Regulating Predatory Style Transfer

By Matthew Sag, Professor of Law in Artificial Intelligence, Machine Learning, and Data Science
Emory University Law School

Abstract

The emerging international consensus that text data mining does not, or should not generally, amount to copyright infringement was predicated on a technological paradigm that now seems quaint. In the United States, the key decisions confirming that the non-expressive use of copyrighted works is fair use, and thus non-infringing, involve software reverse engineering through decompilation, plagiarism detection software, library digitization to enable meta-analysis of millions of books for academic research, and library digitization to facilitate search and other forms of data extraction. The rationale for allowing for-profit and academic researchers to derive valuable data from other people's copyrighted works that emerges from these cases is a necessary implication of the fundamental distinction between ideas and expression. The process of text data mining renders copyrighted text, sounds, and images into uncopyrightable abstractions. These abstractions are not same, or even substantially similar to, the original expression, but they are interesting and useful in combination for generating insights about the original expression. Accordingly, theorists have argued and courts have ruled that technical acts of copying that do not communicate the original expression to a new audience do not interfere with the interest in original expression that copyright is designed to protect.

Recent advances in generative artificial intelligence have challenged this premise. By training unfathomably large machine learning models on a significant fraction of all of the digitized expression in the world—and using as much electricity as it would take to power a small country—"foundation models" such as GPT-3, BERT, and DALL-E 2 can produce much more than information about expression, they are now the engines of new content creation. The remarkable capacity of ChatGPT to synthesize, speculate, elaborate, tell stories, and otherwise mimic human language is now well documented. The ability of image and music generation models to create digital art is just as impressive. For the most part, the copyright implications of the new wave of foundation models are no different to earlier applications of text data mining. Most of the time, when a user enters a prompt into ChatGPT or Midjourney the output carries no resemblance to any particular input or set of inputs, except at the abstract and unprotectable level. Even though these applications produce new content that is often indistinguishable from copyrightable human expression, they still qualify as non-expressive uses because these outputs are not substantially similar to any particular original expression in the training data. Mostly.

Some caution is warranted before we breezily assume that the reasoning of court decisions holding that copying student term papers to check for plagiarism is fair use applies to computer systems that copy millions of works of art to produce yet more works of art. One reason for caution is that because of the staggering size of foundation models, we can no longer take it for

granted that that a model trained on a set of copyrighted inputs won't simply encode and reproduce those inputs. The latest research confirms that systems like Stable Diffusion do sometimes reproduce nearly identical copies of images from the training data, but only under special conditions, and exceedingly rarely. The second reason is that even where the model produces content that is not substantially similar to any specific individual work in the training data, it may fabricate digital artifacts that bear an uncomfortably close resemblance to a particular artist's "style." Generative AI that allows users to create new content that falls short of copyright infringement based on a traditional application of the idea-expression distinction and judicial elaborations of the threshold of substantial similarity may nonetheless undermine the dignitary and commercial interests of artists whose works are well-represented in the training data. The question for copyright law and the creative industries is whether and how such "predatory style transfer" should be regulated.