COMMUNIA

Policy Paper #15 on using copyrighted works for teaching the machine

Background

We have witnessed a proliferation of (so-called) generative artificial intelligence (AI) models since OpenAI made DALL-E 2 available to the public in July 2022. This has gone alongside a renewed interest in questions about the relationship between machine learning (ML) and copyright law — as evidenced by a surge in publications on the topic both in scientific journals and general-audience media, as well as a number of lawsuits.

In this policy paper, we are looking at the input side of the equation within the EU copyright framework.[1] We discuss the considerations of the use of copyright-protected works and other protected subject matter as training data for generative AI models, and provide two recommendations for lawmakers. Here, we leave aside questions relating to the output of AI models (e.g. whether the output of generative AI models is copyrightable and in how far such output can be infringing exclusive rights), which we will address in another, yet to be published paper.

The surge of generative AI raises concerns for creators and their livelihood. It also prompts broader questions about the implications of potentially inauthentic and untrustworthy AI-generated output for social cohesion as well as about the extraction and concentration of resources by only a few tech companies. We are mindful of these challenges, but copyright is not designed to address all of them in a way that does justice to the underlying grievances.

Training generative machine learning systems

This paper is based on the assumption that in order to train generative ML models their developers require access to large amounts of materials, including images and text, many of which – but far from all – are protected by copyright. Currently, the most prominent examples of such models are image generators (such as Dalll-E, Midjourney and Stable

[1] This paper is without prejudice to the position of COMMUNIA or individual COMMUNIA members regarding this discussion in other jurisdictions.

Diffusion) and large language models (such as BLOOM, GPT and LLaMA) that are able to generate text and sometimes software code. But it is only a question of time until the same questions will arise for music or video generators and the use of copyrighted musical or audiovisual materials.

Access to large amounts of training data enables developers of ML models to train their models so that they can generate output. Based on our understanding of the technology, we assume that copies of the works that have been used for training are in no way stored as part of the model weights[2] or the model itself.

What we observe at a high level of abstraction is a situation where ML models are trained on very large numbers of works from a vast number of rightholders. Usually, the generated output will be based on the model as derived from the totality of the training data.

Traditionally, the copyright system has been badly equipped to deal with instances of large numbers of underlying rightholders. Rights clearance for mass digitization projects, for instance, is greatly encumbered by the amount of rightholders and difficulties in obtaining permission from those who either never have or are no longer actively managing their rights. Training ML models from the open internet involves even greater numbers of rightholders.

All of this combined with the novelty and rapid development of ML technologies has resulted in significant legal uncertainty. Unsurprisingly, the use of copyrighted works as part of AI training is already subject to legal dispute in the US and UK.

Most of the discussion about copyright and ML training has been conducted within the parameters provided by the US framework. The question dominating this discussion has been: Do uses of copyrighted works for the purpose of training (generative) ML constitute fair use?[3]

The EU copyright framework does not provide for a fair use defence; users can rely on the system of exceptions and limitations to copyright when they use a work without express permission of rightholders. Therefore the situation is different here.

We argue that questions relating to the input side of ML are sufficiently addressed by the existing EU copyright framework. Since the adoption of the 2019 Copyright in the Digital Single Market Directive (CDSM Directive), the EU copyright framework contains a set of

[2] Model weights are an integral part of the model itself. The model consists of a structure or architecture that defines the way it processes the input data, and the weights are the parameters that are learned during the training process to optimise the model's performance.

[3] For a recent overview, see: Henderson, P. et al. (2023) ‘Foundation Models and Fair Use’ [Unpublished]. Available at: https://arxiv.org/abs/2303.15715.

harmonised exceptions that are applicable to ML as described above. These are the exceptions for text and data mining (TDM) introduced in Articles 3 and 4 of the Directive.

Machine learning and text and data mining

Even though not directly referenced in the Directive, the fight over TDM exceptions during the legislative battle over the CDSM Directive has always been about the ML revolution that was already on the horizon at that time.[4]

The CDSM Directive defines TDM as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations." This definition clearly covers current approaches to ML that rely heavily on correlations between observed characteristics of training data. The use of copyrighted works as part of the training data is exactly the type of use that was foreseen when the TDM exception was drafted and this has recently been confirmed by the European Commission in response to a parliamentary question.

Article 3 of the Directive allows text and data mining for the purposes of scientific research by research organisations and cultural heritage institutions, as long as they have lawful access to the works to be mined, including content that is freely available online.

The Article 4 exception – which is the result of extensive advocacy by researchers, research organisations, open access advocates and technology companies to broaden the scope of the TDM exception in Article 3[5] – allows anyone to use lawfully accessible works, as defined above, for text and data mining unless such use has been "expressly reserved by their rightholders in an appropriate manner, such as machine-readable means."

In sum, these two articles provide a clear legal framework for the use of copyrighted works as input data for ML training in the EU. Researchers at academic research institutions and cultural heritage institutions are free to use all lawfully accessible works (e.g. the entire public Internet) to train ML applications. Everyone else – including commercial ML developers – can only use works that are lawfully accessible and for which their rightholders have not explicitly reserved use for TDM purposes.

[4] The European Parliament's summary published after the adoption of the Directive makes this explicit by noting that "the co-legislators agreed to enshrine in EU law another mandatory exception for general text and data mining (Article 4) in order to contribute to the development of data analytics and artificial intelligence."

[5] This statement by 24 stakeholders stresses "the foundational role that TDM plays in Artificial Intelligence (AI)."

Opt-out and the limits of the copyright framework

The EU's approach constitutes a forward-looking framework for dealing with the issues raised by the mass scale use of copyrighted works for ML training. Importantly, it ensures a fair balance between the interests of rightholders on the one side and researchers and ML developers on the other.

The exception in Article 3 provides much needed clarity for academic researchers and ensures that they have access to all copyrighted works for TDM/ML training purposes. The more limited exception in Article 4 addresses the interests of creators and other rightholders who want to control the use of their works and those who don't.

Creators and rightholders who want to control the use of their works can opt out from TDM/ML either to prevent their works from being used for this purpose or to establish a negotiation position for licensing such uses of their works either collectively or individually. Here, the European Commission should also play an active role in defining technical standards for reserving the right in machine-readable form in order to increase certainty for all parties involved.

What is equally important is that this differentiation also recognizes the fact that for a significant amount of works the rights are not actively managed by their rightholders. This means that under a copyright-by-default regime (i.e works can only be used on the basis of explicit opt-in) these works could not be used for ML learning since obtaining permission from large numbers of rightholders not actively managing their rights would be impossible. Future EU rulemaking should thus maintain the opt-out approach for ML training and ensure that permissionless use remains the default.

Recommendation 1: The EU must maintain the exceptions for text and data mining established in Articles 3 and 4 of the CDSM Directive. The existing opt-out model for commercial uses should be preserved as it establishes a balance between the interests of ML developers on the one hand and creators on the other.

Transparency

However, we should not stop there. For this approach to work in practice it is essential that ML development becomes more transparent. The EU legislator should enact provisions that require providers of generative ML models to publicly disclose the use of all materials used as training data, including copyright-protected works, in a reasonable and proportionate manner so as not to create an undue burden. Creators should be empowered to know which of their works have been used for training and how.

Such a requirement would improve the transparency of ML development and deployment. As such, this requirement would ensure that adherence to EU legal framework governing the use of copyrighted works for ML training can be verified by anyone, particularly those rightholders who seek to control the use of their works.

Finally, a general transparency requirement contributes to the development of trustworthy and responsible AI and is in the public interest.

Recommendation 2: The EU should enact a robust general transparency requirement for developers of generative AI models. Creators need to be able to understand whether their works are being used as training data and how, so that they can make an informed choice about whether to reserve the right for TDM or not.

The two recommendations developed in this paper are closely tied to our Policy Recommendations #2 (Full copyright protection should only be granted to works that have been registered by their authors) and #16 (Creators should have the right to know their audience).

From a copyright perspective, the opt-out mechanism increases legal certainty for all parties involved. On the one hand, it allows creators and rightholders to indicate how their works are to be used in the context of ML training. On the other hand, it provides ML developers with the ability to ensure that their use of training data does not infringe copyright where applicable. Finally, by excluding scientific research from the scope of opt-outs, the system provides an important contribution to academic freedom and to ensuring that ML-related research can flourish in the EU.

We also maintain that transparency is paramount to a copyright and AI system that works for everyone. Creators should be able to track copyright-relevant uses of their works in order to be able to make informed decisions and improve their negotiating position vis-à-vis other actors in the value chain, and the public needs transparency to ensure that as AI continues to progress its benefits flow to society as a whole.

About COMMUNIA

The COMMUNIA association advocates for policies that expand the Public Domain and increase access to and reuse of culture and knowledge. It acts as a network of like-minded activists, researchers and practitioners based in Europe and the United States who seek to limit the scope of exclusive copyright to sensible proportions that do not place unnecessary restrictions on access and use.

COMMUNIA is grateful for the financial support of Arcadia, a charitable fund of Lisbet Rausing and Peter Baldwin.

For more information on COMMUNIA visit our website: www.communia-association.org; or contact us at: communia@communia-association.org.