Insight

Primer: Training AI Models with Copyrighted Work

JOSHUA LEVINE, JOHN BELTON | NOVEMBER 14, 2023

Executive Summary

- Multiple corporate and class action lawsuits claim that artificial intelligence (AI) training practices are violating copyright law; specifically, copyright holders claim that training AI models on copyrighted material is not a "fair use" of their material and should be prohibited.
- No major judgements have been made to date, yet if courts agree with the plaintiffs, it could jeopardize future development of AI technology.
- As Congress will likely consider legislative action to provide clarity to courts and stakeholders, this primer discusses how AI models are trained, the intersection between existing copyright law and training AI models, and how potential judicial and global decisions could influence the United States' development and use of AI.

Introduction

As the use of artificial intelligence (AI) grows, policymakers are grappling with critical issues surrounding the data used to train AI models and intellectual property (IP) rights. Today's leading generative AI models are trained on mass amounts of data – today's largest datasets contain over a billion pieces of human-generated text – and often utilize copyrighted materials as they are the "best examples of high-quality" content. Concerned about potential copyright infringement, however, content producers have filed several lawsuits focused on the IP implications of training generative AI models on copyrighted work. While no major judgments have been made to date, if courts do not find that this a "fair use" of the material, it could hamper the development of AI in the United States.

Congress and the Biden Administration have begun to explore potential policy solutions to balance the value of copyrights with the need for more data to train AI. As policymakers craft a response, they should examine how different countries have approached the issues, as well as potential market-driven solutions. For example, the European Union, Israel, and Japan have rolled out frameworks to increase transparency around the sorts of data that AI uses, and even hold that training on AI is a fair use of the copyrighted material. Similarly, companies such as Adobe have begun introducing a feature allowing creators to opt out of their data being used for AI development, which could obviate the need for major regulatory reforms.

This primer discusses how AI models are trained, how existing copyright law intersects with training AI models, and the broader legislative and regulatory environment surrounding AI. It also examines the impact of judicial and key global decisions influencing the United States' development and use of AI.

Training on Copyrighted Material and Legal Implications

The principal input for AI models is training data, classified as the raw information a model uses to make

decisions. Training data presents itself in many forms. Self-driving cars require photos and videos, allowing a model to interpret road signs and distinguish them from billboards. A customer service chatbot requires authentic voice and chat interactions, showing the robot how to assist a customer properly.

While the quantity of training data is essential – frontier models often utilize over 45 terabytes (TB) of data – the quality is equally important and thus developers use data sets containing copyright-protected material. For example, text generation models such as ChatGPT need materials such as books, which are invaluable because of their length and diversity of content. Open-source datasets today contain almost all literature ever written, providing a plug-and-play solution for AI training. This practice is not limited to just literature. AI models are trained on nearly all forms of copyrighted material, including videos, images, and music.

This approach to training is not without significant controversy. Content creators have already begun to file lawsuits over the potential infringement of their copyrighted material. In July, writer and comedian Sarah Silverman sued OpenAI and Meta for direct and robust copyright infringement, stating the model is using protected work without her permission. In a similar suit filed in late September, several authors allege there is a "systemic theft" in the training of AI models by OpenAI. These suits rely on a similar set of facts: The Large Language Model ChatGPT was able to reproduce and provide information regarding plaintiffs' copyrighted materials. The copyright holders did not authorize this training, and they claim this violates the Copyright Act.

The outcome of these cases will likely depend on the courts' application of the doctrine of "fair use," a major defense for AI developers. Fair use essentially allows for infringement of a copyright when the infringement of that copyright is done for a limited and transformative purpose. To determine whether a fair use defense would be applicable, courts balance four criteria:

1. The purpose and character of the use, including whether it is for commercial or educational purposes.
2. The nature of the copyrighted work.
3. The amount and substantiality of the portion used concerning the copyrighted work.
4. The effect of the use on the potential market for or value of the copyrighted work.

As there have been no rulings on the applicability of fair use for training AI models, both AI developers and copyright holders face uncertainty regarding their perspective industries. Experts and creatives have cited AI's potential to replace artists and transform methods of work as destabilizing for markets of creative products. Conversely, AI also has great potential to support artists' development and accelerate creative industries. A letter published by Creative Comments and signed by artists states, "Just like previous innovations, these tools lower barriers in creating art…" While courts will undoubtedly grapple with these factors in the context of AI training data, it may be prudent for Congress to specifically address AI and copyright issues to give courts and relevant stakeholders clarity.

Looking Forward

As Congress works with relevant agencies and considers legislative solutions, lawmakers could look to their international counterparts as well as private industry to inform a path forward.

The U.S. Copyright Office launched an initiative examining "the copyright law and policy issues raised by artificial intelligence technology," and on August 30, the office issued a formal Notice of Inquiry requesting information to "advise Congress" on potential paths forward. In Congress, the Senate Judiciary Committee has held two hearingsfocused on AI and intellectual property, one entirely focused on copyright. The Federal Trade

Commission is also getting involved, holding a roundtable featuring artists and creatives to discuss the various implications AI could have on their industries. A bipartisan group of senators has also released a discussion draft of a bill focused on digital replication of work. In a recent Senate Judiciary Subcommittee hearing on IP, witness testimony illustrated the negative impact AI can have on human creators. Congress may look to address the issue and has already held two hearings specifically focusing on AI and copyright. Ranking Member Tillis noted that "action is clearly required" and emphasized aligning U.S. policy with that of similar countries.

While no formal action has been taken in the United States, regulatory regimes around the world and market-driven self-governance tools could inform regulators about how to balance the need for AI innovation and copyright protections.

International counterparts have taken differing approaches that could help inform lawmakers. In the first comprehensive piece of legislation regarding AI, the European Union's AI Act requires foundation model developers "to publicly disclose a 'sufficiently detailed summary' of the copyrighted material used as training data." Increasing transparency makes it possible for creators to know if their data was trained on, meaning artists and copyright holders could have more agency over and seek compensation for the use of their work. Problems arise with this approach because models are trained on broad datasets, making it challenging to quantify how much an individual creator's work contributed to an output and the corresponding compensation. Some in Congress are considering similar action, evidenced by the Bipartisan Framework on Artificial Intelligence Legislation – developed by Senators Hawley and Blumenthal – which states "Developers should be required to disclose essential information about the training data, limitations, accuracy, and safety of AI models to users and other companies."

Diverging from the EU approach, Japan and Israel are allowing machine learning enterprises to make unauthorized use of copyrighted materials to train AI systems, generally favoring the development of AI over the rights of the copyright holders. The Japanese government has differentiated between training models on copyrighted materials and the output generated by a trained model, specifically expanding fair use to train models focused on information analysis or applications related to sound or video recording. A recent opinion by Israel's Ministry of Justice contrasts with the Japanese position by expanding fair use to cover training AI models broadly, but specifies instances that are restricted, such as a model exclusively trained on a single artist's work. Japan and Israel's policiesare intended to accelerate the implementation and rate of AI innovation in the two countries.

Finally, market forces could obviate the need for major regulatory changes. For example, private firms are working to create tools that would give creators a way to prevent AI companies from training models on their work. Instead of disclosures, Adobe announced a policy in which creators can tag their content as "do not train" within the content's metadata, notifying developers or data scrapers they do not want their work to be included in a training data set. Implementing this throughout industries and products that AI developers use to train could allow creators to identify and protect their content that they don't want used by AI, while still allowing development to take place. Other industry participants and researchers are developing training blockers, for example, OpenAI and a web3 protocol group have both developed tools that prevent an individuals' data from being scraped, and researchers are experimenting with "data poisoning attacks" which can damage future iterations of image-generating AI models if the poisoned image is included in the training set.

Conclusion

Copyrighted material is critical for the development of AI models, but policymakers must also weigh the interests of copyright holders. If Congress pursues legislation clarifying copyright protections for material used to train AI models, it would be well served to consider how other nations are balancing this trade-off, as well as

technical solutions developed by the private sector and civil society that can empower creators to decide if and how their work is used to train generative models.