Kluwer Copyright Blog

A first look at the copyright relevant parts in the final AI Act compromise

Paul Keller (Institute for Information Law (IViR)) · Monday, December 11th, 2023

On Friday evening, after 38 hours of negotiations, representatives of the European Parliament, EU member states and the European Commission reached a provisional agreement on the proposed AI Act. The deal reached on Friday night now paves the way for the adoption of the AI Act in the first half of 2024, bringing to an end a legislative process that has lasted more than two and a half years and during which the scope of the Act has been significantly expanded.

Among the additions to the Act's scope is a set of rules for general-purpose AI (GPAI) systems and models, added to address issues raised by generative AI models such as GPT4 and Midjourney that have become available over the past two years. Among these rules, which were one of the main areas of contention between Council and Parliament until the very end, the final compromise includes **two provisions relevant to copyright**.

A newly introduced article on "Obligations for providers of general-purpose AI models" includes two distinct requirements related to copyright. Section 1(c)[1] requires providers of GPAI models to:

> put in place a policy to respect Union copyright law in particular to identify and respect, including through state of the art technologies where applicable, the reservations of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790;

And section 1(d) requires them to:

> draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office;

Sufficiently clear?

While the former provision is relatively new, the latter requirement has its origins in the European Parliament’s report, which contained a proposed provision that would have required model providers to “document and make publicly available a sufficiently detailed summary of the use of copyrighted training data” (see previous comments on the blog here and here). Here the final version is a clear improvement over the original Parliament text as it no longer suggests that model providers need to distinguish between copyright-protected and public domain training materials and then apply different transparency standards to both, which would be unworkable.

Since its inclusion in the IMCO report in the summer, the requirement to publish “sufficiently detailed summary” had also been criticized by some observers for its unclear nature. Thanks to a recital that emerged during the last weeks of negotiations, we now have some more clarity on the legislator’s intention here. The recital makes it clear that the “summary should be comprehensive in its scope instead of technically detailed, for example by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used.”. It also emphasizes that the template to be provided by the AI Office should “allow the provider to provide the required summary in narrative form”.

This clarification means that concerns about this provision, which focused on the fact that it might require too much detail and effort from model providers, have been largely alleviated.

A narrative report on data sources seems like a perfectly reasonable requirement, and builds on what is already common practice for open source AI models. It is clearly aimed at some of the larger commercial model providers who have been very reluctant to share meaningful descriptions of their training data (see, for example, the model release information associated with Open AI’s GPT4, Meta’s Llama2.0, or Google’s Gemini models). Nevertheless, this relatively light requirement to publish high-level descriptions should be sufficient to allow third parties to determine whether model providers have trained their models on lawfully accessible data sources, as required by Article 4(1) of the CDSM Directive, and therefore in compliance with EU copyright rules.

It is also worth noting that there is considerable overlap in terms of training data transparency with a separate requirement to “draw up and keep up-to-date the technical documentation of the model”, which requires model providers to provide detailed information about “information on the data used for training” provided in the form of datasheets. However, unlike the requirement to provide sufficiently detailed summaries, this technical documentation requirement does not apply to open source AI models, defined as “AI models that are made accessible to the public under a free and open-source licence whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available” An example of such an open source AI model is the GPT-NeoX-20B released by the non-profit AI research lab EleutherAI.

Training AI models = TDM

While the compliance policy provision in Section 1(c) has received less attention, it is probably the more interesting and impactful of the two. On its face, it does not require much – of course, entities operating in the EU must "respect Union copyright law" regardless of whether there is a provision in the AI Act requiring them to do so.

More interestingly, the provision explicitly links the use of copyrighted works for training AI models to the text and data mining (TDM) exception in Article 4 of the CDSM Directive. While there has been relatively broad agreement that this exception applies to the use of copyrighted works for the purpose of training AI models, this has been disputed by some who have claimed that the legislator did not have these types of uses in mind when the TDM exceptions were discussed and subsequently adopted.

The direct reference to Article 4(3) CDSM in the AI Act should put an end to this discussion. Even if the legislator did not have these particular types of uses in mind, the legislator who is about to adopt the AI Act clearly recognizes that Article 4 CDSM applies to these uses.

This is further underlined by the language in the recitals, which makes it clear that "Any use of copyright protected content requires the authorization of the rightholder concerned unless relevant copyright exceptions apply", and then goes on to state, with reference to Article 4(3) of the CDSM Directive, that "where the rights to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorisation from rightholders if they want to carry out text and data mining over such works."

But the recitals go further than simply restating the existing copyright rules, they also contain language that is clearly intended to preempt discussions about the territorial application of the EU's TDM rules.

> Any provider placing a general-purpose AI model on the EU market should comply with [the obligation to put in place a policy to respect Union copyright law], regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of these foundation models take place. This is necessary to ensure a level playing field among providers of general-purpose AI models where no provider should be able to gain a competitive advantage in the EU market by applying lower copyright standards than those provided in the Union.

This passage is noteworthy because it could be read as an attempt to broaden the scope through a recital in a piece of legislation outside of the copyright framework. However, given the realities of AI model training, where training copies take place in a context that can be far removed (both geographically and temporally) from the actual use of the resulting models, this seems like a necessary intervention (although it remains to be seen if such a requirement can be legally feasible or effective) . It is also an intervention that points to the urgent need for international convergence in the way copyright law deals with the use of copyrighted works for AI model training (see also my previous post on this topic) and it seems that with the AI Act, the EU has doubled down on its claim to be a global rule maker in this field.

Overall, the copyright provisions in the AI Act are a step in the right direction. They further consolidate the existing balanced legislative approach adopted by the EU in the 2019 CDSM Directive. Section 1(c) provides additional clarity for both rights holders and AI model providers, but will only really help them if a generally accepted standard emerges in this area. With the AI

Act on its way to becoming law, the development of such a standard must become the focus of policymakers and other stakeholders. In addition, the requirement to publish sufficiently detailed summaries of the content used for training in section 1(d) will make it easier for third parties to understand the sources of training data – including whether the lawful access criterion has been met.

[1] *There is currently no consolidated text of the AI Act with authoritative article numbers. The quotes in the remainder of the text are taken from a compromise proposal that was published by POLITICO on the 6th of December. The author understands that the language reflects the agreed upon text although it is still subject to small technical corrections that will take place over the coming weeks.*

To make sure you do not miss out on regular updates from the Kluwer Copyright Blog, please subscribe here.

Kluwer IP Law

The **2022 Future Ready Lawyer survey** showed that 79% of lawyers think that the importance of legal technology will increase for next year. With Kluwer IP Law you can navigate the increasingly global practice of IP law with specialized, local and cross-border information and tools from every preferred location. Are you, as an IP professional, ready for the future?

Learn how **Kluwer IP Law** can support you.

79% of the lawyers think that the importance of legal technology will increase for next year.
Drive change with Kluwer IP Law.
The master resource for Intellectual Property rights and registration.
79%
Wolters Kluwer
2022 SURVEY REPORT
The Wolters Kluwer Future Ready Lawyer
Leading change

This entry was posted on Monday, December 11th, 2023 at 2:26 pm and is filed under Artificial Intelligence (AI), European Union, Legislative process
You can follow any responses to this entry through the Comments (RSS) feed. You can leave a response, or trackback from your own site.