

Tomaž Lajovic

**Foundation Models as Infringers: Should
Large-Scale AI Training Trigger Collective
Licensing Obligations under EU Law?**

Ασπασία

SPLATO

Title:

Foundation Models as Infringers: Should Large-Scale AI Training
Trigger Collective Licensing Obligations under EU Law?

Author:

Tomaž Lajovic

Evaluation by Prof. Dr. Thomas Hoeren, University of Münster,
Law of the Digital Economy Certificate Programme.

Series:

Ασπασία (Aspasia)

Editor:

Tomaž Lajovic

Publication Date:

9 October 2025

Edition:

First Digital Edition (Preprint)

Website:

www.splato.eu

Publisher:

SPLATO

Tomaž Lajovic s.p., Ljubljana, Slovenia

© 2025 SPLATO

This work is licensed under CC BY-NC-ND 4.0.

To view a copy of this license, visit

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

This publication is distributed for academic purposes only

and is not available for commercial sale.

ISBN 978-961-07-2932-7 (PDF)

Katalogni zapis o publikaciji (CIP)

pripravili v Narodni in univerzitetni knjižnici v Ljubljani

COBISS.SI-ID 252408067

ISBN 978-961-07-2932-7 (PDF)

Foundation Models as Infringers: Should Large-Scale AI Training Trigger Collective Licensing Obligations under EU Law?

Legal Essay
as part of the certificate course
Law of the Digital Economy

Author:
Tomaž Lajovic
tomaz.lajovic@gmail.com

Date of Submission: 24/08/2025

Table of Contents

A. Introduction	1
I. Foundation Models and AI Systems	1
II. Provision of FM-based AI Systems	3
B. Use of Copyrighted Works in FM-based AI Systems	3
I. Preparation of Training Datasets	3
II. Creation of Foundation Models	4
III. Output Generation	5
IV. Use of AI Generated Output Outside the AI System	7
V. Issues Pertaining to Moral Rights	7
VI. The Question of Lawfulness of Access, Source and User	8
C. Exemptions and Collective Licensing for the Purposes of Generative AI Systems	10
I. Considerations and Approach	10
II. General Principles of the Exemptions	11
D. Conclusion	12

A. Introduction

The term “foundation models” (FM) has been coined in 2021 by a research group at *Stanford Institute for Human-Centered AI*, defining it “as models trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks” and seeing their potential at the time “as the subject of a growing paradigm shift, where many AI systems across domains will directly build upon or heavily integrate foundation models”¹. It has since been commonly used in reference to the AI systems of generative nature (Gen-AI).

I. Foundation Models and AI Systems

In this essay the focus within the range of FMs is set on large language models (LLMs) “because how LLMs use copyrighted material is obviously relevant to the legal analysis of potential liability for copyright infringement”². Textual content that LLMs are trained on is essentially a number of compilations of word sequences and since computers cannot directly comprehend the words of human languages, the text must be represented by numerical values³.

The central components of LLMs are artificial neural networks that define the way the model is able to be trained and “contain an internal representation of the input that has been processed during the model’s training”⁴. During the training process the model effectively develops its own internal language that itself uses to describe the ingested data, hidden to the outside world within the “model’s black box [that] is mystical, both before and after the training.”⁵ Initial step in training is “tokenisation” that transforms natural and artificial languages’ representations into a set of tokens, forming the LLM’s vocabulary used to represent any text as a series of numbers in a lossless manner – it allows for “map[ping] raw text onto a set of numbers, and a set of numbers back into text”⁶. The next step is “word embeddings” which “provide the representation of a word [token] as a high-dimensional vector”⁷ defining individual token’s “probability distribution over enormously large vector

¹ (Bommasani and Liang 2021, p. 1)

² (Gervais, et al. 2024, p. 1)

³ (Gervais, et al. 2024, p. 1)

⁴ (Dornis 2024, p. 16)

⁵ (Dornis 2024, p. 10)

⁶ (Gervais, et al. 2024, p. 3)

⁷ (Gervais, et al. 2024, p. 3)

spaces”⁸ of other tokens and embeddings. In this process the model learns of “the meaning of a word in the context of the words that surround it, as found in the text that is used during training [..., resulting in ...] a mathematical construct that can efficiently capture the meaning of words based on the various contexts (i.e., word sequences) in which a word can be found”⁹.

According to the taxonomy of FM-based AI systems, proposed by Qinghua Lu and co-authors, FM may be used as a “connector” between interaction and

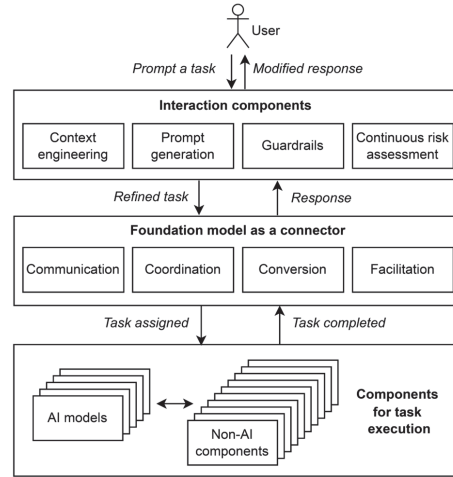


Figure 1: Role of foundation models in implementations of AI systems
Source: (Lu, et al. 2024, p. 4)

task execution components, as well as one of the AI models as a component for task execution¹⁰. As a connector the FM may serve as a communication connector enabling transfer of data between the components, and/or coordination connector coordinating computation of different software (task execution) components, and/or conversion connector adapting data formats communicated to individual software components, and/or facilitation connector optimising interactions between components¹¹.

The distinction between text-and-data-mining (TDM) and Gen-AI systems lies in the fundamentally different ways they use the large-scale data ingested¹². “TDM [...] involves using software to process large volumes of text, images, or other data in order to find patterns [...] while Gen-AI systems] are engineered to process large datasets and algorithmically synthesise outputs.”¹³ Therefore, the purpose of TDM is to “generate information and new

⁸ (Gervais, et al. 2024, p. 4)

⁹ (Gervais, et al. 2024, p. 4)

¹⁰ (Lu, et al. 2024, p. 4)

¹¹ (Lu, et al. 2024, p. 4)

¹² (Lucchi 2025, p. 38)

¹³ (Lucchi 2025, p. 38)

knowledge”¹⁴ and the purpose of Gen-AI is ”the production of data that are similar to the training data.”¹⁵

II. Provision of FM-based AI Systems

An FM-based AI system may be set-up by the user obtaining a copy of FM and implement it in an AI system deployment installed on their own computer infrastructure (either owned or rented) or they can rely on an AI system service provider as an online service, offered to subscribers (paid or free) in the form of Model-as-a-Service (in many flavours, e.g. generalised like Software-as-a-Service or custom deployments like Platform-as-a-Service). The way the AI system is set-up defines the person in whose sphere of influence (i.e. responsibility) an individual downstream process is performed.

B. Use of Copyrighted Works in FM-based AI Systems

The so called “generative-AI supply chain”, as accounted for by Feder Cooper and James Grimmelmann, can be broken-down into eight interconnected stages¹⁶. The first three stages address the original content by (1) creation of works and other information by their respective creators, (2) conversion into digitised data, and (3) collection and curation into training datasets. The next three stages relate to the FM, stating its (4) pre-training, (5) fine-tuning to the specific domain, and (6) public release or embedment in an AI system. The final two stages deal with the use of the AI system by (7) prompt-induced generation of output and (8) alignment of the model with human preferences or usage policies. The authors note that “even to call this a supply ‘chain’ understates its complexity; it is a densely interconnected ecosystem whose stages can branch, recombine, loop, repeat, and feed back into each other.”¹⁷

I. Preparation of Training Datasets

FMs are trained on large datasets of diverse data that may include copyrighted material. Moreover, the AI industry admits that since “copyright today covers virtually every sort of human expression, [... training of FMs] without using copyrighted materials [...] would not provide AI systems that meet the needs of today’s citizens.”¹⁸

¹⁴ (Dornis 2024, p. 9)

¹⁵ (Dornis 2024, p. 9)

¹⁶ (Cooper and Grimmelmann 2025, p. 12)

¹⁷ (Cooper and Grimmelmann 2025, p. 13)

¹⁸ (OpenAI 2023, p. 4)

Regardless of whether a training dataset by itself is copyrighted as a database (copyrighted database) and/or granted sui generis database right protection (sui generis protected database), the copyrighted material collected within – that may include other copyrighted databases – is subject to reproduction (possibly twice, should it be digitised first), and the sui generis protected databases within the datasets are extracted.

Further on, the dataset itself, if it is a copyrighted database, and/or the copyrighted material within, may be distributed or communicated to the public or rented for the purpose of FM training. Such an act constitutes also re-utilisation of the training dataset as a sui generis protected database and of all sui generis protected databases within.

II. Creation of Foundation Models

The stages 4 and 5 mentioned above represent the two phases of FM training with the pre-training phase focusing on quantity of materials and the fine-tuning phase¹⁹ focusing on quality of material used²⁰. In this regard, both data included in the training sets and training set by themselves are deemed materials.

Specifically, LLMs are trained on textual material that is reproduced or translated (or adapted or otherwise altered) and stored within the FM during the training process, using its own language expressed by enormous vector spaces of tokens and embeddings. “The syntax is not filtered or sifted out during generative AI training”²¹ and “the number of possible combinations is astronomical, [... therefore] the only combinations that would have non-negligible weights (i.e., non-zero probability of occurrence) would be the ones that were observed during training.”²² Within the LLM’s black box a lossless reproduction of textual material or “memorization” may (and, in most successful training cases, does) occur, defined as when “(1) it is possible to reconstruct from the model (2) a (near-)exact copy of (3) a substantial portion of (4) that specific piece of training data”²³. “If a generative-AI model has memorized

¹⁹ The fine-tuning phase may be further divided in three sub-phases: supervised fine-tuning, reward modelling, and reinforcement learning. (Gervais, et al. 2024, p. 5)

²⁰ (Gervais, et al. 2024, p. 5)

²¹ (Dornis 2024, p. 21)

²² (Gervais, et al. 2024, p. 4)

²³ (Cooper and Grimmelmann 2025, p. 1), where the authors further explain that they “distinguish ‘memorization’ from ‘extraction’ (in which a user intentionally causes a model to generate a (near-)exact copy), [and] from ‘regurgitation’ (in which a model generates

copyrighted works, the memorized aspects of those works are present in the model itself.”²⁴ It is “clear that copyright-protected works in the training data are *ingested and digested* during generative AI training with regard to *all* information contained [...] and that] the vector spaces inside generative AI models represent elements which, according to legal doctrine, must be considered to be part of the works’ creative and expressive quality.”²⁵ Thus, memorized copyrighted works and copyrighted databases are subject to either reproduction or translation (or adaptation or another kind of alteration), and content of sui generis protected databases is subject to extraction.

The reproductions or translations (or adaptations or some kinds of alterations) and content of sui generis protected databases within the LLM (in the form of its parameters) may subsequently be publicly released and/or embedded in an AI system for deployment in a software service.²⁶ Depending on the actual modalities this may be considered (a) either distribution or communication to the public or rental of the copyrighted works and copyrighted databases memorized within the model and (b) re-utilization of the content of sui generis protected databases extracted to the model.

III. Output Generation

LLMs generate output based on user-invoked prompts. These may be created by the user directly, or automatically generated by software agents following certain programmed procedures and/or using a FM as a connector performing communication, coordination, conversion, or facilitation tasks within the AI system used (as explained in A.I above).

In the process of responding to the user’s prompt an AI system does not just refer to the one LLM embedded within and provide output. Instead, it uses embedded FMs as connectors to address respective FMs and other non-AI components (as components for task execution; see Figure 1 above). These may either be embedded within the AI system or are accessible to it by any

a (near-)exact copy, regardless of the user’s intentions) [...] as ‘memorization’ specifically referring to] ‘reconstruction’ (in which the (near-)exact copy can be obtained from the model by any means, not necessarily the ordinary generation process)”; “memorization is a back-end phenomenon; it describes the characteristics and capabilities of the model itself that directly result from its training, [while] regurgitation and extraction are front-end phenomena; they describe how the model behaves in generating outputs in response to a specific prompt.” (Cooper and Grimmelmann 2025, p. 18),

²⁴ (Cooper and Grimmelmann 2025, p. 5)

²⁵ (Dornis 2024, p. 21)

²⁶ (Cooper and Grimmelmann 2025, p. 13)

means – both hosted on the web²⁷, including those behind paywalls or otherwise technically protected, and any other data sources it may be able to access.

It should be noted that the user’s prompt feed may also include reproductions of copyrighted material and content of sui generis protected databases. Moreover, the prompt or a specific part of it might in certain cases be deemed a copyrighted work/database and/or sui generis protected database by itself. Whether entering a prompt should be treated as an act of publication (in cases of previously unpublished works) depends both (1) on the internal mechanics of the AI system in terms of its treatment of prompts received – either the works/databases within user’s prompts may only be used for alignment limited to the individual user or their impact may exceed these limitations and be effectively ingested into the vector space available to the other users of the AI system or otherwise influencing other users’ use of it – and (2) on the other eventual unattended publication procedures of the AI system in question. Nevertheless, the AI system uses the copyrighted works/databases provided with a prompt by reproducing/altering/extracting them within its components, including vector spaces addressed, in order to come up with an appropriate output.

The LLM’s output as a response to a specific prompt may include copyrighted works/databases and content of sui generis protected databases – either memorized during the model’s training or provided to it with prompts (including the alignment process) or fetched from other components for task execution²⁸ during output generation process – in their original or translated (or adapted or otherwise altered) form. In copyright terms such outputs (or parts thereof) are either direct reproductions or translations (or adaptations or other kinds of alterations) of originals and this is true regardless of whether the LLM’s output is only temporary and used internally within the AI system in its function as a connector. Should the outputs be made accessible to the user (or any other public), an act of either or both communication to the public of the copyrighted works and re-utilization of sui generis protected databases.

²⁷ Apart from generally visible and indexed websites, also publicly non-indexed dark-web sites might be “known” to an AI system.

²⁸ See Figure 1 above.

IV. Use of AI Generated Output Outside the AI System

The output by the LLM containing copyrighted material that has been made accessible to the user may be subject of reproduction, of communication to the public, of translation/adaptation/alteration, and of any other use the user put it to outside of the AI system. Regardless of the output containing copyrighted material in their original or altered form being obtained from an AI system, (1) the copyrighted works within the output retain direct links to the original works and the respective rightsholders their rights and (2) the makers of the databases retain their protection of content of sui generis protected databases²⁹.

V. Issues Pertaining to Moral Rights

The moral rights of authors that must be respected with any use of the copyrighted works (within the Union for the protection of the rights of authors in their literary and artistic works under the Berne Convention for the Protection of Literary and Artistic Works) comprise some form of “the right to claim authorship of the work and to object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his honor or reputation.”³⁰ A relatively straight-forward authorship attribution (or at least source attribution, should authorship itself be undisclosed or otherwise missing) of works is key to respecting the former right to claim authorship, while use of works in AI systems raise complex issues with the author’s right of integrity as subject of the latter right to objection.

Regarding the right to claim authorship the creators of training datasets and developers of LLMs should (jointly) be able to ensure that at least attribution is not intentionally omitted in preparation of the training data, so it can be found within the material ingested and properly accounted for during training, and subsequently stated with the (substantial parts of) copyrighted works in the LLM’s outputs. Whether the developers might be willing to comply with the eventual requirement to make a comprehensive list of all (copyrighted)

²⁹ Protection applies to both extraction and re-utilization of substantial parts of the content and in cases of repeated and systematic extraction and/or re-utilization extends also to insubstantial parts thereof; see Article 7 of the Database Directive (Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases)

³⁰ See Berne Convention for the Protection of Literary and Artistic Works, Article 6bis(1).

material ingested available with the release of their LLMs is questionable, since they “tightly guard their training pipelines as trade secrets [... avoiding] full transparency of dataset composition [as it] would undermine commercial confidentiality and complicate web-scale crawling.”³¹

Ensuring the right of integrity in AI systems may be the easiest in cases when substantial verbatim sections of source material are reproduced in the output. Any depart from it raises concerns – as even relatively modest alterations might lead to the author’s objection, the eventual translations and adaptations are significantly more exposed to such actions. Moreover, “a recent EU study^[32] found that even when AI outputs do not reproduce original content in a recognisable way, they may still infringe moral rights if they mimic the author's style or cause reputational harm. This perspective was endorsed by 67% of surveyed experts, who supported allowing rightsholders to invoke moral rights to oppose AI training, even where economic rights exceptions like TDM might apply.”³³

VI. The Question of Lawfulness of Access, Source and User

“The concept of lawfulness in relation to user status or user acts has been gradually established in EU digital copyright law as a condition for the enjoyment of certain copyright exceptions.”³⁴ If the principle is to be applicable to all cases of exceptions and limitations, regardless of its application being founded on explicit requirements by the law³⁵ or on the jurisprudence by the CJEU³⁶, leads to the conclusion that, in the absence of a license granted by the rightsholder directly, any such use may be deemed illegal³⁷. In this regard, licenses granted by collective management organisations may have effects

³¹ (Lucchi 2025, p. 82)

³² European Commission: Directorate-General for Communications Networks, Content and Technology, Crowell&Moring, IMC University of Applied Sciences Krems, Philippe Rixhon Associates, Technopolis Group and UCLouvain. “Study on copyright and new technologies – Copyright data management and artificial intelligence.” Publications Office of the European Union. 16 March 2022. <https://data.europa.eu/doi/10.2759/570559> (accessed August 10, 2025).

³³ (Lucchi 2025, p. 82)

³⁴ (Synodinou 2025)

³⁵ Such requirement is stated in relation to TDM exception in Articles 3 of the DSM Directive.

³⁶ See CJEU cases *ACI Adam BV and Others v Stichting de ThuisKopie* (C-435/12), *Copydan Båndkopi v Nokia Danmark A/S* (C-463/12), *Vereniging Openbare Bibliotheken v Stichting Leenrecht* (C-174/15).

³⁷ In some jurisdictions (i.e. France, Belgium, Spain, Slovenia) the exclusive right to first publication is also listed among the moral rights of the author, effectively precluding any kind of lawful access to the work beyond extremely narrow private/confidential use of non-published works explicitly allowed by the author himself.

similar to exceptions and limitations due to their obligation to either issue license or provide reasons, based on objective and non-discriminatory criteria, not to.³⁸ Such an obligation does not exist for the rightholder exercising his rights individually, therefore (at least) the cases of mandatory collective management of exclusive rights (otherwise not subject to limitations and exceptions) present an important limitation in the exercise of the right³⁹ and should be treated as such.

Despite being referred to, neither comprehensive definition of the terms “(un-)lawful access”⁴⁰, “(un-)lawfully accessible”⁴¹, “(il-)legal access”⁴², “(un-)lawful acquirer”⁴³, “(un-)lawful user”⁴⁴, “(un-)lawful source”⁴⁵ nor the eventual distinctions between them is set by the copyright law and relevant jurisprudence. “All these different terms cause quite the headache – they all seem related, yet they are differently formulated and some can be understood to refer to completely different things, such as the bona fide character of the user, the status of the copy, or the status of the source from which the copy is obtained.”⁴⁶

Recital 14 of the DSM Directive appears to counter the approach to “lawfulness” as introduced by the CJEU by allowing for a kind of “source-washing” in the case of TDM exception effectively included in the Article 3 of the same directive. It states that “lawful access should also cover access to content that is freely available online”. “A literal interpretation of the free access basis of lawful use [...] could a priori mean that if a copyright protected work is found

³⁸ See Article 16 of the CRM Directive (Directive 2014/26/EU of the European Parliament and of the Council of 26 February 2014 on collective management of copyright and related rights and multi-territorial licensing of rights in musical works for online use in the internal market).

³⁹ As noted also in the Recital 19 of the CRM Directive, referring to rightholders’ freedom of choice in relation to exercising their respective rights: “Where a Member State, in compliance with Union law and the international obligations of the Union and its Member States, provides for mandatory collective management of rights, rightholders’ choice would be limited to other collective management organisations.”

⁴⁰ See Article 3 of the DSM Directive (Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC).

⁴¹ See Article 4 of DSM Directive.

⁴² See Article 6(4) of the Infosoc Directive (Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society).

⁴³ See Article 5(1) of the Software Directive (Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs)

⁴⁴ See Articles 6(1) and 9(1) of the Database Directive.

⁴⁵ See CJEU cases referred to in the note 36 above.

⁴⁶ (Szkalej 2025, p. 312)

on free access (non-technically restricted) online, then it should be lawful to conduct text and data mining, without taking into consideration any contractual restrictions or even the unlawful source of the work. [...] This approach disregards the previous case law of the CJEU in the “lawful source” cases. [...] Therefore], accepting that the “lawful source” requirement is also applicable when mining content that is freely available online would mean that the concept of free access must be more restrictively interpreted. This implies that the free access must have been authorised from the right holder.”⁴⁷

C. Exemptions and Collective Licensing for the Purposes of Generative AI Systems

Use of copyrighted works may be performed either with appropriate licensing secured directly with the authors, other rightsholders, independent management entities, and collective management organisations, or relying on the use within the boundaries of effective exemptions and limitations in accordance with the provisions of the law. Generally, the same applies to exploitation of sui generis protected databases. The approach presented below builds on the considerations of Nicola Lucchi in his study for the EU Parliament⁴⁸ and on the established principles of exemptions and collective management within EU.

I. Considerations and Approach

The continuously growing number of AI system developers and users affects practically every domain of human activity and generates substantial new value (and/or potential harm⁴⁹). With the high volume of works required for FM training, many potential uses of works (both used for training and used on-the-fly) in the process of output generation by AI systems, their authors are right to demand a fair share as compensation for their contribution and loss of income. As even moderately effective system based on individual licensing of billions of copyrighted works used in AI systems seems inconceivable and taking into account that interests of AI developers and users on the one hand and of creators on the other are inherently conflicting, introduction

⁴⁷ (Synodinou 2025)

⁴⁸ (Lucchi 2025)

⁴⁹ So called “market dilution” by AI systems’ generated output may have important consequences in the markets for the types of works included in the training datasets. (Lucchi 2025, page 127)

of a combination of specific exemptions for the AI systems' purposes including creators' remuneration right and either mandatory or extended collective licensing has the potential to provide for a proper balance of rights and obligations for all the stakeholders concerned. The exemption should not apply to at least certain categories of databases should their use in AI systems be deemed either their normal exploitation or the application of exception would be prejudicial to the legitimate interests of the rightsholders (training datasets and LLMs, were they deemed as databases, are the most obvious examples).

Given that legislative introduction of copyright exemptions for development and use of AI systems can generally be justified by (when legislated) securing proper balance between the conflicting interests, they may still be applied only within the boundaries set by the three-step test⁵⁰ and effectively denied should the balance be significantly shifted. The remuneration right should be unwaivable and either managed under extended collective management regime, should an opt-out right be provided for and enforceable, or simply mandatory collectively managed.

II. General Principles of the Exemptions

The analysis in chapter B above has shown that there are three distinct sets of operations in the generative-AI supply chain where, moving downstream, change of operator and eventual commercialisation is most likely. Therefore, the exemptions should apply accordingly, taking their actual proceeds and/or value generated into account.

Purpose of exemption	Rights exempted	Remuneration principle	Reporting and other obligations
Preparation of training datasets	reproduction extraction distribution communication to public rental re-utilisation	tariff per data volume, weighted by category of material levy as share of proceeds, for commercial exploitation (optionally tariff deduced) levy as share of proceeds, for use of datasets in private environments (private use levy) potential special conditions, for non-profits, start-ups ... until commercial exploitation or certain threshold of proceeds	itemised reporting on material included fingerprinting and/or watermarking of dataset inclusion of metadata reporting on proceeds acquisition and transfer of certificates of private use levy paid

⁵⁰ See Article 5(5) of the Infosoc Directive.

Purpose of exemption	Rights exempted	Remuneration principle	Reporting and other obligations
Creation of foundation models	reproduction alteration extraction distribution communication to public rental re-utilisation	tariff per major versions, based on training data volume, weighted by category of material levy as share of proceeds, for commercial exploitation (optionally tariff deduced) levy as share of proceeds, for use of FM in private environments (private use levy) potential special conditions, for non-profits, start-ups ... until commercial exploitation or certain threshold of proceeds	reporting on commercially obtained training sets used itemised reporting on used material included in the training sets fingerprinting and/or watermarking of FMs inclusion of metadata reporting on proceeds acquisition, redemption and transfer of certificates of private use levy paid
Output generation	reproduction alteration extraction communication to public re-utilisation	minimal commercial tariff, based on number of queries received or time of normalised compute×throughput levy as share of proceeds, for commercial exploitation (tariff deduced and certain share of private use levy paid deduced) potential special conditions, for non-profits, start-ups ... until commercial exploitation or certain threshold achieved	reporting on commercially obtained FMs used itemised reporting on material used for output generation reporting on proceeds redemption of certificates of private use levy paid

Table 1: A possible approach to general principles of AI-systems-related exemptions

D. Conclusion

The TDM exception and AI Act are not sufficient to ensure compliance of AI systems’ developers and users with the EU copyright framework and cultural and creative sectors are right to demand appropriate steps to be taken to “safeguard European intellectual property rights in the age of generative AI.”⁵¹ Introduction of specific exemptions with remuneration rights and collective management mechanisms is the right way to secure long-term balance of the stakeholders’ conflicting interests.

⁵¹ (Broad coalition of rightsholders active across the EU’s cultural and creative sectors 2025)

Bibliography

- Bommasani, Rishi, and Percy Liang. “Reflections on Foundation Models.” *Stanford HAI*. 18 October 2021. <https://hai.stanford.edu/news/reflections-foundation-models> (accessed August 9, 2025).
- Broad coalition of rightsholders active across the EU’s cultural and creative sectors. “Joint statement by a broad coalition of rightsholders active across the EU’s cultural and creative sectors regarding the AI Act implementation measures adopted by the European Commission.” *CISAC*. 30 July 2025. <https://www.cisac.org/Newsroom/articles/joint-statement-broad-coalition-rightsholders-active-across-eus-cultural-and> (accessed August 7, 2025).
- Cooper, Feder, and James Grimmelmänn. “The Files are in the Computer: On Copyright, Memorization, and Generative AI.” *arXiv*. 24 March 2025. <https://arxiv.org/pdf/2404.12590> (accessed August 11, 2025).
- Dornis, Tim W. “The Training of Generative AI Is Not Text and Data Mining.” *SSRN*. 19 December 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4993782 (accessed August 8, 2025).
- Gervais, Daniel J., Haralambos Marmanis, Noam Shemtov, and Catherine Zaller Rowland. “The Heart of the Matter: Copyright, AI Training, and LLMs.” *SSRN*. 21 September 2024. <https://ssrn.com/abstract=4963711> (accessed August 3, 2025).
- Lu, Qinghua, Liming Zhu, Xiwei Xu, Liu Yue, Zhenchang Xing, and Jon Whittle. “A Taxonomy of Foundation Model based Systems through the Lens of Software Architecture.” *2024 IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN)*. Lisbon, Portugal: IEEE, 2024. 1-6.
- Lucchi, Nicola. “Generative AI and Copyright - Training, Creation, Regulation.” *European Parliament*. 9 July 2025. <http://www.europarl.europa.eu/supporting-analyses> (accessed August 8, 2025).
- OpenAI. “OpenAI—written evidence (LLM0113).” *UK Parliament*. 5 December 2023. <https://committees.parliament.uk/writtenevidence/126981/pdf/> (accessed August 9, 2025).

Synodinou, Tatiana. “Navigating User Lawfulness in European Copyright Law: From Lawful Use to Lawful Access.” *Kluwer Copyright Blog*. 19 March 2025.

<https://legalblogs.wolterskluwer.com/copyright-blog/navigating-user-lawfulness-in-european-copyright-law-from-lawful-use-to-lawful-access/> (accessed August 10, 2025).

Szkalej, Kacper. „The Paradox of Lawful Text and Data Mining? Some Experiences from the Research Sector and Where We (Should) Go from Here.“ *GRUR International*, 25. March 2025: 307–319.

**Evaluation by Prof. Dr. Thomas Hoeren,
University of Münster, Law of the Digital Economy Certificate Programme.**

Votum on the Essay by Tomaž Lajovic

“Foundation Models as Infringers: Should Large-Scale AI Training Trigger Collective Licensing Obligations under EU Law?”

The essay by Tomaž Lajovic addresses the copyright implications of foundation models (FMs), particularly large language models, under EU law. It examines how training datasets, memorisation, and output generation implicate copyright and database rights, and assesses whether collective licensing regimes or exemptions with remuneration rights could balance the interests of AI developers and rightsholders. The topic is highly relevant, engaging with cutting-edge debates on generative AI and copyright law reform.

Strengths:

The essay is ambitious in scope and demonstrates strong familiarity with both technical and legal aspects of foundation models. It carefully dissects the stages of AI training (dataset preparation, model training, output generation) and links them to copyright doctrines such as reproduction, communication to the public, and extraction of sui generis databases. The discussion of memorisation, regurgitation, and moral rights shows strong engagement with emerging academic literature. The proposal for exemptions coupled with collective licensing is thoughtful and situated within established EU copyright principles, including the three-step test. Stylistically, the essay is detailed, well-referenced, and shows advanced research skills.

Weaknesses:

The main weakness is the essay’s density and tendency towards descriptiveness. Large parts summarise technical processes or statutory provisions without sufficient critical reflection on feasibility or policy trade-offs. For instance, while the licensing proposal is innovative, the essay does not fully assess its economic or institutional viability. Similarly, while moral rights are discussed, the essay does not clearly evaluate whether EU law is likely to expand their role in the AI context. The normative stance is cautious, avoiding clear prioritisation between innovation and protection. Finally, the extensive detail sometimes overwhelms the central argument, which could have been sharpened with more concise framing.

Overall Evaluation:

This is a highly competent and ambitious essay that engages with one of the most pressing issues in copyright and AI law. It excels in doctrinal coverage, integration of technical insights, and originality of its collective licensing proposal. However, its density, descriptive passages, and cautious normative stance prevent it from reaching the top grade. With a sharper evaluative focus and stronger policy analysis, it could have achieved higher distinction.

Final Grade: B

Ασπασία