Equitable Legal Remedies and the Existential Threat to Generative AI

Elizabeth E. Joh

Generative AI programs like ChatGPT and Midjourney can produce text, images, and video because their models have been trained on vast amounts of books, pictures, art, and other media found on the internet. Training a large language model like ChatGPT, for instance, to “learn” patterns and relationships in billions of words allows these models to produce written text responsive to prompts. But these models have trained on digital information like novels and original art without obtaining permission. Unsurprisingly, the release of these generative AI tools has already prompted lawsuits by individuals and companies alleging that the use of their artistic works in the development of these models violates copyright law. Whether reliance on this type of training data violates copyright, and whether the outputs based on the training data constitute infringing derivative works are open and difficult questions. But these copyright claims also raise a separate issue: can plaintiffs seek remedies that would destroy generative AI programs?

Federal copyright law allows plaintiffs to seek actual damages or statutory damages for the infringing works. The two lawsuits filed in July 2023 by the comedian Sarah Silverman, for instance, seek such damages against OpenAI and Meta. The suits allege that both models illegally trained on copyrighted works, including her own book Bedwetter, and when prompted, can summarize and reproduce these works.

But copyright law also permits litigants to seek not just damages but equitable relief. That potential relief is very broad. A court may “order the destruction” of all infringing works, but also of “all plates, molds, matrices, masters, tapes, film negatives, or other articles by means of which such copies or phonorecords may be reproduced.” This means that a court may order the defendant to destroy not just the copyright-infringing works, but also tools that have the capacity to produce a copyright-infringing work. This broad language grants federal courts to use their discretion in deciding whether to order the destruction

---

1 Professor of Law, U.C. Davis School of Law
2 17 U.S.C. 504(b)-(c).
4 17 U.S.C. 503(b).
of tools capable of infringement but also non-infringing uses, including printing presses and other “dual use” tools.\textsuperscript{5}

The availability of this equitable remedy in copyright lawsuits means that courts can order the destruction of a broad range of tools like machines, computers, and other vehicles for infringement. Some federal courts do not even require a connection between the tools ordered to be seized or destroyed and any infringing action.\textsuperscript{6}

What does the destruction clause mean in the generative AI context? If we assume for now that there are legitimate copyright issues in the first generation of lawsuits, then programs like Stability AI or ChatGPT can plausibly be construed as tools capable of producing copyright infringing works under the terms of 503(b)’s destruction clause. These tools rely on copyrighted images, text, and video in order to learn to generate new synthetic data. The resulting outputs can very closely resemble the works in training data sets used by these models. If a large language model can mimic Sarah Silverman’s \textit{Bedwetter}, it is because her work was included in the model’s training data. Researchers have also demonstrated that copyrighted images can be extracted from these models through careful prompting, because some images have been “memorized” by the model.\textsuperscript{7} A court may determine that a generative AI program not only should be retrained because of copyright infringement, but also should be scrapped altogether under the equitable remedy of destruction.

The problem is that models like Chat GPT-3 represent massive amounts of research, investment, and data. A large language model is not a printing press or a hard drive. But the statutory discretion granted to federal courts in these copyright lawsuits leaves a great deal of uncertainty about the scope and nature of a destruction remedy. The question is an existential one for generative AI.

\textsuperscript{5} Anne-Marie Carstens, Copyright’s Deprivations, 96 Wash. L. Rev. 1275, 1279 (2021).
\textsuperscript{7} Nicholas Carlini, et al., Extracting Training Data from Diffusion Models (Jan. 2023), at https://arxiv.org/abs/2301.13188.