Scientific Research Publishing

Open Journal of Social Sciences, 2025, 13(8), 263-277
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

The Training of AI Models in the Context of the EU Copyright Law and the AI Act

Susana Navas Navarro

Faculty of Law, Universitat Autònoma de Barcelona, Barcelona, Spain
Email: Susana.Navas@uab.cat

How to cite this paper: Navas, S. (2025). The Training of AI Models in the Context of the EU Copyright Law and the AI Act. *Open Journal of Social Sciences, 13,* 263-277.
https://doi.org/10.4236/jss.2025.138016

Received: July 15, 2025
Accepted: August 10, 2025
Published: August 13, 2025

Abstract

This paper deals with legalities in the EU associated with AI models, especially during their training phase. It addresses the exceptions for text and data mining in the Arts. 3 and 4 of the Directive (EU) 2019/790 (DSM Directive), and Art. 53 of the AI Act. The latter contains provisions ensuring that AI model providers respect copyrights and related rights. It also explains providers' potential civil liability for breaching the obligations set forth in Art. 53, based on their equivalence with service providers, as referred to in the Digital Services Act (DSA).

Keywords

Artificial Intelligence, Large Generative Models, TDM Exception, Copyright Law, Related Rights, Providers Obligations

1. Introduction

Generative artificial intelligence (GAI), especially large generative AI models (LGAIMs), produces outputs (content) derived from the training of the model on extensive datasets that encompass texts and data, including materials protected by the Copyright Law. This circumstance has provoked legal actions by authors and other rights holders against providers and deployers of these LGAIMs in both the United States. In the United States, several lawsuits and class actions related to LGAIMs have already been filed for alleged infringements of the Copyright Act. They can be tracked in real time at https://www.bakerlaw.com/services/artificial-intelligence-ai/case-tracker-artificial-intelligence-copyrights-and-class-actions/ (Consultation date: 22 July 2025). Also see the updated repository at George Washington University: https://blogs.gwu.edu/law-eti/ai-litigation-database/ (Consultation date: 22 July 2025). See the comprehensive study made by the EUIPO, "The development of generative artificial intelligence from a copyright perspective," at https://www.euipo.europa.eu/es/publications/genai-from-a-copyright-perspective-2025 (Consultation date: 22 July 2025) (EUIPO, 2025) and Europe (Durantaye, 2023).

According to Art. 3 nr. 3 of the AI Act "provider" means: "a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge" and to Art. 3 nr. 4 "deployer" means: "a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity" (Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laid down harmonized rules on artificial intelligence, OJEU, 12.7.2024. Abbreviated as "the AI Act").

A LGAIM serves as a representative example of a general-purpose AI model (GPAIM), which is defined in the AI Act as follows: "[A]n AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market" (Art. 3, section 63).

GPAIMs have been subject to regulation, notably by the AI Act. Within this regulatory framework (to be further addressed in Section 3), particular emphasis is placed on respect for copyrights and related rights. Consequently, this contribution examines the Copyright Law, with specific attention to the text and data mining exception in Europe, particularly Directive 2019/790 on copyrights and related rights in the Digital Single Market (OJEU L 130/92, 17.5.2019, hereafter abbreviated as "the DSM Directive") in relation to the AI Act. It is worth noting that there is no EU Copyright Act applicable to all EU member states; instead, there are 27 different Copyright Acts, which have implemented the "Acquis Communautaire" on copyrights and related rights.

This analysis will distinguish between two primary aspects: the text prompt input by the user (Section 2) and the generated output, which raises a multitude of issues concerning the legal protection of data and the training of the model (Section 4). Finally, a set of conclusions will be presented (Section 5).

2. The Text Prompt

As it is well-established, text prompts constitute the content that the users of a GPAIM input into the service. This input can range from a single word or sentence to more extensive textual passages, images, or even entire documents. When the user is the rights holder for the text, video, or image provided as a prompt, no legal issues arise concerning their exploitation. However, if a third party holds rights for the prompts users are obliged to obtain their authorization unless the usage falls within the scope of application of an exception, such as reproduction for private use or temporary and provisional reproduction as part of a technological process (e.g., Art. 2 Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyrights and related rights in the information society, OJEU L 167, 22.6.2001. Abbreviated as "InfoSoc Directive"). Obtaining the consent of rights holders is the responsibility of the user. In this context, for example, a clause on the OpenAI website regarding the use of ChatGPT stipulates: "You represent and warrant that you have all the rights, licenses, and permissions needed to provide input to our services".

It is worth noting that text prompts are often memorized and stored (Biderman, 2023) as material for the training of the model. They may also be reproduced as output ("plagiaristic output". The term "plagiaristic output" is colloquially used to mean a substantially similar copy of an original work.) (Guadamuz, 2024) potentially affecting the rights of users over their inputs, particularly when these inputs exhibit originality (Cámara Águila, 2019; Spindler & Schuster, 2019). At this juncture, the text and data mining (TDM) exception in Europe merits consideration

3. The Training Phase of the AI Model

This section initially adopts a broad perspective before progressively focusing on the main subject of this paper. Consequently, the legal protection of data will be addressed generally in Section I. Then, the exception for TDM in the DSM Directive will be presented (Section 3.1), and after that, the focus will shift to copyright infringements in the AI training phase (Section 3.2).

3.1. The Legal Protection of Data

As mentioned above, GPAIMs are trained using vast amounts of data, necessitating that the system be trained not only with predefined or synthetic data, but also with freely accessible data available on the internet, and this process occurs continuously in a model with self-supervised learning (Bommasani, 2021). This activity carries inherent risks, as the model may acquire new biases when processing novel data, and it may also involve, in the absence of appropriate safeguards, the infringement of regulations such as those concerning personal data protection and intellectual property (IP) rights.

Before proceeding, it is worth clarifying the definition of "data". For this purpose, reference will be made to the current EU Regulation 2023/2854 on data of December 13, 2023 (OJEU nr. 2854, 22.12.2023). Article 2.1 of this regulation defines "data" as "any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audio-visual recording". The relationship between information and data is significant. Data constitutes unordered and unstructured units of information. When data are organized, analyzed, and correlations are identified, it can be accurately stated that data transform into information (Mendo Carmona et al., 2013). The importance of data lies in their potential for re-use through combination, aggregation, or modification with other data. In this paper, a broad conceptualization of data (López-Tarruella, 2021) will serve as the starting point. Data can be categorized into four distinct types (López-Tarruella, 2019):

1) ***Undisclosed data***: These data are generated by a company, its users, or its products. They may or may not circulate on the digital market, and the owner may employ technological measures to protect them. Such data could also be protected by copyright or trade secrets laws. However, under exceptional circumstances, public sector bodies may access the data and metadata needed to interpret and employ them in the performance of their legal duties in the public interest. Data holders with the status of legal persons, other than public sector bodies, are obligated to make these data available in response to a duly justified request (Art. 14-15 Data Act). Furthermore, if a subset of these data is personal, it must be protected in accordance with the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and the free movement of such data (OJEU L 119/1, of 4.5.2016). Abbreviated as "GDPR".

2) ***Data generated directly by internet users***: Such data may be either personal, in which case they are subject to the GDPR, or non-personal, in which case they are protected by Regulation 2018/1725 of the EU, on a regulation on the free flow of non-personal data (OJEU L 303/59, 28.11.2018) It is also important to note that personal data can be inextricably linked with non-personal data, in which case the GDPR applies to the entire dataset (Art. 2.2).

The Commission's Working Document, Impact Assessment accompanying the document Proposal for a Regulation of the European Parliament and of the Council on a framework for the free movement of non-personal data in the European Union (SWD(2017) 304 final, part 1/2, 3) states that "regardless of the amount of personal data included in mixed datasets, the GDPR must be fully complied with in respect to the personal data part of the dataset".

In fact, Article 2.7 of the AI Act specifies that "Union law on the protection of personal data, privacy and the confidentiality of communications applies to personal data processed in connection with the rights and obligations laid down in this Regulation".

3) ***Data protected by EU Law on copyrights and related rights***: In this regard, the DSM Directive establishes, in Articles 3 and 4, an exception or limitation to copyright, enabling public bodies such as research organizations (Román Pérez, 2019; Evangelio Llorca, 2016) cultural heritage institutions, and private bodies or individuals, to reproduce works for TDM, provided they meet certain requirements. In addition to copyrights, other material may be subject to related rights (for example, photographs, databases, phonograms, audio-visual recordings, and press publications). Particularly relevant are the rights of publishers for the online use of their press publications by information society service providers (Art. 15.1

DSM Directive) and the *sui generis* right of the database makers (Arts. 3 and 4 DSM Directive and Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJEU L 77, 27.3.1996). Indeed, in the Directive is stated that no consent from the rights holder of publishers and of the data base makers is required for TDM.

The distinction between authors' rights and the rights of other parties, such as publishers and database makers, is important, as the former are based on originality, while the latter are based on the investment made. Additionally, Directive (EU) 2019/1024 on open data and the re-use of public sector information (OJEU L 172/56, 26.6.2019, abbreviated as "Open Data Directive") stipulates that the IP rights of third parties over public information must be respected (Art. 1.2 lit. c). However, according to Art. 1.6 of the Open Data Directive, the right of the maker of a database shall not be exercised by public sector bodies in order to prevent the re-use of documents or to restrict the re-use beyond the limits of this Directive." In addition, Art. 3.2 of the same Directive states that for documents for which libraries (including university libraries), museums, and archives hold intellectual property rights and for documents produced by public undertakings, Member States shall ensure that, where the re-use of such documents is allowed, such documents shall be re-usable for both commercial and non-commercial purposes.

This does not preclude such data from being subject to TDM if the requirements set out in Articles 3 and 4 of the DSM Directive are met. On the other hand, according to Article 12.2 of the same directive, if the data rights holder is a public body, the re-use of these data, that is, use for purposes other than those of the public service mission for which they were generated (secondary use of data), is subject to certain conditions, and in any case, the re-use may be reviewed every three years.

4) ***Data unprotected by the Copyright Law***: Unprotected data may be used by any party; however, the absence of copyright protection does not preclude their protection by other (legal) means, such as by technological measures (EUIPO, 2025) or contractual agreements (private ordering) (González Otero, 2019; Mezzanotte, 2017). An example is the inclusion of terms and condition by a website owner concerning the use of site content.

3.2. The Intersection between the Text and Data Mining Exception and the AI Act

This section will first delineate the general regulatory frameworks of the TDM exception (3.2.1). The goal is not to furnish an exhaustive analysis but rather to identify the elements for its application to GAI. Subsequently, the obligation of GPAIM providers to respect IP rights, as stipulated in the AI Act, will be examined in detail (3.2.2).

3.2.1. The Text and Data Mining Exception in Europe

While the European Union ideally promotes IP rights, the openness of all data, trade secrets, and plant breeders' rights, the duty of confidentiality must be respected, implying that the use of such data requires authorization by either the data holders or their assignees, unless the data have been made publicly available under an open license permitting such use (Sartor, Lagioia, & Contissa, 2018).

Text and data mining (TDM) is defined as "any automated analytical technique aimed at analysing texts and data in digital format to generate information, including patterns, trends, or correlations" (Art. 2.2 of the DSM Directive).

The DSM Directive, as is widely acknowledged, includes the TDM exception in Articles 3 and 4 (García Vidal, 2019; Vicente Domingo & Rodríguez Cachón, 2021) without providing for the possible remuneration of authors for this specific mode of exploitation. This point has been subject to debate (for advocacy for remuneration, see: Senftleben, 2024) when considering other exceptions to IP rights by which authors are typically compensated for the use of their works. With regard to Article 3, it is important to note that the beneficiaries of this exception are research organizations and institutions responsible for cultural heritage. The primary objective of a research organization is to conduct research activities, and therefore, TDM must be undertaken with this specific purpose. The Directive refers to organizations that reinvest their profits entirely in their research and activities in accordance with a mission of public interest, but does not stipulate that the scientific research must be conducted for non-commercial purposes, which allows for the potential marketing of results obtained through TDM. On the other hand, companies aiming to develop general-purpose AI models or AI systems employing machine or deep-learning approaches that use TDM for commercial purposes must adhere to the requirements outlined in Article 4. In addition, Art. 2.1 of the DSM Directive covers private for-profit entities that conduct research in collaboration with public bodies; nevertheless, if the former has a decisive influence on the latter, they cannot have preferential access to the results generated by TDM.

Article 4 of the DSM Directive establishes an exception to certain IP rights, which could benefit any entity, be it a natural person, legal entity, or public or private sector organization, when engaging in TDM activities for commercial purposes. Recital 18 specifically enumerates certain circumstances, including, but not limited to, the provision of government services, complex business decisions, or the development of new technology. In any case, TDM for scientific research purposes, as previously stated, must remain permissible (Art. 4.4 of the DSM Directive).

To qualify for this exception, lawful access to the work or other subject matter is a prerequisite (Lucchi, 2023). However, rights holders retain the prerogative to prevent TDM activity by reserving their rights, as stipulated by Article 4.3. The reservation of rights over content made available online can be enforced by machine-readable means ("opt-out"). For example, systems can be implemented to prevent bots from crawling the content of websites (in this vein, see the General-Purpose AI Code of Practice Concerning Copyright at https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (Consultation date: 22 July 2025). Rights can also be reserved through the general terms and conditions of a website, where they are accepted by users (disclaimer). If rights holders have reserved their rights, they can preclude the reproduction of works and other subject matter for the purpose of commercial TDM.

Whether under the provisions of Article 3 or Article 4, the TDM exception concerns the reproduction of works and the extraction of database content. It does not extend to other economic exploitation rights, such as the right to communicate, the right to make a work available to the public during the generation of an output, or the right to transform material to make derivative works.

The exception also encompasses the *sui generis* right of the database maker, meaning that works or other subject matter that are part of a database may be extracted and reproduced, provided the requirements set out in Articles 3 and 4 of the DSM Directive are met. Moreover, Article 4.2 stipulates that such reproductions or extractions may be stored only for the duration of the TDM activity. Once its purpose is fulfilled, they must be deleted.

However, it is crucial to consider that exceptions to the economic exploitation rights of authors should not cause unjustified harm to their legitimate interests. Article 7 of the DSM Directive contemplates the application of the three-step rule by referencing the InfoSoc Directive. Therefore, a restrictive interpretation is mandated, potentially even leading to the non-application of the TDM exception if that is the sole means of protecting the author's interests (Rosati, 2024).

In practice, significant transaction costs for miners are associated with obtaining licenses to process immense volumes of data (Novelli et al., 2024).

Furthermore, the legality of web scraping continues to be a matter of debate in Europe among both courts and legal scholars (Klawonn, 2019). If a website is considered a database, the issue arises as to whether TDM activities infringe the *sui generis* right of the website maker. As it stands, there is no infringement when the use focuses on a non-substantial part of the database. In any case, web scraping may be restricted by contract (i.e., website terms and conditions).

3.2.2. The Obligation for GPAIM Providers to Comply with the EU Copyright Law and Related Rights

Providers that place GPAIMs, whether classified as having systemic risks or not, in the EU market are obliged to ensure compliance with the relevant obligations stipulated in the AI Act, particularly Article 53.1. Furthermore, Article 50.4 establishes a transparency obligation when an artistic work is combined with artificially generated content.

1) Mandatory policy to ensure compliance with the Copyright Law

Providers are required to establish a policy to ensure compliance with Union law on copyrights and related rights, specifically to identify and adhere to the reservation of rights expressed by rights holders pursuant to Article 4.3 of the DSM Directive (for the best possible practices or policy in this regard, see: Peukert, 2024; Quintais, 2024). This raises the question of which Union law on copyrights and related rights. There is no overall EU copyright act, but, as mentioned above, there are twenty-seven copyright acts implementing Arts. 3 and 4 of the DSM Directive in different ways. Therefore, this Union law on copyright should be interpreted as copyright directives.

Any provider placing a GPAIM on the Union market must fulfill this obligation, irrespective of the jurisdiction in which the copyright-relevant acts underpinning the training of the GPAIM take place. This is essential to ensure a level playing field among providers of general-purpose AI models; it prevents any provider from gaining a competitive advantage within the Union market by applying lower copyright standards than those established within the Union. The fulfillment of this obligation is of paramount importance to authors and other rights holders.

Concerning the implementation of a policy or best practices to comply with the Copyright Law (Art. 4 of the DSM Directive. Art. 7.1 of the DSM Directive guarantees the non-enforceability of any contractual clause contrary to the exception provided for by Art. 3 (TDM exception for non-profit purposes), a website owner, for instance, may establish contractual clauses prohibiting the use of TDM techniques, as indicated by the Court of Justice of the European Union (CJEU) in the *Ryanair* case. Very illustrative in this regard is the Ryanair case, C-30/14 (January 15, 2015), in which the CJEU understood that, although the extraction of flight schedules and prices from its website was not a violation of copyright (since it was not an original database) nor an infringement of a *sui generis* right held by the company, there was nevertheless a breach of the conditions of use of the airline's website, which enabled the company to prohibit the data-scraping activity (López-Tarruella, 2021).

Another practice is to implement a tool that helps website owners block access to their content or prevent scraping, such as OpenAI's GPTbot (a web crawler or spider. https://platform.openai.com/docs/gptbot#:~:text=GPT-Bot%20is%20OpenAI's%20web%20crawler,%3A%2F%2Fopenai.com%2Fgpt-bot) (Consultation date: 29 June 2025). See the comprehensive study made by the EUIPO (2025). This represents a proactive measure that may be adopted in the future by other AI deployers, and it aligns with Article 4.3 of the DSM Directive regarding the reservation of rights over content made available to the public when parties interested in TDM are private companies pursuing a commercial purpose (EUIPO, 2025), 41 et seq. The opt-out mechanism does not apply to non-commercial scientific research under Article 3 of the DSM Directive). Other practices could include avoiding the duplication of training data, incorporating human feedback, and ensuring oversight. Ultimately, these measures reflect a principle of *copyright by design and by default*, similar to the provision of Article 25 of the GDPR for personal data protection. Providers should take into account state-of-the-art technologies. Here, the similarity to service providers covered by the DSA could serve as a model for stipulating an obligation to implement a user-friendly copyright notification system, whereby, for example, providers of proper notification could disable certain prompts (Peukert, 2024).

Nonetheless, the interpretation of Articles 3 and 4 of the DSM Directive as covering solely the training phase of the system or model (and not the testing and validation phases) implies that AI providers must delete all content used after training the model. This interpretation highlights the obligation of providers to document all the information needed to develop a model. However, if the datasets used in training a system or model have to be deleted due to Article 4.2 of the DSM Directive, how can providers and deployers guarantee the elimination of biases that infringe fundamental rights? Or verify the correct application of the TDM exception and respect for other exclusive rights or contractual clauses? And how would a victim's right to access documentation, including datasets, be enforced in the event of damage? In light of this, it has been proposed that, for the TDM exception to be truly effective, it should be interpreted as encompassing both the testing and validation phases of the model (Novelli et al., 2024). In fact, this interpretation of the TDM exception aligns with Article 53.1(a) of the AI Act, which states that providers shall draw up and keep up-to-date technical documentation of the model, including documentation relating to its training and testing process and the results of its evaluation, which shall contain, at a minimum, the information set out in Annex XI, so that it can be provided upon request to the AI Office and competent national authorities.

2) Required summary of content used

Furthermore, according to Article 53.1(d) of the AI Act, providers of GPAIMs are required to draw up and make publicly available a sufficiently detailed summary of the content used for training a model, employing a template provided by the AI Office. Hosting services, search engine service providers, and providers of large GPAIMs face a similar burden (Peukert, 2024). Indeed, the transparency obligation stipulated in Article 53.1(d) of the AI Act is reminiscent of Articles 15, 26, and 39 of the Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act. Abbreviated as "DSA"). Such transparency obligations are crucial in cases of damage resulting from the infringement of the author's copyrights and related rights. There is an important distinction between Article 53.1(a) and Article 53.1(d) of the AI Act; the former requires technical documentation, while the latter demands public transparency.

According to Recital 107 of the AI Act, the scope of the summary should be comprehensive rather than technically detailed, to assist parties with legitimate interests, including copyright holders, in exercising and enforcing their rights under Union law, for example, by listing the main data collections or sets that were used in training the model, such as large private or public databases or data archives, and by providing a narrative explanation for any other data sources used. The AI Office provides a template for the summary, which makes reporting simple and effective and facilitates the process of furnishing the required information in narrative form. In fact, Article 53.1(d) of the AI Act does not employ the term "data" when discussing training the model but rather the term "content", which is well-established in the EU Copyright Law (Peukert, 2024). Another consideration is that, if the obligation to draw up a summary of content is performed or interpreted as a pre-marketing obligation, it appears to contradict Article 2.8 of the AI Act, which states that the Act is not applicable to any research, testing, or development activity regarding AI systems or models prior to their being placed on the market or put into service.

3) Obligation of transparency stipulated in Article 50.4 of the AI Act

If the GPAIM is integrated within an AI system that is not classified as high-risk, Article 50.4 of the AI Act should also be applicable. This article concerns transparency obligations for providers and deployers of certain AI systems (e.g., low-risk systems). When the content (the output) forms part of an evidently artistic, creative, satirical, fictional, or analogous work or program, transparency obligations are limited to the disclosure of such generated or manipulated content in an appropriate manner that does not hamper the display, quality, or enjoyment of the work during its normal exploitation and use.

The relationship between this rule and the Copyright Law is problematic, insofar as it may affect the author's right to exploit the work and, above all, the moral right of the author to the integrity of the work. In any case, the author's rights must be respected, and his or her consent must be sought.

In any event, providers of GPAIMs may rely on general codes of practice to demonstrate compliance with the obligations set out in Article 53.1 of the AI Act until a harmonized standard is published (Art. 53.4 AI Act). Consequently, technical standards from industry can serve as the primary sources of governance in this (and other) fields. In this vein, the EU has published the General-Purpose AI Code of Practice on July 10, 2025 (https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (Consultation date: 22 July 2025).

4. Copyright Infringement in the Training Phase: Infringements of the Rights of Exploitation of Copyrighted Works and Other Subject Matters

With regard to the output generated by the AI model, the extent to which the Copyright Law is infringed during the training phase should be considered.

Although a case-by-case analysis (Rosati, 2024) would be necessary to definitively address this issue, it can be stated that training an AI model with protected material does not necessarily imply that the generated content infringes the Copyright Law in regard to the material used, nor can it be automatically classified as derivative work. Infringement of the Copyright Law by a model's output can occur in two primary ways. The first arises when the output bears a close resemblance to protected materials used to train the model ("plagiaristic output"), in which case it can be argued that the right of reproduction (the right of reproduction embraces the direct or indirect, temporary or permanent reproduction, by any means and in any form, in whole or in part, of works or other subject matter (Art. 2 InfoSoc Directive). See the CJEU case *Austro-Mechana*, C-433/30) of the materials has been infringed. It is pertinent to recall that the TDM exception refers to a specific phase of the process, namely, the training phase, and as previously suggested, may encompass the testing and validation phases, but its scope does not extend beyond these stages; that is, it does not cover the output generated by the model. Consequently, within the European context, neither the user nor the provider and/or deployer of the model can invoke the aforementioned exception to defend themselves, stressing that the reproduction right of the pre-existing material has not been infringed (Novelli et al., 2024; Rosati, 2024).

The second way infringement can occur is when a model's output indirectly reflects the pre-existing material through the introduction of adaptations or modifications. In such cases, depending on the specific circumstances, the output might reflect an author's distinctive style, merely be inspired by an author's style, or represent a free version of a protected work. These instances are not protected by the Copyright Law. Conversely, it might be claimed that the right to transform the material used to train the model has been infringed, because the TDM exception solely covers extraction and reproduction but not the right to transform. Nevertheless, users could be held responsible for infringement, as their input would have instigated the generation of that specific output. While in certain cases the causal link between input and output may be readily apparent, in the majority of cases this is not so, and no direct causal link exists between the output and the users' text prompts.

When can GPAIM providers be held liable? As previously highlighted, Article 53.1(c) and (d) of the AI Act establishes two significant provider obligations. First, there is an obligation to implement a policy to ensure compliance with the Copyright Law and related rights, particularly under Article 4.3 of the DSM Directive, and second, there is a transparency obligation requiring the publication of a summary detailing the datasets used for training the model (See Recitals nr. 106 and 107 of the AI Act.) However, the breach of this obligation entails only the imposition of a fine; GPAIM providers are not held directly liable under the AI Act for infringing EU Law on copyrights and related rights.

Therefore, to address this question, one might draw legal inspiration from, on one hand, the Case Law of the CJEU, and, on the other hand, the EU Regulation on digital services (OJEU nr. 277, 10.27.2022).

1) As for CJEU case law, two cases are worthy of mention. The first is *YouTube vs. Cyando* (C-682/18, 2021). The CJEU ruled that if a platform fails to meet three due diligence obligations, it can be held liable for communicating an original work to the public. These obligations are: i) to expeditiously remove or block any content published on the platform that infringes exclusive rights, ii) to implement appropriate technological measures to detect Copyright Law infringements if the platform is aware, as it should be aware, that its users are illegally communicating content protected by the Copyright Law, and iii) to implement on the platform any technological tool needed to prevent the illegal communication of works and their sharing with other users and to establish mechanisms that enable users to report protected content that is illegally communicated and shared.

The second relevant case was ruled on by the CJEU in its decision of April 26, 2022 (C-401/19. See: Communication from the Commission to the European Parliament and the Council. Guidance on Article 17 of the Directive 2019/790 on Copyright in the digital single market, COM(2021) 288 final); it concerned the application of Article 17(4) of the DSM Directive. The Court ruled that this provision establishes sufficient safeguards to protect users' rights when online content-sharing service providers comply with certain requirements. While a general duty to monitor user content is not permissible, from the perspective of the CJEU in its ruling analyzing Article 17, a duty to implement automated recognition mechanisms for illegal content might be imposed by Member States on service providers.

Article 17(4) of the DSM Directive states that if no authorization is granted, online content-sharing service providers shall be liable for unauthorized acts of communicating to the public, including making available to the public, copyright-protected works and other subject matter, unless the service providers demonstrate that they have: i) made the best efforts to obtain an authorization, ii) made, in accordance with high industry standards of professional diligence, the best efforts to ensure the inviolability of specific works and other subject matter for which the rights holders have provided the service providers with relevant and necessary information, and in any event, iii) have acted expeditiously, upon receiving a sufficiently substantiated notice from the rights holders, to disable access to, or to remove from their websites, the designated works or other subject matter, and have made the best efforts to prevent future uploads, in accordance with point ii).

In determining whether the service provider has complied with the obligations under paragraph 4, and in light of the principle of proportionality, the following elements, among others, shall be taken into account: i) the type, audience, and extent of the service and the types of works or other subject matter uploaded by the users of the service, and ii) the availability of suitable and effective means of compliance and their cost to service providers (López Richart, 2023; Espín Alba, 2020).

One may question whether AI model providers should have a duty to implement automated recognition mechanisms in the event of intellectual property rights infringements involving users' text prompts or the training phase. Legal scholars have proposed introducing filters or safeguards to prevent users from publishing or using content protected by the Copyright Law (Rosati, 2024; Peukert, 2024).

2) With regard to the DSA, it is worth considering the due diligence obligations of the service provider to foster a transparent and safe online environment, as established in Articles 13-18 concerning illegal content (Art. 3(h): "Illegal content" means any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law.), which may encompass content that infringes the Copyright Law. As stated in Recital 52 of the DSA,

the rules on notice and action mechanisms should be harmonized at the Union level to provide for the timely, diligent, and non-arbitrary processing of notices, based on rules that are uniform, transparent, and clear, and that provide for robust safeguards to protect the rights and legitimate interests of all affected parties, especially the fundamental rights guaranteed by the Charter, irrespective of the Member State in which the parties are established. These fundamental rights include, for the recipients of the service: the right to freedom of expression and of information, the right to respect for private and family life, the right to protection of personal data, the right to non-discrimination, and the right to an effective remedy. For the service providers, they include the freedom to conduct a business, including the freedom of contract, and for parties affected by illegal content, the right to human dignity, and the rights of children. The right to protection of property, including intellectual property, and the right to non-discrimination (Krokida, 2024).

Providers of hosting services should act upon notices in a timely manner, particularly taking into account the type of illegal content in question and the urgency to take action. For instance, such providers must act without delay when the allegedly illegal content involves a threat to the life or safety of persons. After making a decision about whether or not to act upon the notice, the provider of hosting services should inform the individual or entity designating the specific content without undue delay.

Thus, based on both the CJEU decisions and the DSA, it can be inferred that providers of a GPAIM shall be subject to obligations akin to those imposed on service providers, and consequently, non-compliance should entail civil liability for infringing the Copyright and Related Rights Act. Furthermore, AI model providers, in their capacity as such, and when also acting as service providers, should be held cumulatively liable for infringement if they fail to comply with the aforementioned obligations (Geiger & Jütte, 2021). Platforms and AI providers hold these in common in respect of their handling in their handling of issues related to size, scale, and control (Peukert, 2024).

5. Conclusion

GAI comes with legal uncertainties, predominantly concerning the training phase of AI models. The complexities surrounding the protection of data employed in model training, particularly with respect to copyrights and related rights, could escalate in the future. Based on the output generated by AI models, a variety of issues, such as the infringement of copyrights and related rights by both the users and providers of these models, have been subjected to scrutiny from the European legal perspective. The exception for TDM has been presented, and a proposition has been advanced advocating for the expansion of the TDM exception to encompass not only the training phase but also the testing and validation stages of model development, thereby mitigating potential inconsistencies with the AI Act.

The intricate interplay between the TDM exception and the reservation of rights as delineated in Article 4.4 of the DSM Directive, on one hand, and the provisions of Article 53 of the AI Act, on the other, have been further elucidated.

Regarding the attribution of liability to AI model providers, drawing upon the case law of the CJEU and the framework established by the DSA, it is argued that model providers may be subject to liability for infringements of intellectual property rights corresponding to that of service providers. Moreover, when AI model providers also act as service providers, they must bear the dual burden of both forms of liability cumulatively (e.g., fines under the AI Act for non-compliance and civil liability as a service provider under the DSA framework).

Finally, the deployment of automated recognition mechanisms designed to prevent the illegal use of copyrighted works offers AI model providers a viable strategy to mitigate the risk of liability for Copyright Law infringements.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Biderman, S. (2023). *Emergent and Predictable Memorization in Large Language Models.* https://arxiv.org/abs/2304.11158

Bommasani, R. (2021). On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258.*

Cámara Águila, P. (2019). Los conceptos autónomos sobre el objeto de protección del Derecho de autor: El concepto de obra y el concepto de originalidad. In P. Cámara, & I. Águila (Eds.), *La unificación del Derecho de propiedad intelectual en la Unión Europea* (pp. 49-93). Tirant Editorial.

Durantaye, K. (2023). Garbage in, Garbage Out—The Regulation of Generative AI through Copyright. *Zeitschrift für Urheber- und Medienrecht, 10,* 645-660.

Espín Alba, I. (2020). Online Content Sharing Service Providers' Liability in the Directive on Copyright in the Digital Single Market. *UNIO—EU Law Journal, 6,* 100-114. https://doi.org/10.21814/unio.6.1.2705

EUIPO (2025). *The Development of Generative Artificial Intelligence from a Copyright Perspective.* https://www.euipo.europa.eu/es/publications/genai-from-a-copyright-perspective-2025

Evangelio Llorca, R. (2016). La propiedad intelectual sobre obras creadas por personal investigador al servicio de las universidades y otras entidades públicas de investigación. In J. J. Marín López, R. Casas Vallés, & R. Sánchez Aristi (Eds.), *Estudios sobre la Ley de propiedad intelectual: Últimas reformas y materias pendientes.* Editorial Dykinson, S.L.

García Vidal, A. (2019). *Propiedad intelectual y minería de textos y datos: Estudio de los artículos 3 y 4 de la Directiva UE 2019/790.* ADI.

Geiger, C., & Jütte, B. J. (2021). Towards a Virtuous Legal Framework for Content Moderation by Digital Platforms in the EU? The Commission's Guidance on Article 17 CDSM Directive in the Light of the YouTube/Cyando Judgement and the AG's Opinion in C-401/19. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3889049

González Otero, B. (2019). Las excepciones de minería de textos y datos más allá de los derechos de autor: La ordenación privada contrataca. In C. Saiz García, & R. Evangelio Llorca (Eds.), *Propiedad intelectual y mercado único digital europeo.* Tirant Editorial.

Guadamuz, A. (2024). *Snoopy, Mario, Pikachu, and the Reproduction in Generative AI.*

https://www.technollama.co.uk/snoopy-mario-pikachu-and-reproduction-in-generative-ai

Klawonn, T. (2019). Urheberrechtliche Grenzen des Web Scrapings (Web Scraping under German Copyright Law). *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3491192

Krokida, Z. (2024). Large Language Models and EU Intermediary Copyright Liability: Quo vadis? *European Intellectual Property Review, 46,* Article 361.

López Richart, J. (2023). Sistemas de reconocimiento automatizado de contenidos y ponderación de derechos fundamentales: El difícil equilibrio en la transposición del artículo 17 DDAMUD. In J. López Richart, & C. Saiz García (Eds.), *Digitalización, acceso a contenidos y propiedad intelectual*. Tirant Editorial.

López-Tarruella, A. (2019). Propiedad Intelectual, inteligencia artificial y libre circulación de datos. In C. Saiz García, & R. Evangelio Llorca (Eds.), *Propiedad intelectual y mercado único digital europeo*. Tirant Editorial.

López-Tarruella, A. (2021). *Propiedad intelectual e innovación basada en datos*. Editorial Dykinson, S.L.

Lucchi, N. (2023). ChatGPT: A Case Study on Copyright Challenges for Generative AI Systems. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4483390

Mendo Carmona, C., Fernando Ramos, L., Arquero, R., Del Valle-Gastaminza, F., Botezán, I., Sánchez, R. et al. (2013). Del acceso a la reutilización, del dato al documento: Una visión conceptual de la información pública. *Revista Española de Documentación Científica, 36,* e013. https://doi.org/10.3989/redc.2013.3.957

Mezzanotte, F. (2017). Access to Data: The Role of Consent and the Licensing Scheme. In S. Lohsse, R. Schulze, & D. Staudenmayer (Eds.), *Trading Data in the Digital Economy: Legal Concepts and Tools: Münster Colloquia on EU Law and the Digital Economy III* (pp. 159-188). Hart/Nomos.

Novelli, C. et al. (2024). *Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity*.

Peukert, A. (2024). Copyright in the Artificial Intelligence Act—A Primer. *GRUR International, 73,* 497-509. https://doi.org/10.1093/grurint/ikae057

Quintais, J. P. (2024). *Generative AI, Copyright and the AI Act*. https://ssrn.com/abstract=4912701

Román Pérez, R. (2019). Los organismos públicos de investigación en la Ley sobre reutilización de la información del sector público. *Diariolaley*.

Rosati, E. (2024). Infringing AI: Liability for AI-Generated Outputs under International, EU, and UK Copyright Law. *European Journal of Risk Regulation, 16,* 603-627.

Sartor, G., Lagioia, F., & Contissa, G. (2018). The Use of Copyrighted Works by AI Systems: Art Works in the Data Mill. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3264742

Senftleben, M. (2024). AI Act and Author Remuneration—A Model for Other Regions? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4740268

Spindler, G., & Schuster, F. (2019). *Recht der elektronischen Medien* (4th ed.). Beck-Shop.

Vicente Domingo, E., & Rodríguez Cachón, T. (2021). *Minería de textos y datos como (nuevo) límite al Derecho de autor*. Editorial Reus.