

Kluwer Copyright Blog

Copyright Office as Pivot Between Copyright Holders and AI-Service Providers

Danny Friedmann (Peking University School of Transnational Law) · Wednesday, May 15th, 2024

The forthcoming article “Creation and Generation Copyright Standards” to be published in NYU JIPEL 2024 (see pre-publication version [here](#)) analyses and critiques the different standards for copyright eligibility between expressive works and generative products in the U.S. and China. This blog post focuses on a balanced solution for the emerging problems on the input and output sides of generative Artificial Intelligence (AI), UpstateNYer, CC BY-SA 3.0, via Wikimedia Commons mostly from the perspective of U.S. copyright law.



On the input side, the risk is that AI-service providers first use the copyrighted works in the training data to train their Large Language Models (LLMs) without permission of the copyright holders and, subsequently, that these AI-service providers and their generated products replace those very same copyrighted works, in the process destroying the market for human authors.

The statements by authors that were consulted by the [U.S. Copyright Office](#) and the explosion of litigation (see [here](#)) in the U.S., where many authors are plaintiffs, demonstrate that for many authors their profession is of existential importance and, zooming out, their role as culture creators is indispensable for society to prevent the dilution of human culture (*Friedmann 2024*).

The problem on the output side has been, so far, the impossibility to know what part of an AI-assisted product was created by authors and what part was generated by AI-services. Therefore, the Copyright Office could not determine whether content is a copyrightable work or a non-copyrightable product.

Solution to the input side of the problem

An optimal solution should reconcile the needs of authors and copyright holders on the one hand, and AI-service providers on the other hand. The authors and copyright holders would like to receive a fair and equitable remuneration, while the AI-service providers would like to have access to high-quality data, such as copyrighted works, so that they can further improve their LLMs and promote the progress of innovation.

Instead of fair learning as a variety of fair use, as advocated by [Casey and Lemley](#), or text-and-data mining exceptions, as discussed by [Dermawan](#), this author prefers a more balanced solution, whereby the Copyright Office should start registering copyrighted works in combination with the authors' metadata as training data for LLMs, enabling AI-service providers to use copyrighted works with the metadata and remunerate these authors of the works in the training data.

Metadata

The metadata (data about one or more aspects of the data) include data that can identify the author/copyright holder of the works, the time of creation, and information about whether the author/copyright holder agrees that the work will be used and, if so, under what conditions and licensing rate. The metadata could include a link to a bank account number so that the author/copyright holder could be directly compensated by the AI-service provider via a smart contract.

Criticasters might argue that the metadata cannot technically survive when they are “put through the wringer” (tokenization and “decapitating” of semantics) from input in the training data to output from the AI-service. However, this does not have to be the case. It is arguably a matter of “law by design” (see [here](#), [here](#) and [here](#)). [Lanier and Weyl](#), who coined the term “digital dignity”, pointed out that AI does not have to be a black box regarding the provenance of the output from the input.

Remuneration

Ideally, remuneration would be proportional to the use in the output. Second best would be a lump sum remuneration. In the absence of a proportional or lump sum compensation system, an output-oriented levy system for AI-service providers to the cultural sector, as suggested by [Senftleben](#), would be a good start, although it would lead to an imprecise allocation to the relevant authors.

Solution to the output side of the problem

It is imperative that authors disclose the extent to which the content was created by themselves, and which part(s) and to what extent the content has been generated by AI. The U.S. Copyright Law's reluctance to accept the eligibility for copyright of the award-winning “Théâtre D'opéra Spatial” (see [here](#)), seems partly motivated by the impossibility for the U.S. Copyright law to ensure a clear delineation between creation and generation.

A copyright office should not have to rely on the veracity or honesty of authors. In addition, it would be burdensome for authors to record the process of each of their creations. OpenAI is already recording every single generated content, if only to learn from these interactions generally (what is the break-off point, which could service as a proxy of the AI's success rate) and to personalize the results for the users, unless the user explicitly requests to delete the "memory" (see [here](#)). If the AI-service providers could give the copyright office access to all of their generated output, the office could review and compare each copyright application to the products in the database that were generated by the AI-service provider.

From idea to expression of an idea

Did the user of the AI-service use increasingly precise and fine-grained instructions, so that the ideas become expressions of ideas? (this was arguably the case in "Zarya of the Dawn" "Théâtre D'opéra Spatial" and "Spring Breeze Brings Tenderness"), or did he or she use existing checkpoints and generation data. The first might be considered creative and the second generative (and also not independently created) and it would play an important role in the assessment of whether the content meets the threshold of originality. Thus, the requirement for transparency should not rely solely on the users of AI. AI-service providers have a significant responsibility to make the provenance visible and traceable.

Conclusion

The emergence of generative AI threatens to replace human authors; writers, artists, musicians, etc., thereby undermining human culture. In an optimistic scenario, generative AI will not replace human authors, and will be merely used as a tool. In that case it is still imperative to be able to distinguish the human involvement in and differentiate (which is required for copyrighted works as the *Naruto* and *Urantia* cases make clear) between human creations and AI assisted-works/products. In addition, the copyright office, as an organization trusted by copyright holders, AI-service providers, and beyond, is in a unique position to facilitate copyright holders' remuneration and AI-service providers' access. This requires institutional reform and a paradigm shift.

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

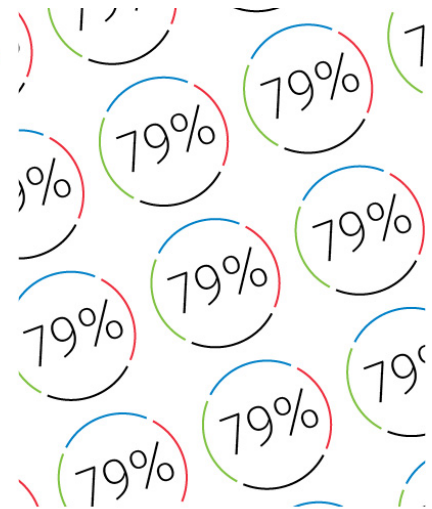
tools from every preferred location. Are you, as an IP professional, ready for the future?

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.



2022 SURVEY REPORT
The Wolters Kluwer Future Ready Lawyer
Leading change

This entry was posted on Wednesday, May 15th, 2024 at 11:31 am and is filed under [Artificial Intelligence \(AI\)](#), [China](#), [USA](#)

You can follow any responses to this entry through the [Comments \(RSS\)](#) feed. You can leave a response, or [trackback](#) from your own site.