Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
November 2023

Foundation Models and Copyright Questions

Peter Henderson, Xuechen Li, Dan Jurafsky,
Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang

FOUNDATION MODELS ARE OFTEN TRAINED ON LARGE VOLUMES OF COPYRIGHTED MATERIAL, including text on websites, images posted online, research papers, books, articles, and more. Deploying these models can pose legal and ethical risks. Under U.S. law, copyright for a piece of creative work is assigned "the moment it is created and fixed in a tangible form that it is perceptible either directly or with the aid of a machine or device." Most data used to train foundation models falls under this definition. For example, the Pile, a massive open source language modeling dataset that has been used by Meta, Bloomberg, and others to train foundation models, contains a dataset of copyrighted, torrented e-books called Books3 that has become the focus of various ongoing lawsuits.

In the United States, AI researchers have long relied on fair use doctrine to avoid copyright issues with training data. The fair use doctrine allows members of the public to use copyrighted materials in certain instances, notably when the output is "transformative." However, amid a class-action lawsuit against Microsoft, GitHub, and OpenAI for training systems on publicly published code without adequate credit; Getty Images suing the Stable Diffusion AI tool for scraping its photos; and other significant AI-related legal actions, existing fair use interpretations are increasingly being challenged.

Key Takeaways

Foundation models—AI models trained on broad data at scale for a wide range of tasks—are often trained on large volumes of copyrighted material. Deploying these models can pose legal and ethical risks related to copyright.

Our review of U.S. fair use doctrine concludes that fair use is not guaranteed for foundation models as they can generate content that is not "transformative" enough compared to the copyrighted material. However, amid still evolving case law, the extent of copyright infringement risk and potency of a fair use defense remain uncertain.

To mitigate copyright risks, policymakers should consider making clarifications to fair use doctrine as it applies to AI training data while also encouraging good-faith technical mitigation strategies that align foundation models with fair use standards. Together, these strategies can maximize the benefits of foundation models while minimizing the moral, ethical, and legal harms of copyright violations.

In parallel, policymakers should investigate other policy mechanisms to ensure artists, authors, and creators are awarded fair compensation and credit, both those who do their work with the assistance of AI tools and those who do not use AI.

...the United States needs a two-pronged approach to addressing these copyright issues—a mix of legal and technical mitigations that would allow us to harness the positive impact while reducing intellectual property harms to creators.

In our paper "Foundation Models and Fair Use," we shed light on the urgency and uncertainty surrounding the copyright implications of foundation models. First, we reviewed relevant aspects of U.S. case law on fair use to identify the potential risks of foundation models developed using copyrighted content. We highlight that fair use is not guaranteed and that the risk of copyright infringement is real, though the exact extent remains uncertain. Second, we discussed four technical strategies to help reduce the risk of potential copyright violations, while underscoring the need for developing more techniques to ensure that foundation models behave in ways that are aligned with fair use.

We argue that the United States needs a two-pronged approach to addressing these copyright issues—a mix of legal and technical mitigations that would allow us to harness the positive impact while reducing intellectual property harms to creators. Fair use is not a panacea. Machine learning researchers, lawmakers, and other stakeholders need to understand both U.S. copyright law and technical mitigation measures that can help navigate the copyright questions of foundation models going forward.

An Analysis of U.S. Case Law

In the United States, the legal doctrine of fair use allows the unlicensed use of some copyrighted material. Whether "fair use" applies depends on four factors:

- Transformativeness: This factor is determined based on the *purpose* and *character* of the use. When the original work is deemed transformative, that weighs heavily in favor of fair use.
- Nature of copyrighted work: Fair use is strongly favored if the original work is factual as opposed to creative.
- Amount and substantiality: Taking smaller amounts and less substantial portions of the original work is more likely to be considered fair use.
- Effect on market: The impact the new product has on the market for the original work could affect fair use.

Each of these considerations comes into play when talking about foundation models, their training data, and potential copyright issues. They are also relevant at various stages of the model development process, during which numerous actors create, collect, and distribute training data and others train, host, and use the resulting model. In this paper, however, we focus on the copyright issues faced by those who create data (and therefore own intellectual property).

In legal analyses, the transformativeness factor tends to carry the greatest weight when determining fair use and is heavily emphasized in legal assessments. When Google copied part of the Java API for its Android operating system, for example, the U.S. Supreme Court ruled it was fair use, considering the relatively small percentage of code copied and the transformativeness of the end product, among other factors.

Fair use could apply to the training of AI systems, including some foundation models, when they function differently from the (copyrighted) input data, particularly when they focus on a different market. Training a model like a search engine or recommender system using books as input data, for example, would likely be sufficiently transformative from the books' original purpose and target audience to qualify as fair use. The new model does not substitute or compete with the books themselves. Rather, it enables new services in different markets.

However, it becomes more complicated when we consider generative use cases of foundation models. Generative foundation models like ChatGPT or DALL-E are likely to count as fair use when they transform input data into totally different, creative outputs. But that argument is diminished if the downstream product is too similar to the input data. For instance, if a language model produces text that is similar to the copyrighted book it was trained on, or targets a highly similar economic market, the court may find that the generated content, the model deployment, and even the model parameters themselves do not constitute fair use.

Researchers, policymakers, and others investigating copyright issues must consider other fair use factors as well, such as the amount of content taken from the original work. For example, in a 2015 case against Google, the court determined that Google Books was covered by fair use because it did not display significant portions of the books on its website. Compare that to a different court ruling in 2015, which held that displaying entire books online with small changes to formatting does *not* constitute fair use. When thinking about language models like the one behind ChatGPT, or large text-to-image models like DALL-E, the amount of data used and presented matters.

...what constitutes fair use is highly contextual and reliant on deeper semantic interpretations.

Our paper discussed many other scenarios with different copyright implications, including mimicry that does not include verbatim copying and parodies that copy some source material to provide commentary. The research also explores the fair use complexities specific to models that generate code or images, which come with their own distinctive case law with varied, often conflicting, outcomes. Ultimately, our analysis found that what constitutes fair use is highly contextual and reliant on deeper semantic interpretations. In particular, the evolving market for licensing training materials may affect a court's analysis of the fourth fair use factor.

Technical Mitigations

Scholars (including one of the authors) have suggested that humans and AI should be held to similar standards when it comes to copyright. The legal complexities of determining the fair use of foundation models highlight the importance for machine learning researchers and developers to design new technical strategies and tools that allow them to create models that meet fair use standards. In doing so, researchers should go beyond focusing on only near-verbatim text matching.

The legal complexities of determining the fair use of foundation models highlight the importance for machine learning researchers and developers to design new technical strategies and tools that allow them to create models that meet fair use standards.

We surveyed existing and potential tools and identified five main technical mitigation approaches that can help foundation models stay in line with fair use: 1) filtering training data and outputs; 2) filtering inputs and outputs at runtime; 3) scoring training data to understand its role in outputs, also known as instance attribution; 4) using differentially private training, which ensures that including or excluding individual data points does not greatly vary a model's parameters; and 5) making models learn from human feedback; in other words, training models to generate outputs aligned with human values. Each of these can work in tandem to help address the copyright questions of foundation models.

Training-time data filtering

The data filtering approach encompasses two main types of data filtering that could help with copyright issues. Researchers could filter training data for underlying licenses, copyright status, and opt-outs, choosing not to train models on copyrighted or restrictively licensed material at all; rather, the model would be trained only on open data. Another type of data filtering is deduplication: If an example occurs multiple times in training data, all but one of them are removed. The idea is that the more a foundation model sees a particular example, the more likely it may be to regurgitate that exact piece of data or information in its final outputs. Doing this kind of filtering could help avoid potential copyright violations by preventing memorization and regurgitation.

Runtime filtering

A different type of filtering would filter the inputs to the model and/or the outputs from the model at runtime. This could apply to the prompts that users supply to the model. For example, Google's MusicLM and OpenAI's DALL-E 3 both reject prompts that ask for generations mimicking a particular artist's style. Though style itself may not necessarily be copyrightable, such filtering mechanisms reduce the risk of outputting something that is potentially infringing and not fair use. Similarly, the outputs of the model could be filtered to prevent verbatim matching of copyrighted material.

Instance attribution

Instance attribution can help to identify the source of a particular foundation model output. Using instance attribution could help to clarify whether the output of a foundation model trained on large volumes of data was influenced by a copyrighted work. This could then be used as a data point to understand the copyright infringement risk associated with that model.

Differential privacy

Differential privacy is a mathematical guarantee used to protect information in training data. Specifically,

in machine learning, a differential privacy guarantee means that no adversary can distinguish, with high probability, between a model trained with a particular training example and one trained without it. Multiple studies have found that foundation models trained with strong levels of differential privacy memorize only limited amounts of training data. In those cases, extracting training data from or reconstructing a model's training data is almost impossible. This could be a strong technical measure to mitigate the risk of copyright infringement through verbatim memorization, but it may not address other more complex forms of infringement.

Learning from human feedback

Lastly, training models to generate outputs aligned with human values could include considering the copyright implications of rating systems. If a model is rated by how well it follows instructions, telling it "read me a Harry Potter book verbatim" would likely lead the model to read the entire book verbatim—and infringe on copyright in the process. Developers working on these model ratings could therefore include a consideration for how well a model respects fair use. Using content that is sufficiently transformative could be rated as a more successful output than using content that copies word for word something under copyright.

Policy Discussion

As the use of foundation models to generate content continues to skyrocket, so does the urgency with which content creators, researchers, companies, legal scholars, and policymakers must address related copyright questions. Even with fair use protections, copyright infringement and resulting litigation are a real risk.

As various stakeholders attempt to address the current uncertainties surrounding these copyright questions, conversations about fair use and machine learning could trend toward extremes. On the one hand, courts could rule that foundation models are widely acceptable under fair use, which would mean copyright owners would not be paid for this use. On the other hand, courts could declare that generative foundation models cannot use unlicensed, copyrighted training data. This could lead to a concentration of power among companies, such as Meta and Google, that have retained licenses to large amounts of data or that own large amounts of user-contributed data. The technical mitigation strategies discussed in this paper could help move the conversation toward a more productive middle ground.

However, policymakers should be wary of public or company policies that involve unreasonably high degrees of data filtering. Other countries' online filtering requirements have been criticized for their impacts on free speech. YouTube's content ID system has also faced criticism for its overaggressive filtering and for not following fair use standards. When addressing copyright issues, factual content, parodies, and short-form regurgitation used for commentary (e.g., quoting a book in a news article) should not necessarily be filtered out of foundation model training data.

Developers and researchers will not always be able to determine the provenance of every piece of data in a large training dataset. The law, including applications of the Digital Millennium Copyright Act (DMCA), may have to evolve to reflect this reality, and policymakers could, for example, consider establishing safe harbors where foundation models that employ sufficiently robust technical mitigations are protected from legal responsibility for copyright infringement claims. These

Developers and researchers will not always be able to determine the provenance of every piece of data in a large training dataset.

are worthwhile steps that could provide more balance and protection than the general exemptions in place in other countries.

Foundation models could also create additional harms well beyond copyright infringement, ranging from the disruption of creative industries to the exploitation of labor. Fair use doctrine and technical mitigations will not solve everything. Policymakers must survey and understand the landscape of available options and consider if and how statutory licensing schemes for data or taxation and redistribution policies for data could fit into the picture.

The law is opaque and continuously shifting around foundation models and their impact on copyrighted works. Machine learning researchers can help shape this conversation through education and by proactively pursuing effective technical mitigations. In tandem, policymakers can work on clarifying unresolved questions surrounding fair use and potential legal protections. As foundation models permeate society, it is critical that maximizing their benefits does not come at the expense of violating the intellectual property of those whose data is used to train the models.

Reference: The original article is accessible at Peter Henderson et al., "**Foundation Models and Fair Use**," arxiv.org, March 28, 2023, https://arxiv.org/abs/2303.15715.

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Peter Henderson is an incoming assistant professor of computer science and of public and international affairs at Princeton University. He received his JD from Stanford Law School and will receive his PhD in computer science from Stanford University.

Xuechen Li is a PhD student in computer science at Stanford University.

Dan Jurafsky is a professor of linguistics and computer science at Stanford University.

Tatsunori Hashimoto is an assistant professor of computer science at Stanford University.

Mark A. Lemley is the William H. Neukom Professor at Stanford Law School. He represents generative AI companies in some of the pending lawsuits discussed in this paper.

Percy Liang is an associate professor of computer science and the director of the Center for Research on Foundation Models at Stanford University.

HAI **Stanford University** Human-Centered Artificial Intelligence