Dornis, Tim W. and Stober, Sebastian, Urheberrecht und Training generativer KI-Modelle - technologische und juristische Grundlagen (September 4, 2024). Available at SSRN:

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4946214

"Copyright law & training of generative AI – technological and legal foundations"

Executive Summary (English)

Copyright infringement

During the training of a generative AI model, a number of acts of copying and reproduction of copyright-protected works within the meaning of the InfoSoc Directive and of the German Copyright Act occur:

(1) It starts with the collection, preparation, and storage of copyrighted works used for the AI training process.

(2) In addition, during pre-training and fine-tuning, copyright-relevant reproductions of copyrighted works materialize "inside" the AI model. This also constitutes a copy and replication in the legal sense.

(3) Furthermore, during the application of generative AI models, particularly by the end users of the fully trained AI systems (e.g., ChatGPT via the OpenAI website), works that have been used for training the AI model may be copied and replicated as part of the systems' output.

(4) Finally, what has been overlooked so far, the making available of generative AI models that have been implemented in AI systems for the users of these systems (again, ChatGPT via the OpenAI website) or for downloading as a whole constitutes a making available to the public of the works replicated "inside" the generative AI models according to Sections 15(2)(2) and 19a of the German Copyright Act and Article 3 of the InfoSoc Directive.

Limitations and exceptions

The current canon of copyright limitations and exceptions under European law covers only some of the acts of copyright infringement that come along and are associated with the training of generative AI models. Yet all of these scenarios are irrelevant in

practice. Contrary to voices in scholarly commentary, the exception for text and data mining does not apply to the training of generative AI models, for several reasons:

(1) The acts of copying and reproduction that take place as part of the collection, preparation, and storage of protected works as training data are not subject to the limitation for temporary acts of reproduction (Section 44a of the German Copyright Act and Article 5(1) of the InfoSoc Directive).

(2) The exceptions for text and data mining are inapplicable as well. The exception for text and data mining for scientific research (Section 60d of the German Copyright Act and Article 3 of the 2019 DSM Directive) does not apply to the commercial training of generative AI models.

(3) The exception for commercial text and data mining (Section 44b of the German Copyright Act and Article 4 of the 2019 DSM Directive) is inapplicable: The statutory language and text of the provision, its systematic conception, and the ratio of the exception indicate that it must not be applied to the training of generative AI models.

> (a) This is unveiled by an examination of the technologies underlying both text and data mining and the training of generative AI models: The training of generative AI models does not limit the use of the training data to a simple analysis of the semantic information contained in the works. It also extracts the syntactic information in the works, including the elements of copyright-protected expression. This comprehensive utilization results in a representation of the training data in the vector space of the AI models and thus in a copying and reproduction in the legal sense. Consequently, the training of generative AI models does not fall under the exceptions for text and data mining.
>
> (b) A historical interpretation of the exception for text and data mining confirms the technological and conceptual perspective: Lawmakers, when enacting the 2019 DSM Directive, did not foresee the technological development of creative-productive AI models and their disruptive socioeconomic effects. The text and data mining exception was specifically designed for the analysis of semantic information. Therefore, it cannot be extended to the comprehensive syntax-extracting functionality of generative AI models. Considering the extent to which circumstances have changed since 2019, as well as the still-existing void of substantive analysis and debate on the technological realities, it is also hardly conceivable that the lawmakers of the AI Act had a clear intention to retroactively create an over-extensive scope for the text and data mining exception in the 2019 DSM Directive.
>
> (c) Furthermore, even if one wanted to apply the text and data mining exception, the training of generative AI models would not pass the three-step test of international and European copyright doctrine (implemented, inter alia,

in Article 9 of the Berne Convention for the Protection of Literary and Artistic Works). The comprehensive extraction of syntactic information by generative AI models must be classified as conflicting with the "normal exploitation" by the right holders.

(d) With respect to acts of AI training that occurred in the past, it is important to note that prior to the enactment of the 2019 DSM Directive, copyright-protected works were exploited without any valid exception or limitation.

(4) With regard to the copying and reproduction of copyrighted works in the course of the application of AI systems (particularly when used to produce AI-generated output), no exceptions or limitations apply: Neither the right to quote nor exceptions for caricature, parody, pastiche, or other qualified and excepted purposes will apply.

Applicable law, international jurisdiction, and the AI Act

With regard to private international law and international jurisdiction, it seems to be generally acknowledged among legal scholars that AI training activities abroad (i.e., outside Germany or Europe) are beyond the scope of national and European laws and beyond the jurisdiction of German or European courts. Yet when AI models are offered for application or download to users in Germany (e.g. ChatGPT via the OpenAI website), due to the reproduction of the copyrighted training data "inside" the AI models, these works are made publicly available within the meaning of Sections 15(2)(2) and 19a of the German Copyright Act and Art. 3 of the InfoSoc Directive. This provides both a choice-of-law and a jurisdictional point of attachment to Germany. Accordingly, German copyright law is applicable and German courts have international jurisdiction.

It must further be noted that the 2024 AI Act requires compliance with European copyright law. Hence, the training of generative AI models without the right holder's consent can be classified as both a copyright infringement and a violation of duties in the AI Act. Depending on the circumstances (legal doctrine is still in flux in this regard), sanctions under private law (e.g., Section 823(2) of the German Civil Code) may exist with respect to violations of the AI Act.

Outlook

Under a policy-oriented perspective, three narratives that are currently widely propagated need to be critically examined:

(1) First of all, we must ask ourselves whether "natural" creativity (by humans) will maintain its dominant position in light of the constantly increasing capacities in the field of AI and, accordingly, with regard to artificial creativity. It is not a stretch to

expect that humans as producers of creative products will increasingly be substituted by AI. In this regard, the world is likely to experience a vast reduction in the number of "niches" where human creativity is "superior" to the capacities of AI systems.

(2) Contrary to common prophesies, we will probably not see an overall increase in creative production by humans as a consequence of the growth of artificially creative capacities. Rather, it is likely that the results of genuinely human creativity in many professions and industries—especially in the news and entertainment sector and with regard to everyday products—will be displaced to a considerable extent by generative AI output.

(3) Finally, common warnings that too much copyright protection might stifle AI innovation are misguided. European lawmakers have always rigidly safeguarded regulatory minimum standards (e.g., labor and workplace conditions, environmental protection, international human rights). The beneficial effects on regulation beyond the European Union, often called the "Brussels effect," have even become a hallmark of modern regulatory instruments in the digital economy. Against this backdrop, European lawmakers must ask themselves whether they want to stand idly by and watch the global race to the bottom that has already begun with regard to the protection of copyrights. The appeal to lawmaker action is not about preventing or stifling AI innovation but about establishing a level playing field for AI innovation, as well as fair and equitable compensation for the resources used by AI innovators.