Generative AI could leave users holding the bag for copyright violations

March 23 2024, by Anjana Susarla

Credit: AI-generated image

Generative artificial intelligence has been hailed for its potential to transform creativity, and especially by lowering the barriers to content

creation. While the creative potential of generative AI tools has often been highlighted, the popularity of these tools poses questions about intellectual property and copyright protection.

Generative AI tools such as ChatGPT are powered by foundational AI models, or AI models trained on vast quantities of data. Generative AI is trained on billions of pieces of data taken from text or images scraped from the internet.

Generative AI uses very powerful machine learning methods such as deep learning and transfer learning on such vast repositories of data to understand the relationships among those pieces of data—for instance, which words tend to follow other words. This allows generative AI to perform a broad range of tasks that can mimic cognition and reasoning.

One problem is that output from an AI tool can be very similar to copyright-protected materials. Leaving aside how generative models are trained, the challenge that widespread use of generative AI poses is how individuals and companies could be held liable when generative AI outputs infringe on copyright protections.

When prompts result in copyright violations

Researchers and journalists have raised the possibility that through selective prompting strategies, people can end up creating text, images or video that violates copyright law. Typically, generative AI tools output an image, text or video but do not provide any warning about potential infringement. This raises the question of how to ensure that users of generative AI tools do not unknowingly end up infringing copyright protection.

The legal argument advanced by generative AI companies is that AI trained on copyrighted works is not an infringement of copyright since these models are not copying the training data; rather, they are designed to learn the associations between the elements of writings and images like words and pixels. AI companies, including Stability AI, maker of image generator Stable Diffusion, contend that output images provided in response to a particular text prompt is not likely to be a close match for any specific image in the training data.

Builders of generative AI tools have argued that prompts do not reproduce the training data, which should protect them from claims of copyright violation. Some audit studies have shown, though, that end users of generative AI can issue prompts that result in copyright violations by producing works that closely resemble copyright-protected content.

Establishing infringement requires detecting a close resemblance between expressive elements of a stylistically similar work and original expression in particular works by that artist. Researchers have shown that methods such as training data extraction attacks, which involve selective prompting strategies, and extractable memorization, which tricks generative AI systems into revealing training data, can recover individual training examples ranging from photographs of individuals to trademarked company logos.

Audit studies such as the one conducted by computer scientist Gary Marcus and artist Reid Southern provide several examples where there can be little ambiguity about the degree to which visual generative AI models produce images that infringe on copyright protection. The New York Times provided a similar comparison of images showing how generative AI tools can violate copyright protection.

How to build guardrails

Legal scholars have dubbed the challenge in developing guardrails against copyright infringement into AI tools the "Snoopy problem." The more a copyrighted work is protecting a likeness—for example, the cartoon character Snoopy—the more likely it is a generative AI tool will copy it compared to copying a specific image.

Researchers in computer vision have long grappled with the issue of how to detect copyright infringement, such as logos that are counterfeited or images that are protected by patents. Researchers have also examined how logo detection can help identify counterfeit products. These methods can be helpful in detecting violations of copyright. Methods to establish content provenance and authenticity could be helpful as well.

With respect to model training, AI researchers have suggested methods for making generative AI models unlearn copyrighted data. Some AI companies such as Anthropic have announced pledges to not use data produced by their customers to train advanced models such as Anthropic's large language model Claude. Methods for AI safety such as red teaming—attempts to force AI tools to misbehave—or ensuring that the model training process reduces the similarity between the outputs of generative AI and copyrighted material may help as well.

Role for regulation

Human creators know to decline requests to produce content that violates copyright. Can AI companies build similar guardrails into generative AI?

There's no established approaches to build such guardrails into generative AI, nor are there any public tools or databases that users can consult to establish copyright infringement. Even if tools like these were available, they could put an excessive burden on both users and content

providers.

Given that naive users can't be expected to learn and follow best practices to avoid infringing copyrighted material, there are roles for policymakers and regulation. It may take a combination of legal and regulatory guidelines to ensure best practices for copyright safety.

For example, companies that build generative AI models could use filtering or restrict model outputs to limit copyright infringement. Similarly, regulatory intervention may be necessary to ensure that builders of generative AI models build datasets and train models in ways that reduce the risk that the output of their products infringe creators' copyrights.

Provided by The Conversation