would prevent authors' works from being used in AI training, absent the author's express permission; see https://authorsguild.org/news/model-clause-prohibiting-ai-training/.
7. See 2 Nimmer on Copyright § 8.02; Capitol Records, LLC v. ReDigi Inc. 934 F. Supp. 2d 640, 656-60.
8. https://www.cequence.ai/blog/bot-management/the-danger-of-web-scraping-and-how-to-prevent-it/.