UNIVERSITY OF PIRAEUS

DEPARTMENT OF DIGITAL SYSTEMS

Postgraduate Programme in

"LAW AND INFORMATION AND COMMUNICATION TECHNOLOGIES"

Academic year 2022-2023

POSTGRADUATE THESIS of KONSTANTINA KARRA (ΜΔΙ2216)

Text and Data Mining and Artificial Intelligence in the European Union

Εξόρυξη Κειμένου και Δεδομένων και Τεχνητή Νοημοσύνη στην Ευρωπαϊκή Ένωση

Supervisor: Marina Markellou

Piraeus, August 2024

Abstract (EN)

The EU Directive 2019/790 on Copyright and related rights in the Digital Single Market was set in force to align with the demands of the contemporary digital landscape, encompassing 32 Articles and 84 Recitals. This thesis examines the provisions related to text and data mining (TDM), in particular Article 3 and 4 of the Directive 2019/790, which introduce mandatory exceptions for the use of TDM by research organisations and cultural institutions as well as other entities under specific conditions. Both provisions are examined in the context of applicability, implementation as well as the legal implications they pose. In addition, the interplay between TDM and emerging technologies such as generative artificial intelligence is explicitly discussed as well as the implications of the recently enacted EU Regulation 2024/1689 known as AI Act and EU Regulation 2023/2854, the so-called Data Act. Overall, the aim of this study is to provide a comprehensive understanding of the fast-paced evolving landscape of copyright and TDM in the EU.

Περίληψη (GR)

Η Οδηγία της ΕΕ 2019/790 για δικαιώματα πνευματικής ιδιοκτησίας και τα συγγενικά δικαιώματα στην ψηφιακή ενιαία αγορά τέθηκε σε ισχύ για να ευθυγραμμιστεί με τις απαιτήσεις του σύγχρονου ψηφιακού περιβάλλοντος, περιλαμβάνοντας 32 Άρθρα και 84 Αιτιολογικές Σκέψεις. Η παρούσα διπλωματική εργασία εξετάζει τις διατάξεις που σχετίζονται με την εξόρυξη κειμένου και δεδομένων (Text and Data Mining-TDM), ιδίως τα Άρθρα 3 και 4 της Οδηγίας 2019/790, τα οποία εισάγουν υποχρεωτικές εξαιρέσεις για τη χρήση της TDM από ερευνητικούς οργανισμούς και ιδρύματα πολιτιστικής κληρονομιάς, καθώς και άλλες οντότητες υπό συγκεκριμένες συνθήκες. Και οι δύο διατάξεις εξετάζονται στο πλαίσιο της εφαρμογής, της υλοποίησης καθώς και των νομικών συνεπειών που επιφέρουν. Επιπλέον, συζητείται η αλληλεπίδραση μεταξύ της TDM και των αναδυόμενων τεχνολογιών, όπως η γενετική τεχνητή νοημοσύνη (GenAI), καθώς και οι επιπτώσεις του προσφάτως δημοσιευμένου ΕΕ Κανονισμού 2024/1689 για την Τεχνητή Νοημοσύνη (AI Act) και του ΕΕ Κανονισμού 2023/2854 για τα Δεδομένα (Data Act). Συνολικά, ο στόχος αυτής της μελέτης είναι να προσφέρει μια ολοκληρωμένη κατανόηση του ραγδαίως εξελισσόμενου πεδίου των πνευματικών δικαιωμάτων και της εξόρυξης κειμένου και δεδομένων στην ΕΕ.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AI ACT	Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)
CDSM DIRECTIVE	Directive 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, OJ L 130, 17.5.2019, p. 92-125
CJEU	Court of Justice of the European Union
DATABASE DIRECTIVE	Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77, 27.3.1996, p. 20-28
DL	Deep Learning
EU	European Union
GENAI	Generative Artificial Intelligence
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 right in the information society, OJ L 167, 22.6.2001, pages 10-19
LLM	Large Language Model
ML	Machine Learning
LNP	Natural Language Processing
TDM	Text and Data Mining

INTRODUCTION

The evolution of Artificial Intelligence (**AI**) from performing basic tasks to generating human-level output blurs the boundaries between human and machine, hence the principle that only human-created works are eligible for copyright protection is at stake.[1] In addition, the development and learning of AI models raises significant concerns regarding the potential infringement of the content of a third-party.[2] Since the launch to the market of Chatbots such as ChatGPT in November 2023, there is a plethora of applications of generative artificial intelligence systems (**GenAI**) which is available to the public and very easily accessible. These evolved systems, especially those functioning with GenAI technology, require large amounts of data to be trained.[3] The actual value of data is understood through the extraction of insights.[4] AI and GenAI specifically leverage text and data mining techniques (TDM) to analyse extensive datasets and discover new patterns and relationships, a process that cannot be accomplished manually.[5] It is widely known that the training of AI systems involves the utilisation of publicly available data. For example, OpenAI, the company that brought to the market the revolution with the ChatGPT states on its website that their tool is being trained on, among other things, information that is publicly available on the web.[6] Midjourney has also updated its privacy policy by clarifying that the data collected from third-party sources may also include information found on public databases and the public internet.[7] This technique is what is broadly called as data-scrapping.

Data-scrapping or otherwise web-scrapping is the automated process of gathering data from online sources such as websites, databases, and documents[8]. In many jurisdictions, extensive data-scrapping of personal information may constitute a reportable data breach and fall under

[1] McKinsey and Company "What is Generative AI?" (2024). https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai [last access 17 July 2024].
[2] World Economic Forum, "Will copyright law enable or inhibit generative AI?" (2024). https://www.weforum.org/agenda/2024/01/cracking-the-code-generative-ai-and-intellectual-property/ [last access 17 July 2024].
[3] Mariani Marcello and Dwivedi Yogesh Kumar, "Generative artificial intelligence in innovation management: A preview of future research developments" (2024), Journal of Business Research, Volume 175.
[4] Rosati Eleonora, "An EU Text and Data Mining Exception for the Few: Would it Make Sense?" (2018), Journal of Intellectual Property Law & Practice, 13 (6), 429-430.
[5] World Intellectual Property Organization (WIPO) "Generative Artificial Intellligence. Patent Landscape Report" 2024), Geneva:WIPO. https://doi.org/10.34667/tind.49740
[6] Open AI, "How ChatGPT and our language models are developed" https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed
[7] Midjourney. Policies- Privacy Policy. https://docs.midjourney.com/docs/privacy-policy , [last access 19 July 2024].
[8] Tsaone Swaabow Thapelo et al., "Sasscal websapi: A web scraping application programming interface to support access to sasscal's weather data" (2021), Data Science Journal, vol. 20, no. 1, pg. 1 and 9.

the General Data Protection Regulation (**GDPR**). Accordingly, data-scrapping could be subject to copyright law if it involves duplication and utilisation of copyrighted material without permission, and therefore constitute copyright infringement. Unlike data-scrapping, TDM is regarded as a means of research and as it will be analysed below there are exceptions on the EU legal framework that permit this technique.[9]

To further explain, TDM is the application of complex algorithms to transformed alphanumeric datasets with the aim to extract hidden information.[10] From a copyright perspective, TDM is essential for the analysis of vast amounts of digital data, including images, text, and sound, found in extensive industrial data streams, known as "Big Data focusing on words, themes or topics".[11] Through this process new knowledge is obtained and patterns are identified, therefore resulting to the advancement of GenAI.[12] The creation of outputs with GenAI requires TDM stages like accessing, extracting, and mining of content, which may necessitate authorisation from rightholders.[13] To balance exclusivity and TDM, the EU implemented the Directive 2019/790 (**CDSM Directive**), introducing two key TDM exceptions. Thus, despite the completion of five years anniversary of the CDSM Directive, there is yet concern among experts about whether these exceptions truly encourage innovation.

This thesis aims to contribute to the ongoing discourse on the modernisation of copyright laws in the digital era by answering the following matters. How does TDM provisions set in the CDSM Directive impact the use of copyrighted material for the development of AI tools and what are the legal implications for the stakeholders? What is the benefit for research organisations? Which are the uncertainties for the rest of the researchers? What are the implications of lawful access,

[9] Karathanasis Theodoros "Eu Copyright Directive: A 'Nightmare' For Generative Ai Researchers And Developers?" (2023), AI Regulation. com. MIAI Grenoble Alpes. https://ai-regulation.com/eu-copyright-directive-a-nightmare-for-gai/

[10] Blessing Elisha et al., "Utilizing AI and data analytics to derive insights from large datasets, aiding in decision-making processes" (2023).

[11] The notion of Big Data can be described as a large amount of data (volume), which consists of many different types of sources from which the data is collected (variety), processed and analysed extremely rapidly (velocity) and where the question of quality becomes crucial (veracity), resulting in a significant income and revenue (value) (Cano Jenn, "The V's of Big Data: Velocity, Volume, Value, Variety, and Veracity" (2014)).
As a term it first appeared in 2012, along with the notion of "Industry 4.0".

[12] Christensen Kristina, "A European Solution for Text and Data Mining in the Development of Creative Artificial Intelligence: With a Specific Focus on Articles 3 and 4 of the Digital Single Market Directive" (2021), Stockholm, Intellectual Property Law Review, Volume 4, Issue 2. https://stockholmiplawreview.com/wp-content/uploads/2022/01/TA-European-solution-for-Text-and-Data-Mining_ryck_IP_nr-2_2021_A4.pdf

[13] Rosati Eleonora, "Copyright as an obstacle or an enabler? A European perspective on text and data mining and its role in the development of AI creativity" (2019), Asia Pacific Law Review, Volume 27, Issue 2, pages 198-217.

contractual measures, and technological protection measures? Does the EU legislator manage to foster innovation properly through the aforesaid restrictions?

To answer the above, the TDM provisions as well as their legal and practical implications and the challenges posed by technical measures and the lawful access are analysed in the below chapters.

To begin with, Chapter 1 offers a high-level review of the CDSM Directive with focus on the TDM exceptions in the particular provisions.

Chapter 2 provides with an extensive analysis of the technology behind TDM, as well as the actual procedure, the areas of application and the evolution of such areas such as modern AI.

Followingly, Chapter 3 explores the complexities of AI training from the use of TDM and provides with ground-breaking case law that seems to give some clarity to some concerns.

Chapter 4 is divided in 4 subtitles and explains the challenges found in various levels from the use of TDM. Hence, the first subtitle analyses the challenges related to the stakeholders, while the second one explains what constitutes lawful access for the purposes of TDM as well as the relationship with lawful use. The third and fourth subtitle explain the limitations imposed by contractual agreements and the role of TPMs in controlling access and usage of digital works for TDM accordingly. Following the analysis of this chapter the question arising is what are the potential areas of improvement?

In addition, Chapter 5 examines the EU AI Act provisions for copyright compliance in AI model training, highlighting transparency requirements and unresolved conflicts with the CDSM Directive. The importance of developing clear standards and international cooperation to effectively address these challenges.

Finally, the last Chapter proposes EU Data Act as a potential solution aiming to improve data accessibility by excluding machine-generates databases from the *sui generis* protection. The further revision of Article 43 of this Act is proposed as a measure to provide better support to scientific research and data-driven innovation.

1. HIGH-LEVEL REVIEW OF THE EU CDSM DIRECTIVE: FOCUS ON TDM PROVISIONS

Nowadays, online platforms such as search engines, social networks and video-sharing platforms serve as the primary source of information and content.[14] From one perspective, these platforms offer indeed significant economic and social advantages, however on the other side they also facilitate the movement of content which is illegally used. Among the years, the EU bodies have proposed several adjustments to various laws to safeguard basic rights within the digital landscape. As far as it concerns copyright framework, discussions started around 2014 with a purpose to create a digital unified or single market that will provide optimal online access for individual and businesses and constitute a safe harbor for the protection of the rights of the content or database creators.[15] After several deliberations and voting sessions from the different Committees of the European Parliament, in April 2019, the Directive 2019/790 was finally published. [16] The Directive 2019/790 or simply CDSM Directive consists of 32 Articles supplemented by no less than 86 Recitals, aiming to modernise EU copyright law and to adjust the existing rules to the needs of 21st digital century. In particular, it establishes fairer remuneration and increased transparency for creators and rightholders, while also tries to safeguard EU citizens' online freedom of expression and to enable broader use of copyrighted materials for education, research, and cultural heritage.[17] The main objectives of the Directive as described in Recital 3[18] are the adaption of key exceptions to copyright to the digital and the cross-border environment, the improvement of licensing practices and the wider access to content as well as the establishment of a well-functioning marketplace for copyright.[19] The Directive covers exceptions and limitations (Articles 3-7), out-of-commerce works and licensing practices (Articles

[14] Özkent Yasemin, "Social media usage to share information in communication journals: An analysis of social media activity and article citations" (2022), PLoS ONE 17(2): e0263725. https://doi.org/10.1371/journal.pone.0263725

[15] Giannopoulou Alexandra, "Proposed Directive on Copyright in the Digital Single Market: A Missed Opportunity?" (2018), Alexander von Humbolt Institute for Internet und Gesellschaft: Digital Society Blog. https://www.hiig.de/en/proposed-directive-on-copyright-in-the-digital-single-market-a-missed-opportunity/

[16] Quintais João Pedro and Schwemer Sebastian Felix, "The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright?" (2022), European Journal of Risk Regulation. 13(2), 191–217. doi:10.1017/err.2022.1

[17] Kyriakidis, Harris, "Cyprus transposes the EU Copyright Directive in Cyprus law" (2022), Harris Kyriakides. https://www.harriskyriakides.law/insights/news/cyprus-transposes-the-eu-copyright-directive-in-cyprus-law

[18] CDSM Directive, Recital 3.

[19] European Union, "Copyright and related rights in the Digital Single Market, Summary of Directive (EU) 2019/790 on copyright in the Digital Single Market" (2019), Eur-Lex. https://eur-lex.europa.eu/EN/legal-content/summary/copyright-and-related-rights-in-the-digital-single-market.html [last access 9 July 2024].

8-12), the reproduction of works of visual art in the public domain (Article 14), and a whole chapter dedicated to the fair remuneration of authors and performers (Title IV, Chapter 3). The majority of the provisions have been extensively discussed by scholars as they created debates due to their controversial nature including inter alia, the text and data mining exceptions (Articles 3 and 4), the press publishers' right (Article 15), and the liability of user-upload platforms (Article 17), all of which seem to be considered until now unclear or controversial. Article 17, for instance, which is related to certain information society service providers defined as online content-sharing service providers as per Article 2(6) of the CDSM Directive, has been thoroughly commented by the European Copyright Society as to how its provisions should be implemented to the national regime of each EU Member State.[20] Another example is Article 15 which introduced a new related right within the EU landscape, i.e. the protection for press publications so that the publishers may claim remuneration for the online use of their work.[21]

On the other hand, Articles 3 and 4 of the CDSM Directive, which refer to TDM, seemed to attract far less attention in the drafting phase of the Directive, while the latter was introduced very late in the legislative process following a proposal by the Dutch delegation.[22] More specifically, Article 3 introduces a mandatory exception under EU copyright legislation which actually allows acts of reproduction (for copyright subject matter to which they have lawful access) and extraction (for the *sui generis* database right)[23] conducted by research organisations and cultural heritage institutions (hereinafter **research and cultural organisations**) in order to perform TDM for the purposes of scientific research and definitely for non-commercial purposes, without the need for the rightholders permission.[24] Article 4 copies to a great extent the previous article by

[20] L'Institut des jurists d'entreprise, "The Eu Copyright Directive: The Three Most Controversial Provisions" (2020), Partnerblog. https://ibj.be/fr/news/partnerblog/the-eu-copyright-directive-the-three-most-controversial-provisions [last access 09 July 2024].
[21] Tambiama Madiega, "Modernisation Of European Copyright Rules: Directive On Copyright In The Digital Single Market" (2019), European Parliament, Legislative Train. https://www.europarl.europa.eu/legislative-train/package-better-access-to-digital-goods-services/file-jd-directive-on-copyright-in-the-digital-single-market text updated 20 June 2024.
[22] Hugenholtz P. Bernt, "The New Copyright Directive: Text and Data Mining (Articles 3 and 4)" (2019), Kluwer Copyright Blog. https://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ [last access 09 July 2024].
[23] The protection of databases is known as the *sui generis right* — a specific property right for databases that is unrelated to other forms of protection such as copyright. There are cases where copyright and the *sui generis* right are both applicable when the conditions of protection for each right are permit it. The provisions of the Directive 96/9/EC on the legal protection of databases apply to both analogue and digital databases.
European Commission, Protection of databases (2022) https://digital-strategy.ec.europa.eu/en/policies/protection-databases [last access 01 July 2024].
[24] Keller Paul and Warso Zuzanna, "Defining best practices for opting out of ML training" (2023), Open Future policy brief #5 https://openfuture.eu/wp-content/uploads/2023/09/Best-_practices_for_optout_ML_training.pdf

permitting any type of beneficiaries including individuals and institutions beyond research and cultural organisations to reproduce and extract works to which lawful access is granted, regardless of the reason behind the TDM activities (therefore, any activity for commercial purposes), unless rightsholders have explicitly reserved their rights in a machine-readable format. Hence, in case TDM is conducted without permission, it would infringe copyright.[25] By way of an example, a safeguard for the rightholder may be the adoption of an *'opt-out'* or *'contract-out'* mechanism.[26]

TDM particularly, is defined in the CDSM Directive as "*any automated analytical technique aiming to analyse text and data in digital form to generate information such as patterns, trends and correlations*" (Art. 2(2)) as well as "*the automated computational analysis of information in digital form, such as text, sounds, images or data*" enabled by new technologies (Recital 8). This broad definition seems to be clearly related with the ability of the artificial intelligent (**AI**) tools to analyse enormous amounts of data either autonomously or semi-autonomously. In particular, a sub-class of AI called machine learning (**ML**) is utilised to extract information and generate meaning, therefore, the definition of TDM techniques as prescribed in the CDSM Directive definitely applies to most areas of AI/ML either current or prospective, providing that they rely on data analytics.[27]

2. TDM: CONCEPT, SCOPE AND AREAS OF APPLICATION

2.1. General information

The term TDM was firstly introduced by Marti Hearst in the 1990s, as he differentiated it from other concepts, such as the retrieval and natural language processing.[28] Nowadays, thanks to the close relationship between TDM and AI, TDM has conquered the research and commercial sector

[25] Ozkayagan Hande, "Discover insights from the Copyright office hours on artificial intelligence" (2023), Europeana Pro. https://pro.europeana.eu/post/discover-insights-from-the-copyright-office-hours-on-artificial-intelligence [last access 09 July 2024].

[26] Laurent Le Meur, "TDM Reservation Protocol (TDMRep). Final Community Group Report" (2024), W3C Community and Business Groups. https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/

The '*opt-out*' or '*contract-out*' mechanism will be discussed in further detail in the Chapter 4.3 of this thesis, below.

[27] See note 22.

[28] Hearst Marti A., "Untangling Text Data Mining. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics" (1999), College Park, Maryland, USA. Association for Computational Linguistics. pages 3–10. Available at: https://dl.acm.org/doi/10.3115/1034678.1034679

and constitutes a principal element for the development of intelligent applications driven by AI, by using a vast amount of information and data in order to learn from it and even further make decisions.[29] Apparently, the development of AI tools depends on the TDM techniques.[30]

Both mining of text and data constitute interrelated procedures, but their main difference focuses on the type of data they process. Data mining is on the top of pyramid and text mining constitutes a subfield of it which analyses plethora of documents in order to extract insightful information and identify relationships that may be useful for certain purposes.[31]

The technique of text mining has to deal with large amounts of unstructured and without format text data, such as information found in emails, social media posts, documents from the internet or videos, which is identified, extracted and analysed so that this type of unstructured data becomes ultimately a structured format for further use.[32] Text mining is also highly related to other fields, inter alia ML and AI, as will be seen below.

Data mining, on the contrary, constitutes a more analytical process with the extraction of patterns and knowledge from highly formatted and structured data stored in a large database, but it is not the mining of data itself.[33] Also, data mining assists with the "discovery" of hidden patterns or information that may be further used, through search in databases, a procedure which is often a struggle for many experts. Through this technique often businesses may make positive and knowledge-based decisions.[34]

---

[29] Varese Elena "*Can* generative artificial intelligence rely on the copyright text and data mining (TDM) exception for its training?" (2023), GamingTechLaw. https://www.gamingtechlaw.com/2023/02/artificial-intelligence-training-copyright-tdm-exception/

[30] CIPPIC the Samuelson-Glushko Canadian Internet Policy and Public Interest Clinic, "Text & Data Mining On Copyright Protected Works For Use By Generative AI | TDM versus Training: The difference between making and using a dataset" (2024). https://www.cippic.ca/articles/the-difference-between-tdm-and-training 25/04/2024

[31] Lokesh Kumar et. al., "Text Mining: Concepts, process and application" (2013), Journal of Global Research in Computer Science. Volume 4, Issue 3, page 37. Available at https://www.researchgate.net/publication/277160258_TEXT_MINING_CONCEPTS_PROCESS_AND_APPLICATIONS ;
See further Clark, Jonathan, "Text Mining and Scholarly Publishing" (2012), Publishing Research Consortium, page 5-6. Available at: https://www.stmassoc.org/2012_01_01_PRC_Clark_Text_Mining_and_Scholarly_Publishing.pdf

[32] Gary Miner et. al. "Practical Text Mining and statistical analysis for non-structured text data" (2012), First Edition Academic Press, page 55;
See also Lokesh Kumar et. al. *cit ibid*, page 36.

[33] Jiawei Han et. al. "Data Mining – Concepts and Techniques" (2012), Third Edition, Elsevier Inc., page no. 6. e-book available at: < https://archive.org/details/the-morgan-kaufmann-series-in-data-management-systems-jiawei-han-micheline-kambe/page/n19/mode/2up

[34] See note 31, *cit* page 37.

As per the Court of Justice of the European Union (**CJEU**), the term database is so broad that it may encompass diverse forms, either electronic or non-electronic. These include literary, artistic, musical, or other compilations of work, as well as collections of other materials such as sound, images, numbers, facts, and data.[35] Considering that both text mining and data mining are applicable to protected works and other subject matter or are related with data extracted from databases, and since relevant legal questions may arise, it could be considered too restrictive to treat them as entirely separate processes, within the context of this thesis. Therefore, rather than examining them individually, it is suggested that they be analysed jointly, acknowledging their similarities and potential legal consequences.

2.2. Where text mining meets data mining

As aforesaid, from the use of TDM techniques legal issues may occur, therefore it is highly necessary to primarily understand the actual operation of these techniques. The mining of large amount of text or data is conducted through a chain of activities. According to Prof. Dr. Eleonora Rosati there are 3 fundamental common steps for the TDM techniques that include amongst others (1) the access to the subject matter (either text or data); (2) the copying of substantial quantities of the material along with/or the extraction of data; (3) the (actual) mining of either text or data which leads to discovery of knowledge.[36]

In her explicit analysis, she states that the first step refers to accessing the content and for this stage it is very important to identify whether this content is publicly accessible as an open source or whether permission should be sought. Thus, the freedom of access does not necessarily mean that the input material may not be subject to legal restrictions like for example at the stage of extraction or copying. Issues might also appear in relation to works by rightholders not identified or not located, the so-called orphan works.[37] Overall, at this stage the crucial question is to

[35] CJEU, C-490/14, *Freistaat Bayer v. Verlag Esterbauer*, EU:C:2015:735, para 13-14, referring to CJEU, C444/02, *Fixtured Marketing Ltd v. Organismos prognostikon agonon podosfairou AE (OPAP),* EU:C:2004:697, para 23; See further Recital 17 of the Database Directive.
[36] See note 13, *cit* page 8.
[37] Directive (EU) 2012/28/EU of the European Parliament and of the Council of 25 October 2012 on certain permitted uses of orphan works. Article 2 (1) OJ L OJ L 299, pp 5-12).

identify whether the content is freely accessible or not and if not, what kind of permission is required.[38]

The following step is related to the extraction and/or copying of the material, also constituting a preparatory stage, however, not a prerequisite for all TDM techniques. According to the exceptions and limitations provided by Article 5(1) of the Information Society Directive (**InfoSoc Directive**),[39] the transient or incidental copies whose purpose is to enable lawful use (e.g. browsing) do not constitute acts of reproduction that require the relevant authorisation by the rightholder. Back in 2013, the UK Supreme Court supported the above by stating that "*[m]erely viewing or reading it is not an infringement*", meaning that someone who is making temporary copies has lawful access to the content being copied.[40] In fact, Article 2 of the InfoSoc Directive harmonizes and offers a very broad scope for the right of reproduction, by offering an exclusive right to direct or indirect, temporary or permanent reproduction irrespective of the means and form to the rightholders, while temporary acts of reproduction may benefit from the exception stated in Article 5(1) if all conditions are met and therefore not constituting an infringement against Article 2.[41]

According to Recital 33 of the InfoSoc Directive "*a use should be considered lawful where it is authorized by the rightholder or when it is not restricted by law*", therefore it seems that the authorisation of the rightholder is expressly or implicitly needed. A good example of implicit consent of the rightholder could be considered a work offered online without any restrictions for use (and not just mere consumption) without directly stating this.[42] Respectively, on the said matter, it has been affirmed by the CJEU on multiple instances that "*a use should be considered lawful where it is authorised by the rightholder or where it is not restricted by the applicable legislation*".[43] It is arguable that if the reproduction fails to meet the criteria for the exemption outlined in Article

[38] See note 13, *cit* page 8-10.
[39] Directive (EU) 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.
[40] UKSC, *Public Relations Consultants Association Ltd v The Newspaper Licensing Agency Ltd*, (2013) UKSC 18, para 1 (Lord Sumption).
[41] Case C-5/08 Infopaq, *International A/S v Danske Dagblades Forening,* (2009) at paras 54–55.
[42] Vischer, Part 10, "Copyright and AI: Responsibility of providers and users" (2024), https://www.vischer.com/en/knowledge/blog/part-10-copyright-and-ai-responsibility-of-providers-and-users/ [last access 09 July 2024].
[43] CJEU, *Stichting Brein v Jack Frederik Wullems*, C-527/15, EU:C:2017:300, para 65, referring CJEU, *Football Association Premier League and Others*, C-403/08 and C-429/08, EU:C:2011:631, para 168, and CJEU, *Infopaq International,* C-302/10, EU:C:2012:16, para 42.

5(1) of the InfoSoc Directive, then the interpretation of the reproduction by the CJEU in its case law seems quite broad and qualitative evaluation is required.[44]

In case that the extracted/copied content is incorporated in a database, then *the sui generis* right and the copyright apply at the same time. In accordance with the Database Directive[45] the term "*database*" refers to databases "*in any form*" and it is irrelevant whether their format is electronic or not or whether they provide texts, sound, images, or any data with literary, artistic, or musical content[46]. Therefore, intellectual property rights may stand as a boundary to the activities underlying the second step.

Finally, the third step which is the actual mining activity does not constitute an extraction procedure as such rather than a knowledge discovery, through the identification of patterns, trends, and relationships. During this process when the identification and securing of access to the content occur, two more procedures take place ; the pre-processing of the material by turning it into a machine-readable format and simultaneously the removal of unnecessary or unwanted information, in order for TDM techniques to be applied and the uploading of the pre-processed content on a platform, something that happens often.[47] The first procedure is in fact the cleaning of the document so that even unstructured data can be finally read and edited. Also, through this step the unwanted object or information is removed and when the material is finally clear and structured, then the extraction commences, including the segmentation of the document into component terms. This process includes specifically the identification of synonyms, the transformation of the text and the highlighting of equal classes. The main objective is to discover patterns and relations between previously unrelated pieces of material.[48]

[44] See note 4.
[45] Directive (EU) 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.
[46] See, e.g., CJEU, *Freistaat Bayern v Verlag Esterbauer GmbH*, C-490/14, EU:C:2015:735, paras 13-14, referring to CJEU, Fixtures Marketing, C-444/02, EU:C:2004:697, para 23.
[47] Geiger Christophe, Frosio Giancarlo and Bulayenko Oleksandr, "The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects" (2018), Directorate General For Internal Policies, Policy Department For Citizens Rights and Constitutional Affairs https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf
[48] Rosati Eleonora, The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market: Technical Aspects, (2018), Briefing requested by the JURI committee, Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament, page 2.

2.3. Areas of Application

2.3.1. Classification of AI and relation with TDM

In order to be able to identify and assess the risks from the use of TDM from copyright and related rights perspective, it is advisable to first understand the context in which these techniques are employed. TDM is part of AI, ML, Big Data, Natural Language Processing (**NLP**), semantic analysis and other research activity where programming techniques are used to analyse data and contribute with valuable insights.[49] NLP is an area of AI and Linguistics and refers to computer systems that analyse, attempt to understand, interpret, or produce one or more human languages.[50] NLP has recently gained so much attention for this ability to represent and analyse human language computationally.

Overall, AI is an umbrella term for computer science that focuses on the development of computational systems that can mimic human intelligence.[51] Followingly, ML is a subdomain of AI which has led to the development of mathematical and statistical algorithms that efficiently learn from data and imitate the way that humans learn, gradually.[52] Deep Learning (**DL**) is an even further subset of ML technique that uses multiple learning layers, the so-called *neural networks*,[53] to process data and make decisions. A DL model consists of an input layer, at least one hidden layer that performs a nonlinear feature transformation, and an output layer. This ability of mining and following patterns from data has assisted the development of diagnosis prediction and prognosis classification, in addition to tailoring treatments to enhance individual patient

[49] University of Birmingham, Text and data mining (TDM) https://intranet.birmingham.ac.uk/as/libraryservices/library/copyright/text-and-data-mining/text-and-data-mining.aspx

[50] Diksha Khurana & Aditya Koli & Kiran Khatter & Sukhdev Singh, "Natural language processing: state of the art, current trends and challenges" (2022), Springer.

[51] Edd Gent, "What is Artificial Intelligence (AI)?" (2024), Live Science https://www.livescience.com/technology/artificial-intelligence/what-is-artificial-intelligence-ai [last access 09 July 2024].

[52] Badillo et al. "An Introduction to Machine Learning" (2020), Clinical Pharmacology & Therapeutics. 107. 10.1002/cpt.1796.

[53] *Artificial Neural Networks (ANNs), also known as neural networks (NNs) are computer systems modeled after the biological neural networks that make up animal brains. Artificial neurons are linked units or nodes in an ANN that loosely replicate the neurons in a biological brain. Like synapses in the brain, each link may send a signal to other neurons*.
Kabbay, Harcharan, "Artificial Neural Network Concepts and Examples" (2022). Theses. 402. https://irl.umsl.edu/thesis/402

outcomes.[54] Furthermore, GenAI, a subset of DL, is algorithms that following the deductive method of ML, use prompts or existing data to imitate the style of human work and create new content. Basically, by using the human work as a training material, it generates textual, visual, or auditory content based on a computational analysis.[55] In addition, a Large Language Model (**LLM**) is a subset of GenAI, designed in a way to offer efficiently natural language outputs. Its primary emphasis lies in language modeling, a process centered on constructing probabilistic models capable of accurately predicting the subsequent word in a given sequence, leveraging the context provided by the preceding words. The model is trained on vast amounts of text data, permitting it to learn the probability of word occurrences and the patterns in language usage. However, the language modeling task solely depends on form as training data and is inherently incapable of reaching the learning of meaning. By using the term language model, it applies to any system that is trained on the task of string prediction, operating with characters, words, or sentences, while the word meaning is connected with the relation between a linguistic form and a communicative intent.[56] Prof. Luciano Floridi accurately characterised these models for their modus operandi to "*agere sine intelligere*", meaning that they act without understanding exactly the outcome they produce. Nevertheless, their ability to perform intricate tasks and highly generate results, even in the absence of a thorough comprehension of the underlying processes, remains impressive.[57] The aforesaid techniques often interest and complement each other, so for example LLMs are a key component in many NLP applications, and GenAI can leverage both LLMs and NLP techniques to create new, coherent content.[58]

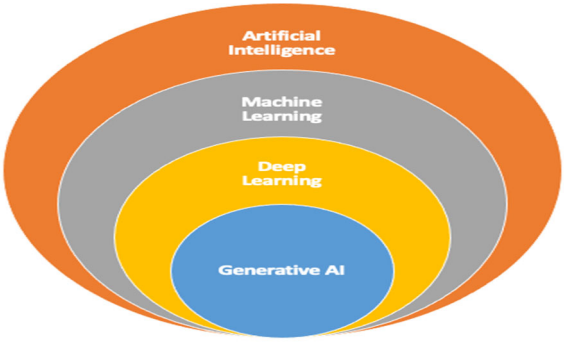
Finally, as aforesaid, TDM is a tool of significant importance for the development of AI, especially creative AI, which functions by the processing of Big Data faster and cheaper.[59] In order to mine, a system usually needs to extract and copy large volumes of digital data, found through various sources such as social media platforms, databases, websites and even digital depositories. While

[54] Johnson KB et al. "Precision Medicine, AI, and the Future of Personalized Health Care" (2021), Clin Transl Sci.;14(1):86-93. doi: 10.1111/cts.12884.
[55] Senftleben, Martin, "Generative AI and Author Remuneration" (2023), International Review of Intellectual Property and Competition Law 54 (2023), pp. 1535-1560, Available at http://dx.doi.org/10.2139/ssrn.4478370
[56] Bender Emily M. and Koller Alexander, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data" (2020), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198, Online. Association for Computational Linguistics.
[57] Floridi Luciano, "AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models" (2023), Philosophy & Technology. 36. 10.1007/s13347-023-00621-y.
[58] Fast Science, "Large language models (LLM) and NLP: A new era of AI and ML has begun" (2024) https://fastdatascience.com/generative-ai/llm-nlp/ [last access 10 July 2024].
[59] Dahlstedt Palle, "Big Data and Creativity" (2019), European Review, Department of Computer Science and Engineering, European Review, Volume 27, Issue 3, page 411-439.

data per se is not copyright-protected, it is the "creative form", namely the "work" that is protected by copyright. Copyright-protected works are converted into a format that forms the basis of TDM.[60]

These acts may cause infringement of the exclusive right of reproduction if not authorised by the rightholder unless the material is in the public domain.[61] Nonetheless, the performance of TDM on non-original compilations of data may also lead to breach of the *sui generis* database right, which operates separately from the copyright protection and is granted to databases that are considered the author's own intellectual creation due to the selection or arrangement of their contents.[62]

On the subsequent chapters it will be analysed whether TDM activities fall under the exclusive rights of reproduction and *sui generis* database right in the EU when mining Big Data that includes protected works and subject matter for AI-driven procedures.



- ***Diagram depicting the relationship between these applications via https://krrai77.medium.com/demystifying-ai-ml-dl-and-generative-ai-oracle-cloud-ai-foundations-associate-2023-cheat-afc310f311bf [last access 3 April 2024].***

2.3.2. From early innovations to modern AI

It was not only until 1945, when at the end of World War II the field of AI started to attract substantial attention. The eagerness for international trade, inspired the ambition to create a

[60] Hugenholtz P. Bernt , "Auteursrecht op informatie (1989), Kluwer, Deventer as discussed in Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, ''Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU (2019), Centre for International Intellectual Property Studies (CEIPI), 6.https://papers.ssrn.com/abstract=3470653
[61] Schlackman Steve, "Who holds the Copyright in AI created art?" (2020), Artrepreneur. Available at: https://alj.artrepreneur.com/the-next-rembrandt-who-holds-the-copyright-in-computer-generated-art/ [last access 3 February 2024];
[62] Database Directive, Art. 3.

machine capable of language translation. The first try was conducted in 1959 by Arthur Samuel, developer of IBM,[63] with the creation of a computer program for playing checkers. Later on, Frank Rosenblatt, inspired by Samuel's work, created the first artificial neural network called the "Mark Perceptron".[64] Between the years 1974 and 1980, AI seemed to have been stuck since there was a limit in data storage and the processing speeds were characterised as slow. This was the period when neural network research seemed to be fading out, leading to a division between ML and AI.

But it was not only until 1980 when IBM developed the first (small) language models that were designed to predict the next word in a sentence.[65] During the 1990s, computational power noted a significant increase and the use of statistical models for NLP analyses was also tremendous, while at the same period DL, the form of ML with additional layers, made its appearance.

Near the end of 2022, OpenAI[66] revolutionised the field of AI with the launch of ChatGPT, an AI chatbot[67] with NLP at its core, able to generate content based on a huge number of pre-existing texts from the internet and to engage with users in a human-like manner.[68] Indeed, they managed to lead the technology to another level as OpenAI's "smarter chatbots" swiftly emerged as a powerful tool, proving assistance to research endeavors, drafting, and generating automatic text-based responses, images or videos.[69]

This famous tool offers blended information that derives from diverse sources such as books, journals, websites, and articles, ultimately resulting in the formation of a unique and engaging discourse, as the technology behind it undertakes to facilitate the understanding and analysis of

[63] Wiederhold, Gio & McCarthy, John, "Arthur Samuel: Pioneer in Machine Learning" (1992), IBM Journal of Research and Development. 36. 329 - 331. 10.1147/rd.363.0329.
[64] Lefkowitz Melanie, Professor's perceptron paved the way for AI – 60 years too soon" (2019), Cornell Chronicle https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon
[65] Foote Keith D., "A Brief History of Natural Language Processing, (2023), Dataversity https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/
[66] San Francisco Company- OpenAI "is an AI research and deployment company, with a mission to ensure that artificial general intelligence benefits all of humanity" as described in their web page. https://openai.com/about
[67] Chatbots are computer programs that use natural language processing and machine learning algorithms to understand and interpret user input and respond with appropriate pre-programmed messages or actions designed to simulate human-to-human conversation, typically via text-based interfaces such as messaging apps, websites, or mobile applications. See Dale R., "The return of the chatbots" (2016) Natural Language Engineering ; Adamopoulou E. and Moussiades L., "An Overview of Chatbot Technology" (2020) Artificial Intelligence Applications and Innovations.
[68] Jianyang Deng and Yijia Lin, 'The Benefits and Challenges of ChatGPT: An Overview' (2022) 2 FCIS 81, 82.
[69] Meyer et al. "ChatGPT and large language models in academia: opportunities and challenges", (2023), BioData Mining. 16. 10.1186/s13040-023-00339-9.

vast amounts of data in a way that mimics human language.[70] Gemini (formerly Bard) and Microsoft Copilot constitute similar GenAI tools, able to generate content respectively based on their training received by neural networks and various datasets.

Despite the outstanding conversational capabilities of a GenAI tool, its operation is limited to understanding syntax, meaning that it solely analyses the sentence's syntactic structure and consequently identifies relationships between its elements.[71] In addition, the outputs generated by the tool are influenced by the data it was trained on, and computational methods employed in it meaning that the results may not always be appropriate or approvable to every specific situation. Finally, it should be highlighted that its AI-powered functions are created without human intervention and despite its exceptional accuracy, it is not free of limitations. Hence, it is advisable that individuals review and modify the designs to ensure that they meet established standards for accuracy and efficiency for specific usage scenarios, to prevent potential issues or errors.

Nonetheless, such machines overall have no intelligence. AI relies on a frustrating learning mechanism as it lacks its own knowledge. It requires a massive amount of data to establish and apply concepts. For example, to recognize a cat, AI needs to analyse around 100,000 cat images, far more than what a human needs. The machine's learning process is repetitive and lacks intelligence, making it less powerful than the human brain. A child, on the other hand, may recognise an animal after seeing it just twice, retaining that recognition for life.[72]

Currently, the main focus of discussions about IP and AI is centered on the question of the authorship and ownership of the output, however there seems to be a lack of attention in relation to the legal issues that may arise when managing the data used to train AI systems.[73] Does the use of copyrighted material for the training of AI programs constitute a copyright infringement?

[70] Jurafsky Daneil and Martin James H., "Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition" (2008), Prentice Hall, Upper Saddle River, New Jersey 07458.
[71] Brown et al. "Language Models are Few-Shot Learners" (2020).
[72] Julia Luc, « L'Intelligence artificielle n'existe pas » First éditions 2019, p.119
[73] Strowel Alain, "ChatGPT and Generative AI Tools: Theft of Intellectual Labor?" (2023), SprinkerLink IIC 54, 491–494 https://doi.org/10.1007/s40319-023-01321-y

3. THE CHALLENGE OF AI'S LEARNING PHASE

As already stated above, AI does not constitute a new project in the market. The latest technological development has been huge and since then we passed from the theoretical stage to the applied technology. There are three essential components making an AI tool operational: i) data analysis algorithms ensuring identification and comparison, ii) massive amount of data and strong computing power,[74] allowing the tool to execute complex computations and iii) data processing tasks.[75]

The internet and the interconnections between databases facilitate increased access to large volumes of diverse data sources. This capability has allowed progress beyond the initial stages where neural networks had previously encountered limitations.[76] A better understanding of the learning process in practise may be given through the example of Compte de Belamy.

Back in 2018, Obvious group[77] with the use of AI called GANs (**Generative Adversarial Networks**) generated a portrait sold at Christie's[78] art auction for $432,500. The painting was marketed by Christie's as "the first portrait generated by an algorithm to come up for auction,'' and with the title Edmond De Belamy. The painting demonstrates a blurry face of a European man of indistinct origin and as per the owners of the Obvious group the system was fed up with a data set of 15,000 portraits painted between the 14th century to the 20th to generate this content. "*The Generator makes a new image based on the set, then the Discriminator tries to spot the difference between a human-made image, and one created by the Generator. The aim is to fool the Discriminator into thinking that the new images are real-life portraits. Then we have a result*".[79] The owners of the group followed the exact same procedure with the "discriminator" and the "generator" on the basis of

[74] Computing power https://computer.howstuffworks.com/computing-power.htm
[75] Portnoff AY and Soupizet, "Artificial intelligence: opportunities and risks. Futuribles" (2018) doi: 10.3917/futur.426.0005
[76] Heudin, Jean-Claude, "Intelligence artificielle et intelligence humaine" (2019), Futuribles. N° 428. 93. 10.3917/futur.428.0093.
[77] A team of French entrepreneurs called OBVIOUS which produces art using artificial intelligence https://obvious-art.com/
[78] Christie's is a world-leading art and luxury business, founded in 1766 https://www.christies.com/about-us/welcome-to-christies
[79] Is artificial intelligence set to become art's next medium? https://www.christies.com/en/stories/a-collaboration-between-two-artists-one-human-one-a-machine-0cd01f4e232f4279a525a446d60d4cd1

the experiment of University of Montreal back in 2014, when GAN systems were used and managed to identify tumours by analysing MRI images.[80]

Indeed, machine learning heavily relies on vast amounts of training data to achieve accurate results. This is where Big Data appear,[81] as these two are inevitably linked. On one hand, Big Data has developed based on the methods and techniques AI uses and on the other hand AI is the one who needs plethora of data. The cooperation of AI with data analytics is the reason why AI cannot perform without Big Data, while each input is extremely important.[82]

As it was analysed in Chapter 2 above, for the effective training of AI, techniques like TDM and generative deep learning are used. TDM involves analysing large amounts of data to find meaningful patterns and insights, improving AI model performance, and allowing it to learn from these patterns and be able to make accurate predictions, enabling content creation and innovation.[83] Therefore, the future of AI relies heavily on the ability of TDM to extract and analyse content such as texts and images on a large scale. However, a significant challenge arises from the fact that AI systems cannot learn from the training material in the same way humans do, as they require an exact replica of the work provided in their training dataset.[84]

To create a training set, millions of examples are required, which involves copying copyrighted images, videos, audio, or text-based works. This raises questions about whether machine copying falls under EU copyright exceptions or in case of the US jurisdiction under fair use. Established companies like Google and Facebook have an advantage in AI due to their access to large language and image datasets.[85] This creates a legal problem for new entrants as dataset ownership and licensing can be complex and the cost of building of licensing datasets from scratch might be prohibitive for smaller companies.[86] Furthermore, if dominant players have the power to control

[80] Wang Z. et al, "Applications of generative adversarial networks (GANs) in radiotherapy: narrative review" (2022), Precision Cancer Medicine (PCM) A Journal Aiming to deliver long lasting blow to cancer. Accessible at https://pcm.amegroups.org/article/view/7431/html
[81] See note 11.
[82] Ahmadi Sina, "A Comprehensive Study on Integration of Big Data and AI in Financial Industry and its Effect on Present and Future Opportunities" (2024) International Journal of Current Science Research and Review, 07 (01), pp.66-74. ff10.47191/ijcsrr/V7-i1-07ff. ffhal-04456267
[83] E Alpaydin, "Introduction to Machine Learning" (2004), Cambridge, MA, MIT Press 2004, page 2. (explaining that TDM is an essential tool for machine learning and data mining, particularly in cases where the data are too numerous or too complex for humans to analyse manually).
[84] MA Lemley and B Casey, "Fair Learning" (2021). Law Review 743, page 775.
[85] W Patry, "Andy Warhol Foundation for the Visual Arts, Inc. v Goldsmith: did the U.S. Supreme Court tighten up fair use?" (2023) 18 Journal of Intellectual Property Law & Practice jpad060.
[86] Vesala Juha, "Developing Artificial Intelligence-Based Content Creation: Are EU Copyright and Antitrust Law Fit for Purpose?" (2023) 54 International Review of Intellectual Property and Competition Law 351.

access to datasets, it could stifle both innovation and competition, giving rise to antitrust concerns. Therefore, it is crucial to prioritise fair and open access to training data in order to effectively address these challenges.

In the United States, Google Books was granted permission to search entire libraries and provide search functions and excerpts from books.[87] However, it remains uncertain whether these permissions extend to data collection and input for machine learning, as there is no copyrightable output. In addition, there is no guarantee that courts will apply the same rulings to comparable technologies.[88] Data collection for TDM may be allowed in the US if it is considered transformative use,[89] but it is not immediately clear if a copyrighted work is being transformed into another copyrighted work. In the Google Books case, the court acknowledged that Google's digitisation of copyrighted books constituted fair use, since the purpose was to enhance users' ability to find and access the original works, rather than to compete with or substitute them. The general intention was to improve the discoverability of the books.

On the other hand, GenAI technology presents a different scenario. These systems can generate new works based on existing content, potentially competing with the original material. Unlike Google's indexing and search functionality, generative AI raises complexities and challenges regarding copyright protection. Several ongoing court cases in the US for instance, seek to clarify the definitions of "derivative work" and "transformative use" under intellectual property law, particularly in relation to copyrighted material used to train AI systems.[90] OpenAI and other GenAI platforms are currently being sued for copyright infringement due to allegations of training their AI systems using unlawfully obtained datasets. For example, In Tremblay v. OpenAI Inc.,[91] the plaintiffs allege that OpenAI utilised their copyrighted books without authorisation to train ChatGPT, implying that the chatbot has assimilated information from the literary works. Similarly, in Silverman et al. v. OpenAI Inc.,[92] the authors claim that ChatGPT can

[87] See *Authors Guild v. Google, Inc*., 804 F.3d 202, 214–15 (2d. Cir. 2015) (The Court granted Google permission to digitise all books available on the market, marking the first step towards creating a book search system that could provide exact excerpts of copyrighted text to users).
[88] See note 87.
[89] A transformative use is one that "alter[s] the first [work] with new expression, meaning, or message". See *Campbell v. Acuff-Rose Music, Inc*., 510 U.S. 569, 579 (1994).
[90] See *Getty Images (US), Inc. v. Stability AI, Inc*., No. 1:23-cv-00135-GBW (D. Del. Mar. 29, 2023);
Also see *Silverman et al. v.OpenAI, Inc. et al.*, No. 4:23-cv-03416 (N.D. Cal. Jul. 7, 2023);
Tremblay et al. v. OpenAI, Inc. et al., No. 4:2023-cv- 03223 (N.D. Cal. Jul. 7, 2023).
It is expected that the outcome of these cases will hinge on the interpretation of the fair use doctrine.
[91] See *Tremblay et al. v. OpenAI, Inc. et al.*, No. 4:2023-cv- 03223 (N.D. Cal. Jul. 7, 2023).
[92] See *Silverman et al. v. OpenAI, Inc. et al.,* No. 4:23-cv-03416 (N.D. Cal. Jul. 7, 2023).

produce summaries of their novels because it was trained using their copyrighted material. Finally, Getty Images Inc. v. Stability AI[93] involves allegations that the software developer of the AI art tool namely Stable Diffusion unlawfully scraped a significant number of Getty Images' photos to train the system, and also modified Getty's watermark, potentially violating copyright management information regulations. Getty Images has filed a similar complaint in the UK, seeking an injunction against Stability AI' s, for selling its AI image generation technology.[94]

Nonetheless, the recent ruling by the US Supreme Court in a non-technological case has raised concerns about the impact on intellectual property rights of works generated by AI.[95] The case involved a dispute over copyright infringement related to a photograph of the musician Prince taken in 1981, which was incorporated by artist Andy Warhol in a series of prints without permission. The Andy Warhol Foundation for the Visual Arts invoked the fair use doctrine to justify the creation of derivative works. However, the Supreme Court ruled against the Foundation, potentially limiting the scope of the transformative use doctrine. This ruling could have implications for the licensing of AI training input. If courts determine that data ingestion, which involves acquiring and modifying data for AI training, constitutes infringement, it could pose legal challenges for the entire AI system.[96] Indeed, GenAI systems have assimilated a significant amount of data, including copyrighted content, without explicit authorisation. The question arises whether using copyrighted works as training data automatically constitutes infringement or if the distinct purpose of training data qualifies for a fair use defense.

On the opposite side, the current TDM framework in the EU seems to be quite restrictive than in other jurisdictions as the EU legislator seems to be taking a protective approach when it comes to training data.[97] The CDSM Directive provides a broad exception for TDM through Article 4(1), allowing commercial AI system developers and educators to make copies of works or databases for extracting information. However, according to Art. 4(3) of the CDSM Directive, rightholders have the option to omit TDM exemptions from their contracts, ensuring they can protect their commercial interests. This provision has faced criticism for being overly restrictive, as it includes

[93] See *Getty Images (US), Inc. v. Stability AI, Inc.*, No. 1:23-cv-00135-GBW (D. Del. Mar. 29, 2023).
[94] Tobin Sam, "Getty asks London court to stop UK sales of Stability AI system" (2023), Reuters https://www.reuters.com/technology/getty-asks-london-court-stop-uk-sales-stability-ai-system-2023-06-01/
[95] See *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 143 S.Ct. 1258 (2023).
[96] See note 88.
[97] Ziaja Gina Maria, "The text and data mining opt-out in Article 4(3) CDSMD: Adequate veto right for rightholders or a suffocating blanket for European artificial intelligence innovations?" (2024), Journal of Intellectual Property Law & Practice, 2024, Vol. 19, No. 5. https://doi.org/10.1093/jiplp/jpae025

factual information and data.[98] The implementation process and the level of compliance from AI developers regarding the opt-out option are yet to be determined.

Compliance with EU data protection legislation, in particular the GDPR is crucial for data aggregation in AI training. The processing of personal data within the EU is followed by rigorous requirements and limitations. These challenges require further exploration and resolution in both legal doctrine and policy to strike a balance between protecting intellectual property rights and facilitating AI development.

Finally, the implementation of EU data protection legislation poses an additional issue for data aggregation, which is crucial for training GenAI models.[99] The gathering and merging of diverse data sources are essential for enhancing the model's capabilities. However, the processing of personal data within the EU is subject to strict requirements and limitations under the GDPR.[100] These challenges require further exploration and resolution in both legal doctrine and policy.

4. CHALLENGES ARISING FROM TDM

4.1. Challenges in relation to the stakeholders

In accordance with article 3(1) of the CDSM Directive the "*Member States shall provide for an exception to the rights provided for in article 5(a) and article 7(1) of Directive 96/9/EC, article 2 of Directive 2001/29/EC and article 15 (1) of this Directive for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, TDM of works or other subject matter to which they have lawful access*".[101] Thus, the CDSM Directive, including with the principles outlined in Recital 11, aims to resolve legal uncertainties regarding TDM.[102] The ambiguity caused may be resolved by introducing a compulsory right and the right to prevent extraction from a database, benefiting universities, research bodies as well as cultural

[98] Margoni Thomas and Kretschmer Martin, "A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology" (2022), 71 GRUR International 685.
[99] Hacker Philipp et al., "Regulating ChatGPT and Other Large Generative AI Models" (2023), in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1112–23 (New York, Association for Computing Machinery 2023) https://doi.org/10.1145/3593013.3594067 [last access 5 March 2024].
[100] 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 on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Apr. 27, 2016).
[101] The relevant article under Greek Law is Article 8 – Law 4996/2022.
[102] Carine Bernault, "Droit d'auteur et open access" (2016) Larcier, No. 198.

institutions. The widespread use of TDM tools by numerous research entities to leverage databases is the origin of the legal uncertainty. Hence, the Directive opts for aligning with what has practically become established without conducting a more thorough assessment of the situation.

The number of the beneficiaries of this exception seems tremendous, including universities and libraries that constitute the primary research organisations that would develop a research tool (CDSM Directive, Recital 11). Additionally, this applies to other entities conducting scientific research (e.g. hospitals and laboratories) or other institutions carrying out educational activities (CDSM Directive Art. 3(1)), which is also related to scientific research on a non-profit basis or under the scope of a public interest mandate acknowledged by a Member State (CDSM Directive Art. 2(1)).

However, an issue arises to organisations over which commercial companies have significant control and this control could give them better access to the results of scientific research and as a result this falls under the scope of definition of research organisations (CDSM Directive Recital 12). This example is unclear in terms of meeting the not-for-profit or public interest requirements. The impact of private donations on an organisation's research status is uncertain as well. Member States have the authority to define research organisations within their national legislation, raising questions about access to data across states with different definitions. While some scholars express concerns about the vague legal terms used in the definition of Art. 2 (1), potentially burdening the CJEU with interpretation,[103] Senftleben believes that an entity may enjoy the perks of a "research organisation" as long as its commitment is focused to public interest, or its non-profit orientation remains unchanged. To this extent, Article 3 of the CDSM may be applicable for scientific research purposes, even if the research project receives industry funding. Moreover, in relation to Article 3 of the CDSM, the primary criterion for assessing potential incompatibility is the requirement for scientific independence, as outlined in Article 2(1) of the CDSM. If an industry gains "decisive influence upon such organisation," the research institution loses its status

[103] Max Planck Institute for Innovation and Competition, "Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules: PART B Exceptions and Limitations: Chapter 1 Text and Data Mining" (2017), 4 https://www.ip.mpg.de/fileadmin/ipmpg/content/stellungnahmen/MPI_Position_Statement_Part_B_Chapter_1_Update23022017.pdf

as a "research organisation" eligible for the TDM privilege.[104] This derives from the explanation of what a research organisation is as per Recital 12.[105]

The Directive's exception is limited to not-for-profit and public interest scientific research.[106] The option of limiting beneficiaries to non-commercial entities was considered by the Commission but deemed counterproductive as valuable discoveries are often made by commercial entities. Distinguishing between commercial and non-commercial in public-private partnerships can be challenging.[107] The exception covers both non-commercial and commercial work done for public-interest purposes as well as the work performed by commercial organisations as part of a public-interest mission. Instead of excluding commercial entities, the EU legislator has chosen to define the eligible institutions for the exception, avoiding any challenges from the interpretation of the term "commercial." Nonetheless, this approach still excludes potential contributors such as start-ups, individual researchers, and journalism.[108] This double limitation is seen as almost equivalent to a restriction on non-commercial purposes Initially, limiting the exception to research organisations aimed to compensate rightholders for the lack of a non-commercial requirement, ensuring that others would pay for a license to mine data. However, this would have negatively impacted unaffiliated researchers and small and medium size enterprises, this is why it has been criticized as a movement against freedom of expression and information,[109] and the freedom to conduct business.[110] The latter may constitute a result of the division of the stakeholders into researchers, corporate research users and rightholders made on the Impact Assessment[111] and therefore this may have led to a diverse group that includes those who use TDM but are not researchers. Overall, the definition of "*research organisation*" appears to have a traditional view of scientific research, limiting innovation in the field and it is important the "who" is conducting the TDM rather than for "what" purpose.

[104] Written by Martin R.F. Senftleben, "Study on EU copyright and related rights and access to and reuse of data" (2022), European Commission. Available at https://archivio.unicas.it/media/7514527/study-on-eu-copyright-and-related-rights-and-access-KI0822205ENN.pdf
[105] CDSM Directive, Recital 12
[106] See note 106.
[107] Bottis Maria and others, "Text and Data Mining in the EU Acquis Communautaire Tinkering with TDM & Digital Legal Deposit" (2019) 12, Erasmus Law Review, 180.
[108] See note 47.
[109] Fundamental Rights Charter art. 16
[110] See note 101.
[111] European Commission, Directorate-General for Education, Youth, Sport and Culture, "Impact Assessment of the European Copyright Framework on Digitally Supported Education and Training Practices" Final report (2016), Publications Office, 116.

Extending the group of beneficiaries to include those with lawful access, as done in the UK, would have been an effective measure.[112] This criterion should not be controversial compared to other IP laws. Similar provisions exist in trade secret and patent laws, allowing for accessing information and increasing scientific knowledge through experimental use.[113] The Software Directive also permits observing and testing program functioning (Software Directive, Art. 5(3)). Copyright holders should anticipate lawful access users engaging in TDM, just as software rightholders expect reproductions.[114] Copyright should accommodate follow-on creativity and fundamental rights exercised through research, protected by limitations and exceptions.[115]

To conclude, the elements of the regime suggested by the Directive appear somewhat incidental considering the scope of the provision. A regime in relation to retaining copies of data extracted from a database for scientific purposes, is declared in Recital 15 of the CDSM Directive, following the example of the Directive on the protection of trade secrets.[116] Member States will decide on the retention of copies, potentially appointing trusted bodies for storage (Recital 15). These arrangements should be proportionate and necessary for secure retention and prevention of unauthorised use. Additionally, Article 5(3)(a) of Directive 2001/29/EC should allow for activities involving the use of copies for scientific research, including scientific peer review and collaborative research, not just TDM, which demonstrates that the outcomes of such uses may only be used for data mining and TDM training.

4.2. Challenges surrounding lawful access

Both TDM exceptions in the CDSM Directive require lawful access to the work, including open access, licenses, and freely available online works.[117] However, these examples, as set in Recital

[112] See note 47.
[113] 76/76/EEC: Convention for the European patent for the common market (CPC 1975) [1975] OJ L 17
[114] See note 106.
[115] Sean Flynn and João Pedro Quintais, "Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for Action at International Level" (21 April 2020), Kluwer Copyright Blog.
[116] Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure.
See adoption of this Directive under Greek Law No. 4605/2019.
See also Würtenberger Gert, Protection of trade secrets and know-how in the European Union: the EU Trade Secrets Directive (EU) 2019/943, (20 August 2019) https://ip-iurisdictio.org/protection-of-trade-secrets-and-know-how-in-the-european-union-the-eu-trade-secrets-directive-eu-2019-943/
[117] CDSM Directive art. 3(1) and 4(1) and Recital 14.

14, do not provide legal certainty as they simply provide an indication of what lawful access should be[118] and in fact the CDSM Directive does not address whether access to a public library is sufficient for mining legal deposits for instance.[119] It is also unclear whether researchers can send copies for mining across borders, even if both nations permit TDM. Questions may arise such as what happens if an EU researcher wishes to transfer a legally obtained database to a research partner in the US for mining?[120] Or what is the connection between "*lawful access*" and "*lawful user*" from the temporary reproduction exception (InfoSoc Directive Art. 5(1) and Recital 33)? Indeed, using the same phrase for both concepts may have provided clarity.[121]

The notion of "lawful user" first appeared in the Computer Programs Directive,[122] though it lacked a clear definition, since no clear terminology was used to describe those entitled to exceptions. At a later stage, the Database Directive[123] used the term "lawful user" to describe the person who is entitled for the exceptions set in this specific Directive. In fact, the interpretation in both directives reflects a progress in copyright law, recognising users as individuals entitled to certain legal prerogatives in the form of obligatory copyright exceptions.[124] In addition, both directives established exceptions which neither shall be upheld by the Member States, nor shall be overridden by contractual terms.[125] Later on, the InfoSoc Directive introduced the concept of "lawful use" in Art. 5. by stating that the mandatory temporary copy exception applies to reproductions required for network transmission by an intermediary or for any lawful use of a work or other subject matter. The CJEU has interpreted "lawful use" variably, initially adopting a broad approach but later linking it to the author's consent and lawful access. More specifically, in Svensson,[126] the CJEU decided that free access to content that was made available to the Internet with the authorisation of the right holder indicates a lawful access. While in GS Media,[127]

[118] Synodinou Tatiana Eleni, "Lawfulness for Users in European Copyright Law: Acquis and Perspectives" (2019) JIPITEC 20 para 1. https://www.jipitec.eu/archive/issues/jipitec-10-1-2019/4876/JIPITEC_10_1_2019_20_Synodinou
[119] See note 110.
[120] See note 118.
[121] See note 47.
[122] Directive (EU) 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version) OJ L 111, 5.5.2009, p. 16–22.
[123] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77,27.3.1996, p. 20–28.
[124] See note 121.
[125] See Article 8 of Directive 2009/24/EC (codified version of Directive 91/250) and Article 15 of Directive 96/9/EC.
[126] Case C-466/12, *Nils Svensson and Others v Retriever Sverige AB*, [2014], ECLI:EU:C:2014:76
[127] Case C-160/15, *GS Media BV v Sanoma Media Netherlands BV and Others*, [2016], ECLI:EU:C:2016:644

the Court declared that the liability of the user is important and depends on whether the user knew or ought to have known that a work was made available without the rightholder's consent.

The aftermath from the aforesaid cases is that the "freely available" criterion is a fundamental condition for lawful use in the TDM exceptions.[128] Thus, concerning the CDSM Directive, a lot of discussions have been raised for allowing rightholders to forbid TDM other than simply granting licenses or applying extra charges for TDM.[129] There are also concerns about the rise of licensing fees and potential disparities between research institutions and Member States.[130] Some scholars argue that legality of access should not matter if it doesn't harm the market, while others believe the opposite, meaning that unlawfully accessing a work is a complete harm to the market.[131] The lawful access requirement aims to protect private actors and ensure payment to rightholders, striking a balance with users of the TDM exception.[132] This basically means that rightholders are compensated through this requirement for lawful access, and they cannot exclude TDM from license terms for scientific research exceptions at least.[133]

It is challenging to determine how lawfulness was enforced in the implementation of a TDM exception in different Member States prior to the CDSM Directive. Thus, France requires lawful access, while the Estonian Copyright Act indirectly suggests a requirement through attribution. Germany's copyright law does not address the matter explicitly.[134] Outside of Europe, national copyright laws often focus on the purpose and commercial nature of the data use.[135] A good example of the aforesaid is the fair use doctrine that applies in the US, however it should be stated that in the US, TDM has not been extensively ruled on since the legality of unlicensed TDM has not been explicitly decided by their courts.[136]

---

[128] Kretschmer Martin, Eleni Synodinou Tatiana, Margoni Thomas, "The Paradox Of Lawful Access" (2024) European Copyright Society (ECS). Available at https://europeancopyrightsociety.org/wp-content/uploads/2024/06/kretschmer-synodinou-margoni.pdf

[129] See note 124.

[130] See note 60.

[131] See note 118.

[132] European Commission, "Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining: Report from the Expert Group" (2014) (Publications Office of the European Union) 51 https://op.europa.eu:443/en/publication-detail/-/publication/d12e3edd-0960-46d1-a7ea-bda1b9cec42d/language-en

[133] Jondet Nicolas, "The Text and Data Mining Exception in the Proposal for a Directive on Copyright: Why the European Union Needs to Go Further than the Laws of Member States" (2018), 67 Propriétés Intellectuelles 25, 19.

[134] Caspers M and Guibault L, "Deliverable D3.3 Baseline Report of Policies and Barriers of TDM in Europe" (2016), FutureTDM https://www.futuretdm.eu/knowledge-library/ page 8.

[135] See note 135, *cit* page 9.

[136] See note 13.

Finally, there are a lot of discussions around the use of TDM for research, and a strong argument is that it lies in the public interest to prevent the rightholders from the control over its use.[137] The development of an exception that does not indirectly empower rightholders to the extent that researchers are compelled to rely on platforms like Sci-Hub and LibGen to bypass paywalls, seems crucial.[138] However, it is important to bear in mind that exceptions generally provide specific uses for defined purposes, rather than granting unrestricted access.

4.3. Challenges surrounding contractual protectional measures

The Member States are to provide for an exception or limitation for reproductions and extractions of legally accessible works and other protected subject matter for the purposes of TDM. While Article 3 clearly designates its mechanism as an exception, the Directive does not specify the category to which the provision in Article 4 belongs.

Rightholders may often restrict TDM access and use, either with a license or by making it subject to additional payment.[139] However, any contractual provision in a license that is contrary to the scientific research exception by excluding TDM is unenforceable.[140] On the other side, the general exception introduced in Article 4(3) of the CDSM Directive can be used unless the rightholder has explicitly reserved his right to restrict use by stating it in the terms and conditions of his website, through contractual agreements or by machine-readable methods for example (CDSM Directive Recital 18). This would create problems to an independent researcher who needs access to sources from major publishing houses if the license agreement excludes TDM. Researchers affiliated with institutions may use the scientific research exception, while individual researchers or journalists might need to pay for an extended license. Unfortunately, for sources not protected

[137] European Copyright Society, "General Opinion on the EU Copyright Reform Package" (2017) page 5 Available at: https://europeancopyrightsociety.org/wp-content/uploads/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf
[138] Balázs Bodó, "The Science of Piracy, the Piracy of Science. Who Are the Science Pirates and Where Do They Come from: Part 1" (6 March 2019, Kluwer Copyright Blog. https://copyrightblog.kluweriplaw.com/2019/03/06/the-science-of-piracy-the-piracy-of-science-who-are-the-science-pirates-and-where-do-they-come-from-part-1/ [last access 1 April 2024].
[139] See note 22.
[140] Rosati Eleonora, "No step-free copyright exceptions: The role of the three-step in defining permitted uses of protected content (including TDM for AI-training purposes)" (2024) European Intellectual Property Review https://www.diva-portal.org/smash/get/diva2:1825121/FULLTEXT01.pdf

by copyright or *sui generis* rights, contractual limitations create a general issue.[141] In the case of *Ryanair v PR Aviation,*[142] the CJEU clarified that if a database fails to meet the conditions for protection outlined in the Database Directive (Database Directive Art. 1(2)), the provisions regarding copyright or *sui generis* protection do not apply. This means that the rightholder of an unprotected database is not restricted by the directive and can limit its use through contracts,[143] as long as they comply with other provisions in each national law.[144] In fact, the applicability of the directive does not prevent the rightholder from imposing contractual limitations on the use of an unprotected database. Therefore, TDM exceptions simply do not apply if the terms and conditions of a website exclude TDM.

In summary, the Directive proposes an inventive formulation, by departing from the idea of creating exceptions to intellectual property rights and by introducing the idea that the owner of intellectual property may opt for a model resembling the widely known "*opt-out mechanism*".[145] When the opting-out from TDM is performed either at the same time with the decision to deny access under Article 3 of the DSM, or subsequently in the case of Article 4, there follows the determination of how to generate income from it. Licenses often seem to be the solution to this issue. It also noted that contractual frameworks that are specifically designed for the licensing of TDM or AI uses are also becoming more prevalent in practice.[146]

In fact, users of protected databases have more extensive possibilities to conduct TDM than users of unprotected databases, and authorisation from the rightholder is required for lawful access to mine such data.[147]

---

[141] Rossana Ducato and Alain Strowel, "Limitations to Text and Data Mining and Consumer Empowerment. Making the Case for a Right to "Machine Legibility" (19 March 2019), Kluwer Copyright Blog http://copyrightblog.kluweriplaw.com/2019/03/19/limitations-to-text-and-data-mining-and-consumerempowerment-making-the-case-for-a-right-to-machine-legibility/

[142] C-30/14, *Ryanair Ltd v PR Aviation BV*, par. 35

[143] C-30/14, *Ryanair Ltd v PR Aviation BV*, par. 39

[144] See note 13.

[145] A. Alemanno, G. Helleringer, Geneviève and A.-L. Sibony, «Brève introduction à l'analyse comportementale du droit», D. (2016), p. 911

[146] Schirru Luca, Margoni Thomas, "Arts 3 and 4 of the CDSM Directive as regulatory interfaces: shaping contractual practices in the commercial scientific publishing and stock images sectors" (22 August 2023) Kluwer Copyright Blog https://copyrightblog.kluweriplaw.com/2023/08/22/arts-3-and-4-of-the-cdsm-directive-as-regulatory-interfaces-shaping-contractual-practices-in-the-commercial-scientific-publishing-and-stock-images-sectors/

[147] See note 137.

4.4. Challenges surrounding technical measures

As previously analysed in subtitle 4.2., according to Recital 33 of InfoSoc Directive "lawful use" is broadly defined as any use authorised by the right holder or not restricted by law. The assessment of lawful use, particularly regarding copyright exceptions, hinges on the use of technological protection measures (**TPMs**) and contractual agreements.[148] Indeed, often, rightholders use TPMs to control access and usage of their digital works.[149] Similar to contractual means, the impact of TPM on TDM depends on the available exceptions. As previously analysed, rightholders have the option to opt-out from the general exception (CDSM Directive art. 4(3) and Recital 18) by explicitly reserving use through technological means such as metadata, paywalls, password control systems, time-limited access, encryption measures, etc.[150] For example, an online database may allow downloading until a predefined limit is reached, after which the user is technically prevented from downloading more. If the rightholder has reserved use through technical means, beneficiaries of the general exception cannot create a workaround code to bypass the restriction. The scientific research exception, in theory, may seem non-overridable by TPM, but in practice, it may provide with the same result (CDSM Directive Art. 3(3)). The EU legislator, by including the possibility for the opting-out mechanism seems to be trying to strengthen the rights of the rightholders.[151] On one hand, the reservation of rights could lead to an enhancement in the bargaining power of rightholders and could potentially result in licensing agreements with technology companies that involve remuneration.[152] Thus, this could also result to the increased market concentration and exploitation of creators. This situation is already noticed in some sectors, and it already makes artists to waive their "training rights" in exchange of reduced compensation.[153] Overall, while the Recitals of the CDSM Directive clarify the purpose and

[148] See note 121.
[149] Rosati Eleonora - Adrian Aronsson-Storrier, "Contractual override and the new exceptions in the Copyright in the Digital Single Market Proposal" (2018).
[150] Nobre Teresaand Mileszyk Natalia "Implementing the new EU protections against contractual and technological overrides of copyright exceptions" (9 December 2019), Communia https://communia-association.org/2019/12/09/implementing-new-eu-protections-contractual-technological-overrides-copyright-exceptions/ [last access 1 April 2024].
For a detailed analysis, please read: Teresa Nobre and Natalia Mileszyk , 'Article 7: Contractual and Technological Override' (Guidelines for the Implementation of the CDSM Directive) https://www.notion.so/Article-7-Contractual-and-technologicaloverride-7f20f72c9aec484194067946c9dbd43f [last access 1 April 2024].
[151] See note 100.
[152] Quintais J, "Generative AI, Copyright and the AI Act" (09 May 2023), Kluwer Copyright Blog. Available at https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/ [last access 18 July 2024].
[153] Ibid.

limitations of TPM, there are serious concerns that the strong protection of TPM under the InfoSoc Directive may provide greater leeway for their application than initially perceived in the CDSM Directive.[154]

At this point, it is necessary to examine the protection provided in relation to the technological measures in copyright law. Article 6(1) of the InfoSoc Directive prohibits the circumvention of technological measures. It should be highlighted that this protection extends not only to measures designed to prevent copyright infringement, but also to any unauthorised use, regardless of whether it constitutes a copyright-relevant act or not.[155] Actually, this means that the rights of the copyright holder are limited by what is technologically possible, rather than what is prescribed by copyright law. This is sometimes referred to as "*paracopyright*".[156] Consequently, the extensive prohibition of circumvention poses difficulties to users, both in terms of non-restricted uses, such as enjoying the work, and benefiting from copyright exceptions in general.[157] In the case of TDM, the application of technological measures to ensure security and integrity can create a barrier with potentially significant consequences.[158]

Furthermore, it should be highlighted that Article 6(4), sub-paragraph 1 of the InfoSoc Directive distinguishes the relationship between circumventing technological measures and copyright exceptions of the article 5.[159] In case that the copyright holder has not implemented voluntary measures to allow for exceptions, then the Member State must ensure that the rightholder does so. The provision aims to ensure effective application of exceptions once legal access is obtained.[160] An apt illustration is a teacher unable to make copies from dictionary on a locked

---

[154] Samuelson Pamela, "The EU's Controversial Digital Single Market Directive - Part II: Why the Proposed Mandatory Text- and Data-Mining Exception Is Too Restrictive" (12 July 2018), Kluwer Copyright Blog http://copyrightblog.kluweriplaw.com/2018/07/12/eus-controversial-digital-single-marketdirective-part-ii-proposed-mandatory-text-data-mining-exception-restrictive/ [last access 1 April 2024].

[155]Dusollier Séverine, "Tipping the Scale in Favor of the Right Holders: The European Anti Circumvention Provisions" (2003), Eberhard Becker and others (eds), Digital Rights Management: Technological, Economic, Legal and Political Aspects, Springer 2003, 465–466. https://doi.org/10.1007/10941270_29, also see note 163, page 253.

[156] See note 137.

[157] See note 47.

[158] See note 110.

[159] Casellati Alvise Maria, "The Evolution of Article 6.4 of the European Information Society Copyright Directive" (2001), 24 Columbia - VLA Journal of Law & the Arts 369, 374.

[160] Dusollier Séverine, "The Protection of Technological Measures: Much Ado About Nothing or Silent Remodeling of Copyright?" (2014).
Rochelle Cooper Dreyfuss and Jane C Ginsburg (eds), Intellectual Property at the Edge (Cambridge University Press). Available at: https://www.cambridge.org/core/books/abs/intellectual-property-at-the-edge/protection-of-technological-measures-much-ado-about-nothing-or-silent-remodeling-of-copyright/C56E32086F907C9D97334B8E249B2EC0

CD-ROM for educational purposes.[161] The initial intention was to make TDM exceptions non-overridable by TPM, but at the end they were simply added to the list of exceptions guaranteed by Member States.[162] So far, some states use traditional court procedures, while others require user to file complaints to relevant authorities.[163] The UK's relevant authority has received few requests, and it has already been argued that users may believe that they cannot exercise their rights or that the process is uncertain.[164]

Given the above, it seems that through Article 6(4) sub-paragraph 1, the EU legislator considers that the best solution for the rightholders is to voluntarily facilitate the practice of exceptions to their rights. However, this system creates an anti-theft alarm that disregards user intentions and applicable exceptions. It does not require rightholders to remove TPMs, instead forcing users to contact them or go through a lengthy process for each work they want to use.[165] Upon the results of a survey, it was found that it takes approximately a month for the users to access blocked by TPMs content, while it was also reported that some never manage to receive access despite contacting the rightholder. Additionally, it resulted that rightholders often respond to mining by imposing sanctions and technical obstacles. This actually leads scientists to circumvent TPMs or use platforms like SciHub, a website connected with lawsuits for copyright violations.[166] Member States should be aware that these actions send negative signals to the research community. As a result, restrictive implementation of TPMs by Member States may force scientists to resort to unauthorised methods to access the data they need for research, and the current legal provisions may not adequately protect them.

Member States' solutions to ensure exceptions following Article 6(4) of the InfoSoc are considered as disproportionate, deficient, and limited by technological barriers.[167] Beneficiaries lack authority to circumvent illegal TPMs, putting them in a weak position.[168] The provision seems to

[161] See note 158.
[162] See note 135.
[163] Gasser Urs, "Legal Frameworks and Technological Protection of Digital Content: Moving Forward Towards a Best Practice Model" (2006), Berkman Klein Center for Internet & Society https://papers.ssrn.com/abstract=908998
[164] Gillen Martina and Sutter Gavin, "DRMS and Anti-Circumvention: Tipping the Scales of the Copyright Bargain?" (2006) 20 International Review of Law, Computers & Technology 287, 291.
[165] Keller Paul, Research Librarians: New TDM exception can be undermined by technical blocking from publishers" (10 March 2020), Communia. https://communia-association.org/2020/03/10/reserach-librarians-new-tdm-exception-can-undermined-technical-blocking-publishers/
[166] LIBER Europe, "Europe's TDM Exception for Research: Will It Be Undermined By Technical Blocking From Publishers?" (10 March 2020) https://libereurope.eu/article/tdm-technical-protection-measures/
[167] See note 153.
[168] See note 133, page 30.

turn exceptions into negotiations and contractual relationships with rightholders.[169] If the procedures for exceptions are burdensome, they become inefficient. Ineffective rights regimes are worse than no rights at all.[170] Based on empirical research, it is evident that individuals who could potentially benefit from exceptions feel limited by the prohibition. Additionally, there are indications that rightholders are transitioning from a strategy of complete prevention to one of moderation and control. This shift involves implementing advanced TPMs that enable users to utilise the work within certain boundaries, such as restricting the number of reproductions for purchased music files.[171] However, the legal aspect of this issue remains unsolved, leading to the concern about the outcome of influencing user behavior in a manner that contradicts the fundamental principles of copyright, such as the free flow of information and ideas.

In light of the above and in relation to scientific research, the CDSM Directive introduces a new provision where Member States are encouraged to promote the establishment of commonly agreed best practices between rightholders and research institutions regarding storage and technical measures (CDSM Directive Art. 3(4)). This resembles the approach taken for extensive technical measures (of InfoSoc Directive Art. 6(4) 1st sub-para), allowing rightholders to voluntarily implement measures before state intervention. Member States have the obligation to ensure that rightholders enable mining in line with the exception in case that voluntary measures are not implemented. The ability of rightholders to apply TPMs is limited to protecting platform security and performance, with Member States having the flexibility to determine the level of such security and implementation. The encouragement of "best practices" can be a one-time effort or subject to continuous revision under national law. Assessing the efficiency of the exception at this stage is challenging. The implementation of the new exceptions must adhere to the three-step test and maintain a fair balance between rightholders and users (CDSM Directive Recital 6). However, the best practice solution has already faced criticism for not fully addressing the issue of rightholders preventing or restricting TDM through technological measures.[172] It will be interesting to observe the approaches taken by Member States as they seek to achieve a harmonious equilibrium in this matter.

[169] See note 158.
[170] Iain Hargreaves, "Digital Opportunity - Review of Intellectual Property and Growth" (2011), Department for Business, Innovation & Skills, Independent Report 11/968 47 page 5 https://www.gov.uk/government/publications/digital-opportunity-review-of-intellectual-property-and-growth
[171] See note 163, page 182.
[172] See note 110.

In regard to the general exception, it could be argued that it was drafted for the benefit of the rightholder considering the possibility to opt-out using technical means and prohibiting circumvention. However, the provision related to the best practices does not apply to the general exception, but it is included in the list of exceptions that Member States should ensure access to when blocked by technological measures (InfoSoc art. 6(4) sub-para 1 through CDSM Directive art. 7(2)). This situation is comparable to the one described in InfoSoc Directive Article 6(4) sub-paragraph 4, which has been criticised as a major weakness in the provisions of InfoSoc regarding the circumvention of technical measures (InfoSoc Art. 6(4) 1st sub-para).

However, this benefit for the rightholders does not extend to works made available to the public through agreed contractual terms, such as subscription services (e.g. Westlaw) and streaming services such as Spotify.[173] As the digital industry adopts a library-like business model, the power of rightsholders increases, favoring their interests.[174] This rule prevents further development of accessed content without the rightsholder's consent, hindering the flow of information and contradicting the purpose of copyright as a means for intellectual and creative exchange in the public. This provision definitely stipulates freedom of contract over copyright law.[175]

Fortunately, InfoSoc Directive Article 6(4), sub-paragraph 4 does not apply to any of the TDM exceptions (as per CDSM Directive 7(2)) and Member States have the authority to take appropriate measures for making the TDM exceptions available, including even for "works and other subject matter made available to the public through on-demand services" (CDSM Directive Recital 7). This development seems to be beneficial for the scientific research exception as well, but it still allows rightsholders to maintain control over their works through technological means. This means that beneficiaries who want to use material available on-demand or through streaming will still face restrictions, similar to the effects of InfoSoc Article 6 sub-paragraph 4.[176] This could hinder the efficient application of TDM, as online access is a common way to access sources today. Furthermore, rightholders now can technologically protect any type of source from TDM.

To summarise the aforesaid, the following points should be highlighted. While the rightholders cannot restrict beneficiaries of the scientific research exception, they are allowed to use technical

[173] See note 167.
[174] See note 158.
[175] See note 162, page 288.
[176] See note 153.

means to protect the functionality of their services. The prohibition of circumventing technical measures should not impede users from accessing content when the rightholder has implemented measures that go beyond what is required. However, unless Member States make significant changes during the implementation framework of the best practices provision, it is likely that users will still face challenges and TPMs will continue to pose obstacles for TDM beyond the scope of copyright law.[177] This is particularly relevant considering the emergence of newer control techniques like blockchain technology, which further hinder the situation.

The rightholder may utilise technical measures to restrict access for individuals who are not researchers but intend to perform TDM, creating a situation similar to the one outlined in InfoSoc Directive Article 6(4) sub-paragraph 1. While Article 6(1) sub-paragraph 1 of the InfoSoc is applicable, it is challenging to determine when Member States actually have the opportunity to intervene and assist beneficiaries of the general exception in exercising their rights, since rightholders can reserve usage through the same technical means. This raises concerns about the effectiveness of the exception when the need for authorisation cannot be bypassed.[178] Those who are unable to rely on the scientific research exception are compelled to negotiate contractual agreements with the rightsholder, without a satisfactory justification for this requirement. Start-ups, journalists, and information intermediaries possess the same potential as researchers affiliated with research organisations[179] (M Planck, 2017). Moreover, this chosen approach is more restrictive compared to jurisdictions with more open clauses, potentially diminishing the competitiveness of EU Member States.[180] It cannot be denied that a clearer solution would have been to explicitly state that TPMs cannot override either of the CDSM exceptions and provide users with effective means to remove them.[181]

Lastly, as explicitly elaborated above, the TDM process involves the analysis and creation of a dataset. For the purposes of this thesis, these procedures were considered and analysed in Chapter 2 as a whole. Apart from the TDM exceptions introduced through CDSM Directive, it is worth noting that the mandatory exception for temporary reproductions as per Art. 5(1) of the InfoSoc Directive still applies to TDM (CDSM Directive Recitals 9 and 18). The exception may be

[177] See note 144, page 19.
[178] See note 163.
[179] See note 106, page 3–4.
[180] See note 133, 30-35.
[181] See note 133, page 21-22 (184).

advantageous for the analysis step of the TDM process when employing a technique that automatically deletes copies from the computer's RAM memory. In this case, the creation of these copies becomes an integral and essential part of the technological process, allowing the analysis step to benefit from the exception. Nevertheless, the remaining stages of the TDM process fail to fulfill the criteria of the temporary reproduction exception and this is because they involve manual work which cannot be easily eliminated. As a result, this exception cannot serve as a defense for the entirety of the TDM process.[182] Furthermore, TDM is not applicable to databases (as per InfoSoc art. 1(2)(e)) as well as no corresponding exception, indicating the absence of a temporary reproduction exception that is applicable to databases, is proposed by the Database Directive either.[183]

For the above reasons, the temporary reproduction exception may be considered to have limited effectiveness and to be providing little legal certainty for TDM and therefore it is necessary to refer to the CDSM Directive for further guidance.[184]

5. AI ACT AND COPYRIGHT COMPLIANCE

5.1. High level analysis of AI Act

In December 2023, the European Commission (**EC**) announced the ground-breaking provisional agreement on what is claimed to be the world's first-ever comprehensive legal framework on AI, the European Union Artificial Intelligence Act (**AI Act**).[185] This agreement is the outcome of a continuous effort that began in April 2021, with the publishment of the proposal for the Act by the EU Commission.[186] The majority of the obligations are anticipated to become effective in the beginning of 2026. Nevertheless, it is noteworthy that prohibited AI will need to be gradually

[182] See note 137, page 15-22.
[183] See note 137.
[184] See note 133, page 30-35.
[185] Council of the European Union, "Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world" (3 December 2023), Press release, updated 2 February 2024. Available at https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/
[186] Harrington Madelaine, Hansen Marty, Peets Lisa, Drake Marianna & Young Mark, "EU Parliament Adopts AI Act" (15 March 2024), Covington https://www.insideglobaltech.com/2024/03/15/eu-parliament-adopts-ai-act/

discontinued six months after the AI Act is enacted. The regulations governing general-purpose AI are expected to be applicable in early 2025.[187]

Overall, the focus of the EU is to ensure that the use of AI systems is safe, secure, transparent, fair and environmentally friendly. The rules set outline responsibilities for both providers and users based on the level of risk associated with AI.[188] Among the topics included in the scope of the Act are provisions related to general-purpose AI (**GPAI**) systems and models. Such regulations were proposed in order to address various concerns raised by GenAI models such as Chat GPT4 and Midjourney.[189] The said provisions raised a lot of discussions due to the disagreement between the Council and the Parliament until the very last moment, however the final compromise included two provisions related to copyright.

More specifically, as per the final text published on 12 July 2024 on the Official Journal of the European Union,[190] in section 2 "*Obligations for providers of general-purpose AI models*" were introduced, with two requirements concerning copyright law. Article 53.1(c) stipulates that providers of GPAI models shall establish a policy to comply with EU copyright law, in order to identify and respect the preservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790. Following, according to section 53.1(d) the providers are also required to create and then publicly share a detailed summary of the content used for the training of general-purpose AI models, based on a template provided by the AI Office. This provision is also accompanied by the relevant Recital 107 according to which this summary shall be "*sufficiently detailed*".

Although the former provision is quite recent, the latter can be traced as a requirement back in the European Parliament's report from June 2023. This report included a suggested provision that

[187] European Commission, Artificial Intelligence — Questions and Answers (December 12, 2023). [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683

[188] European Parliament, "EU AI Act: first regulation on artificial intelligence" (last updated 18 June 2024, European Parliament https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[189] Midjourney is an example of generative AI that can convert natural language prompts into images. With Midjourney, you can create high-quality images from simple text-based prompts.
See Wankhede Calvin, "What is Midjourney AI and how does it work?" (March 6, 2024). Available at: https://www.androidauthority.com/what-is-midjourney-3324590/ [last access 7 July 2024].

[190] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) Text with EEA relevance. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

would oblige model providers to publicly share a detailed summary of their use of copyrighted training material. The finally adopted version is a significant enhancement over the preliminary text of the Parliament, since it no longer differentiates the model providers between the ones using protected training material and the ones using public domain content in order to apply different transparency standards to each provider.[191] Such a measure would have not been practical and indeed after its publishment it has been negatively discussed.[192] The current text provides more clarity, with the Recital stating that the "*summary should be comprehensive in its scope instead of technically detailed, for example by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used*". In addition, it also pays attention to the template to be provided by AI office, which should "*allow the provider to provide the required summary in narrative form".*

This narrative report seems reasonable and in fact it follows the already established practice for open-source AI models. The intention is to focus on larger commercial model providers who have already show reluctance in providing substantial explanations regarding their training data.[193]

In addition, this measure aims at ensuring the collaboration between copyright holders and providers of GenAI systems, in relation with this new method of exploiting work.[194] Through the publication of high-level descriptions, third parties can evaluate whether the model providers have used data sources that are legally accessible and follow the regulations set forth in Article 4(1) of the CDSM Directive, therefore comply with the EU copyright rules.

---

[191] Artificial Intelligence Act Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))1
https://artificialintelligenceact.eu/wp-content/uploads/2023/06/AIA-%E2%80%93-IMCO-LIBE-Draft-Compromise-Amendments-14-June-2023.pdf

[192] Committee on Internal Market and Consumer Protection (IMCO), "Activity Report 2019-2024", https://www.europarl.europa.eu/cmsdata/283430/IMCO%20Activity%20Report%20-%202019-2024.pdf

[193] See Gemini Team, Google, "Gemini: A Family of Highly Capable Multimodal Models" Google Deep Mind. Available at: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf
See also GPT-4 System Card, OpenAI, (March 23, 2023) https://cdn.openai.com/papers/gpt-4-system-card.pdf

[194] Geiger Christophe, Iaia Vincenzo, (Luiss Guido Carli University) "Generative AI, Digital Constitutionalism and Copyright: Towards a Statutory Remuneration Right grounded in Fundamental Rights—Part 1" (17 October 2023), Kluwer Copyright Blog. Available at https://copyrightblog.kluweriplaw.com/2023/10/17/generative-ai-digital-constitutionalism-and-copyright-towards-a-statutory-remuneration-right-grounded-in-fundamental-rights-part-1/ [last access 19 July 2024].

The EU legislator has indeed considered the impact of the transparency obligation and clarified also that the size of the company who acts as a system provider shall be taken into account. As a result, more simple compliance methods will be required for small or medium sized enterprises and start-up companies as the objective is to omit any procedure that could impede the capacity of such entities to develop AI systems.[195]

Finally, although the TDM exceptions are not covered specifically through the AI Act, some reference to the copyright law is also noticed through Recital 48 whereby it is explained that AI systems may negatively affect the fundamental rights protected from the Charter, including amongst others, the intellectual property rights. In fact, the Recital highlights the importance of the safeguarding of such rights.[196]

5.2. Implications of the AI Act and CDSM Directive on AI Model Training

As already mentioned, the compliance policy provision set in Article 53.1(c) is directly connected to the TDM exception set in Article 4(1) of the CDSM Directive showing that the aim of the legislator of AI Act is quite clear. In addition, there is mention related to the above in the Recital and more specifically it is underlined that "*any use of copyright protected content requires the authorisation of the rightsholder concerned unless relevant copyright exceptions and limitations apply*" (AI Act, Recital 105) and then it continues with reference to Article 4(3) of the CDSM Directive, that "*where the rights to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorisation from rightholders if they want to carry out text and data mining over such works*."

To continue with, the following Recital i.e. 106, not only restates the existing copyright rules but also includes language that aims to prevent discussions about the territorial application of the EU's TDM rules. In other words, the Recitals are designed to address and potentially limit debates about where the EU's TDM rules should apply.[197] In particular, the Recital states that "*any provider*

[195] AI Act, Recital 8 and 143
[196] AI Act, Recital 48
[197] Keller Paul, "A first look at the copyright relevant parts in the final AI Act compromise" (Monday, December 11th, 2023), Kluwer Copyright Blog, (Institute for Information Law (IViR))

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

As per P. Keller, the above Recital could be interpreted as an attempt to expand the scope of the legislation beyond copyright rules. However, considering the nature of AI model training, where training occurs in a different context from the actual use of the resulting models, the passage argues that such intervention may be necessary, despite of the uncertainty whether such a requirement can be legally feasible or effective. Furthermore, the passage highlights the pressing need for international alignment in copyright law regarding the use of copyrighted works for AI model training. It suggests that the EU, through the AI Act, is asserting its role as a global rule maker in this field.[198]

To conclude, the new provisions adopted in the AI Act, along with the above-mentioned Recitals, build positively upon the legislative approach previously adopted by the EU in the 2019 CDSM Directive. On one hand, Section 2, Article 53.1(c) of the AI Act is highlighted as providing additional clarity for both rightholders and AI model providers. However, this will only be truly beneficial if a generally accepted standard emerges in this area. Therefore, there is emphasis in the importance of policymakers and other stakeholders focusing on the development of such a standard as the AI Act progresses towards becoming law. Furthermore, Section 2, Article 1(d) of the AI Act, which requires the publication of sufficiently detailed summaries of the content used for training, may eventually make it easier for third parties to understand the sources of training data, including whether the lawful access criterion has been met.[199]

Neither the term AI nor GenAI are mentioned specifically in the CDSM and in the absence of any relevant court decision some believe that the exception of Article 4 of CDSM specifically covers

https://copyrightblog.kluweriplaw.com/2023/12/11/a-first-look-at-the-copyright-relevant-parts-in-the-final-ai-act-compromise/

[198] Ibid.

[199] See note 24.

the systematic and extensive use of creators' protected works and performances, with the intention of generating synthetic content for commercial use, while others state that this interpretation seems unfair.[200]

According to Emanuilov and Margoni, the AI Act proposal makes clear that the TDM exceptions contained in the CDSM Directive apply to the development and training of GenAI models, allowing the creation of necessary copies made during their creation under the relevant Articles 3 and 4 of the Directive.[201] The CDSM Directive seems to assume that these copies occur during preparation and aren't present in the final model, thus not addressing memorization.[202] The authors also assert that Articles 2-4 of CDSM Directive are broad enough to potentially cover permanent copies in the model, which means that any memorization is excluded. Nevertheless, models with copyrighted training data cannot be publicly shared because Articles 3 and 4 of the said Directive only permit reproductions and adaptations, not public communications. In addition, the same seems to apply to the outputs generated by GenAI applications and the inevitable conclusion is that if the model produces outputs, such as an answer, that resemble portions of the training material, these outputs cannot be shared with the public without violating copyright laws, since the output simply reflects the input, the tool was fed with.

Unfortunately, the conflict between copyright and the utilization of protected material by AI systems remains unresolved by the EU AI Act, as it merely makes reference to the sections of the EU Copyright Directive addressing the copyright exception for text and data mining.[203] It is believed that a solution to this conflict could be considered the Article 43 of the Data Act,[204] which defines the scope of the *sui generis* right under the Database Directive by excluding certain types of databases from this protection. Consequently, the last chapter of this thesis will examine how effectively Article 43 of the regulation can address this situation.

---

[200] ECSA, Joint Statement on Generative Artificial Intelligence and the EU AI Act (25 April 2024) https://composeralliance.org/news/2024/4/joint-statement-on-generative-artifical-intelligence-and-the-eu-ai-act/

[201] Emanuilov Ivo, Margoni Thomas (Ku Leuven Centre for IT & IP Law), "Memorisation in generative models and EU copyright law: an interdisciplinary view" (March 26t 2024), Kluwer Copyright blog https://copyrightblog.kluweriplaw.com/2024/03/26/memorisation-in-generative-models-and-eu-copyright-law-an-interdisciplinary-view/

[202] Ibid.

[203] Coraggio Giulio, "AI Act – What Is the Scope of the TDM Copyright Exception?" (April 2 2024) https://www.linkedin.com/pulse/ai-act-what-scope-tdm-copyright-exception-giulio-coraggio-yawcf/

[204] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act).

6. DATA ACT AS AN ALTERNATIVE SOLUTION TO THE TDM REGULATION

6.1. High level analysis of the Data Act

In the recent years, the global market has noticed a swift increase in the availability of internet connected products, constituting as above-mentioned the network of IoT. Their existence expands tremendously the amount of data available for reuse within the EU resulting in a huge potential for innovation and competitiveness in the region.[205] At this point, the European Commission introduced the implementation of an act that will promote the flow and use of data by making them more accessible and usable, to enhance the EU data's economy.[206]

More specifically, the Regulation on harmonised rules on fair access to and use of data — also known as the **Data Act** — entered into force on 11 January 2024 and will become applicable in September 2025. Alongside the CDSM, DMA[207] and DSA,[208] and the recently introduced AI Act, it constitutes a key element of the European data strategy, and its main goal is to release in the market large amounts of digital data held by a few entities, ensuring fair access and use.[209] In order to unlock these, it is required that the data access for consumers and businesses is improved by clarifying the Database Directive,[210] also by allowing public sector to make use of data held by enterprises in specific situations as well as by facilitating switching between cloud and edge services,[211] establishing safeguards against unauthorised data transfers by cloud providers and by developing interoperability standards for data.[212]

The Data Act grants users with the right to access and the right to share data generated by products or related services[213] as well as it obliges private companies to confer data to public

[205] European Commission, "Data Act Explained" (2024) https://digital-strategy.ec.europa.eu/en/factpages/data-act-explained
[206] CMS Law Now, "An overview of the Data Act, (Germany, 15 January 2024) https://cms-lawnow.com/en/ealerts/2024/01/an-overview-of-the-data-act
[207] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance) [2022] OJ L 265, 1-66.
[208] 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) (Text with EEA relevance) [2022] OJ L 277, 1-102.
[209] Derclaye Estelle and Husovec Martin, "Why the sui generis database clause in the Data Act is counter-productive and how to improve it?" (2022). Available at SSRN: https://ssrn.com/abstract=4052390
[210] Data Act, Chapters I-IV, and X.
[211] Proposed Data Act, Chapter VI.
[212] Proposed Data Act, Chapter VIII.
[213] Data Act, Art 5. and Chapter IV.

bodies for exceptional needs, including measures to prevent a public emergency.[214] This regulation places a particular focus on prioritizing the expansion of access to, and utilization of, data collected by advanced sensors and machines, including those employing cutting-edge technologies such as the internet of things (IoT).[215] Moreover, it revises the Database Directive to stay relevant with the current data-driven society and ensure data accessibility within the EU.[216]

Overall, the main purpose of this act is to eliminate legal uncertainties about the *sui generis* right to machine-generated data, specifically through Art. 43, which excludes the protection of databases containing data obtained from or generated by the use of a connected product or a related service.[217] According to the Data Act a "connected product" is described as *an item that obtains, generates or collects data and that is able to communicate product data ..and whose primary function is not the storing, processing or transmission of data on behalf of any party other than the user,*[218] while a "related service" refers to a digital service such as software *which is connected with the product in such a way that its absence would prevent the connected product from performing one or more of its functions, or which is subsequently connected to the product by the manufacturer or a third party to add to, update or adapt the functions of the connected product.*[219] However, only Recital 112 is connected with the Article 43 of the Data Act which further clarifies that the *sui generis* right does not apply to the above-mentioned databases, as the requirements for protection would not be fulfilled.[220]

Hence, it seems that this act intents, amongst others, to specify what is not covered by the *sui generis* regime without explicitly amending the Database Directive.[221] It is obvious that Article 43 of the Data Act positively intends to reduce excessive IP protection on certain databases, in order to encourage innovation and research within the legal framework of the CDSM.[222]

---

[214] Data Act, Art 14 and 15.
[215] Colangelo Giuseppe, "European Proposal for a Data Act-A First Assessment" (2022) Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4199565 [last access 19 July 2024].
[216] ibid.
[217] Data Act, Art. 43.
[218] Data Act, Art. 2(5).
[219] Data Act, Art. 2(6).
[220] Data Act, Recital 112.
[221] See note 212, page 1-2.
[222] European Commission, Commission Staff Working Document, "Impact Assessment Report' Accompanying the document Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act)" (2022) SWD, 34 final, 16, 70 and 139. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022SC0034

6.2. TDM research and Art. 43

As aforesaid, the sole Article in the Data Act related with TDM research seems to be very beneficial for professionals conducting TDM research. For example, researchers who use ChatGPT to undergo TDM research on medical databases, might process different types of databases comprised of data generated by humans,[223] machines,[224] or a combination of both.[225] Thus, Article 43 establishes a legal safeguard for individuals requiring access to, utilization of, or sharing of databases comprised of data generated through the utilization of a product or its related service (IoT data),[226] by allowing them to have a free access to analyse databases which consist of data generated by, for instance, wearable devices for healthcare monitoring such as fitness trackers, smartwatches, blood glucose monitors, smart technology clothing and others. Such an analysis could bring substantial information for the development of new treatments and the improvement of diagnosis or even the disease prevention.[227]

However, although the databases composed of machine-generated data are excluded from the *sui generis* database protection under the provision 43, to ensure that users retain the rights to access, use, and share such data as outlined in Articles 4 and 5 of the Data Act, the key question is how strict this exclusion is and/or if only applies when it affects the defined users' rights. Recital 112 of the Data Act may provide some clarification, by stating that databases created from data collected by connected products should not fall under the protection of Article 7 of the Database Directive as they do not meet the *sui generis* right criteria, hence they would not satisfy the necessary protection requirements. The interpretation of Recital 112 (former Recital 84 of the proposal for Data Act)[228] has led to scholars' debate, with some of them to believe that it is implied that databases with machine-generated data could be protected under certain conditions, while

[223] For instance, an electronic health record is a database containing patients' health information (e.g., diagnoses, allergies, prescribed medication etc,), which is created and maintained by medical personnel.
[224] For instance, wearable health devices (e.g., fitness trackers, smartwatches, smart inhalers, surgical robots, and others).
[225] For instance, biobanks are collections of biological samples, which may consist of human-generated data (e.g., patient's medical history) and machine-generated data (e.g., genetic sequencing).
[226] Data Act, Art. 4, 5 and 43.
[227] De Michele Roberta and Furini Marco, "IoT Healthcare: Benefits, Issues and Challenges" (2019), GoodTechs '19: EAI International Conference on Smart Objects and Technologies for Social Good. Available at: https://dl.acm.org/doi/10.1145/3342428.3342693
[228] Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act), Brussels [2022] COM (2022) 68 final (proposed Data Act).

the opposite party claims that such databases were never meant to be protected under the *sui generis* regime.[229] Nonetheless, both views fail to fully clarify the scope of the *sui generis* right for databases in Article 43 of the legislation under discussion governing data.

More specifically, it was stated that the proposed Data Act and consequently the final text as it was published by the EU legislator, indicates that merely operating an automated process that continuously generates data does not qualify for *sui generis database* rights.[230] In fact, the *sui generis* database right protects the creator of a database who has made a substantial investment from a qualitative or quantitative perspective, [231] However, to better understand in practise, the following example is provided. What would happen if the installation of sensors or software in wearable health monitoring devices requires a substantial investment? Could this be seen as an investment in acquiring data rather than creating it, and thus be eligible for protection under the *sui generis* database regime? The above is related with the "obtaining-creating dichotomy" of the CJEU's legal test regarding the scope of the *sui generis* right.[232] The CJEU interprets "obtaining" as the use of resources to find and collect existing materials for a database, not to create new data.[233] Distinguishing between "collecting" and "creating" data, especially in machine-generated

[229] See for instance Senftleben Martin, "Study on EU Copyright and Related Rights and Access to and Reuse of Data" 2022, European Commission. Research and Innovation. Independent Expert Report 2022) 50. Available at: https://pure.uva.nl/ws/files/85957820/KI0822205ENN.en.pdf%20p.50

[230] Rossana Ducato and Strowel, Alain M., "Ensuring Text and Data Mining: Remaining Issues With the EU Copyright Exceptions and Possible Ways Out" (2021), European Intellectual Property Review, 322, 333. Available at SSRN: https://ssrn.com/abstract=3829858

[231] See Database Directive, art 7 (1);
See also Kuschel Linda and Dolling Jasmin, "Access to Research Data and EU Copyright Law" (2022), 13 JIPITEC 247. Available at https://www.jipitec.eu/archive/issues/jipitec-13-3-2022/5558

[232] Guido Noto La Diega and Derclaye Estelle, "Opening Up Big Data for Sustainability: What Role for Database Rights in the Fourth Industrial Revolution?" (2022), Ole-Andreas Rognstad, Taina Pihlajarinne and Jukka Mähönen (eds), Promoting Sustainable Innovation and the Circular Economy: Legal and Economic Aspects, 26. Available at SSRN: https://ssrn.com/abstract=4220534
See also Europäische Kommission Generaldirektion Kommunikationsnetze, Inhalte und Technologien Copyright, Norbert Maier, Federico De Michiel, Viola Peter, María del Carmen Calatrava Moreno, Chiara Pancotti, Francesca Monaco, Maurice Shellekens, Inge Graef, Nadya Purtova, Andreas Wiebe, European Commission. Directorate General for Communications Networks, Content and Technology, Technopolis, Centro studi industria leggera, Tilburg University, "Study to Support an Impact Assessment for the Review of the Database Directive Final Report" (2022), Publications Office of the European Union, 2022,1.
See *Fixtures Marketing Ltd v Svenska Spel AB, Fixtures Marketing Ltd v Oy Veikkaus AB and The British Horseracing Board Ltd and Others v William Hill Organization Ltd*.
See also note 46 and 220.

[233] See Case C-444/02 *Fixtures Marketing Ltd v Organismos Prognostikon Agnon Podosfairou* ECLI:EU:C:2004:697, para 44; Case C-338/02 *Fixtures Marketing Ltd v Svenska Spel AB* ECLI:EU:C:2004:696, para 28; Case C-46/02 *Fixtures Marketing Ltd v Oy Veikkaus Ab* ECLI:EU:C:2004:694, para 38; Database Directive recitals 7, 39 and 40; Case C-203/02 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* ECLI:EU:C:2004:695, paras 31-49.

contexts, can be complex.[234] For example, data generated from wearable health devices should be considered as "collected" since it already exists. Researchers who record measurements like sugar levels or heart rate are not creating the data itself but are documenting it using the devices.[235]

To finalise, this legislation aims to make machine-generated data more accessible to users, business, and trade professionals, and, in exceptional cases, public sector bodies.[236] Consequently, researchers may benefit from provisions that permit users to share machine-generated data with third parties, as 'third parties' refers to research organisations or non-profit organisations.[237] Data holders are required to provide this data to these beneficiaries promptly, without undue delay, free of charge to the user, and with the same quality as is available to the data holder.[238] In addition, researchers may be able to rely on Article 14 of the Data Act, which requires data holders to provide machine-generated data to public sector bodies in cases of exceptional need. Recital 63 further explains that "*research-performing organizations and research-funding organizations can also be classified as public sector bodies or entities governed by public law*". As a result, publicly funded medical research institutions may benefit from this provision when conducting TDM on databases containing data generated by health monitoring devices. Furthermore, these institutions are allowed to share this data with *'individuals or organizations for the purpose of conducting scientific research*,' provided that these entities operate on a non-profit basis or within the context of a public-interest mission recognized by the State.[239] But, at the same time this indicates that independent individual researchers and private research institutions would be unable to gain even indirect access to databases composed of machine-generated data through collaboration with public entities. Also, it should be highlighted that the provisions in the Data Act that permit public bodies to access and share such data with other parties would only apply in cases of exceptional need, such as public emergency situations.[240]

---

[234] Leistner Matthias and Antoine Lucie, "IPR and the Use of Open Data and Data Sharing Initiatives by Public and Private Actors" (2022), Study commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs at the request of the Committee on Legal Affairs) 50 Available at: https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2022)732266 [last access 7 July 2024]; See also note 245, page 36; See note 235, page 15;
Case C-202/12 *Innoweb BV v Wegener ICT Media BV and Wegener Mediaventions BV* ECLI:EU:C:2013:850,para 39.
[235] ibid, paras 33-34, 38.
[236] Data Act, Art. 1.
[237] Data Act, Art. 4 & 5.
[238] Data Act, Recital 57 & 69.
[239] Data Act, Recital 76.
[240] Data Act, Art. 14 & 15 & 21.

To summarise, for those who conduct research using TDM on databases, the sui generis database infringement under the TDM exceptions of the CDSM Directive may not seem as advantageous as initially anticipated. The TDM exceptions of the CDSM Directive are drafted in such way to demonstrate that copyright and database protections could indeed hinder research and innovation.[241] For this reason, researchers might try to approach the topic form a different perspective to circumvent potential infringements. At first glance, Article 43 of the Data Act appears as a good alternative solution.[242] However, after extensive analysis, it becomes evident that the Data Act does not adequately address the needs of researchers either, and as a result it seems insufficient as regards strong guarantees for access and use of data for scientific research.[243] The data access provisions outlined in Articles 14 and 21(1) of the Data Act are limited in scope, offering minimal benefits to researchers using TDM tools for scientific inquiry. Given this context, Article 43 should be revised to allow for more efficient and comprehensive access to and use of machine-generated raw data collections for the scope of scientific research. It is essential to include researchers as beneficiaries of this provision and to consider their specific needs within the framework of the proposed regulation, as this would support data-driven research activities and promote a research-friendly environment at a time when the importance of data is continually increasing.

[241] See note 233, page 351.
[242] See note 225, page 6.
[243] See note 212, page 2.

CONCLUSION

The legislative copyright landscape of EU was insufficient as it was not synced with the digital evolvement. The EU legislator proceeded with the drafting of EU Directive 2019/790 on Copyright and related rights in the Digital Single Market as part of a broader effort to modernise copyright provisions in line with the current digital advancements. The main aim is to protect the creators of content through placing various obligations on the online platforms that profit from the sharing of copyrighted material and cover the legal gaps that have been created by the time of the publication of the CDSM Directive.[244]

This thesis has focused on the TDM exceptions which were introduced through Articles 3 and 4 on the CDSM Directive and whose objective is to balance the needs of research organisations, cultural institutions, and other entities with the rights of copyright holders. Overall, TDM techniques are not allowed except for the case of Article 3 of the CDSM Directive, under which TDM may be conducted by research organisations and cultural heritage institutions for the purpose of scientific research. In this case no permission by the rightholder is required. TDM is also allowed under Article 4, under which this exception is extended to anyone who has lawful access to the material, but the rightholders may opt-out by reserving their rights.

The fact that the rightholders may opt-out using either contractual or technical measures, seems to be an obstacle for TDM practices. The requirement for lawful access and the sensitivity of exceptions to TPMs further complicate the legal landscape and may hinder the innovation and effective use of TDM in the future. The general exception in Article 4 can be overridden under Article 4(3) by contract or TPMs, creating lawful access problems and adding burdens for beneficiaries of the exception.[245] The issue with the TPMs remains a significant barrier, as the CDSM Directive does not adequately address their impact on TDM activities.

In Chapters 5 and 6, there was an extensive analysis on the implications of emerging technologies, such as GenAI and the recently enacted EU AI Act and Data Act. On one hand, the AI Act stipulates that AI model providers must adhere to the EU copyright legislation and publish detailed summaries of the data employed in the training of their models. While these provisions aim to ensure transparency and legal compliance, they also highlight the imperative for

[244] European Commission, "New EU copyright rules that will benefit creators, businesses and consumers start to apply" (4 June 2021). Available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1807
[245] See note 100.

international harmonisation with copyright law regarding the use of copyrighted works for AI model training.

On the other hand, the Data Act aims to enhance the accessibility of machine-generated data by excluding it from the *sui generis* database right. Specifically, Article 43 exempts databases that are trained on data generated using connected products or associated services from *sui generis* protection. This provision provides extensive privileges for researchers involved in TDM, since it facilitates unrestricted access to databases containing machine-generated data. Nonetheless, the scope of the Data Act is also limited, and its provisions related with data access for scientific research purposes are inadequate to fully support data-driven research initiatives.[246]

To conclude, even though CDSM Directive and subsequent regulations such as AI Act and Data Act constitute significant advancement towards the modernisation of copyright law and the adoption to the new technological reality, they also highlight persistent challenges. The balance between protecting rightholders' interests and promoting innovation remains delicate. To fully realize the potential of TDM and AI technologies, further refinements and harmonisation efforts are essential, to ensure that legal frameworks support rather than constitute an impediment scientific research and technological progress. The thesis insists on the importance of having clear and non-overridable exceptions for TDM to foster innovation while safeguard the rights of the rightholders at the same time. As the technological landscape moves forward faster than ever, it is of fundamental importance to always keep up-to-date and make legislative amendments to achieve a balanced and effective copyright protection within EU level.

[246] Manteghi Maryna, "Overcoming Barriers to Text and Data Mining in the Era of ChatGPT: The Proposed Data Act as a Game-Changer" (2024), GRUR International, Volume 73, Issue 1, Pages 34–44, Available at: https://doi.org/10.1093/grurint/ikad098

BIBLIOGRAPHY

Legislation

1. Directive (EU) 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.
2. Directive (EU) 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version) OJ L 111, 5.5.2009, p. 16–22.
3. Directive (EU) 2012/28/EU of the European Parliament and of the Council of 25 October 2012 on certain permitted uses of orphan works. Article 2 (1) OJ L OJ L 299, pp 5-12.
4. Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure.
5. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77, 27.3.1996, p. 20–28.
6. 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 on the free movement of such data (General Data Protection Regulation) (Apr. 27, 2016).
7. Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance) [2022] OJ L 265, 1-66.
8. 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) (Text with EEA relevance) [2022] OJ L 277, 1-102.
9. Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act).
10. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) Text with EEA relevance.

Case Law

1. Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, 143 S.Ct. 1258 (2023).
2. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994).
3. Case C-160/15, GS Media BV v Sanoma Media Netherlands BV and Others, [2016], ECLI:EU:C:2016:644.
4. Case C-466/12, Nils Svensson and Others v Retriever Sverige AB, [2014], ECLI:EU:C:2014:76.
5. Case C-5/08 Infopaq, International A/S v Danske Dagblades Forening, (2009)
6. C 30/14, Ryanair Ltd v PR Aviation BV.
7. CJEU, C-490/14, Freistaat Bayern v Verlag Esterbauer GmbH, EU:C:2015:735, paras 13-14, referring to CJEU, Fixtures Marketing, C-444/02, EU:C:2004:697.
8. CJEU, C-527/15, Stichting Brein v Jack Frederik Wullems, EU:C:2017:300, para 65, referring to CJEU, Football Association Premier League and Others, C-403/08 and C-429/08, EU:C:2011:631, para 168, and CJEU, Infopaq International, C-302/10, EU:C:2012:16.
9. CJEU, C444/02, Fixtured Marketing Ltd v Organismos prognostikon agonon podosfairou AE (OPAP), EU:C:2004:697, para 23.
10. Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135-GBW (D. Del. Mar. 29, 2023).
11. See Authors Guild v. Google, Inc., 804 F.3d 202, 214–15 (2d Cir. 2015).
12. Silverman et al. v. OpenAI, Inc. et al., No. 4:23-cv-03416 (N.D. Cal. Jul. 7, 2023).
13. Tremblay et al. v. OpenAI, Inc. et al., No. 4:2023-cv-03223 (N.D. Cal. Jul. 7, 2023).
14. UKSC, Public Relations Consultants Association Ltd v The Newspaper Licensing Agency Ltd, (2013) UKSC 18, para 1 (Lord Sumption).

EU Publications

1. Council of the European Union, "Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world" (3 December 2023), Press release, updated 2 February 2024. Available at https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/ [last access 7 July 2024].
2. European Commission, Artificial Intelligence — Questions and Answers (December 12, 2023). [Press release]. [last access 7 July]. 2024].https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683

3. European Commission, Commission Staff Working Document, "Impact Assessment Report' Accompanying the document Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act)" (2022) SWD, 34 final, 16, 70 and 139. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022SC0034 [last access 19 July 2024].
4. European Commission, "Data Act Explained" (2024) https://digital-strategy.ec.europa.eu/en/factpages/data-act-explained [last access 19 July 2024].
5. European Commission, Directorate-General for Education, Youth, Sport and Culture, "Impact Assessment of the European Copyright Framework on Digitally Supported Education and Training Practices" Final report (2016), Publications Office, 116. [last access 31 March 2024].
6. European Commission, "New EU copyright rules that will benefit creators, businesses and consumers start to apply" (4 June 2021). Available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1807 [last access 18 July 2024].
7. European Commission, Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act), Brussels [2022] COM (2022) 68 final (proposed Data Act).
8. European Commission, Protection of databases, Shaping Europe's digital future (7 June 2022) https://digital-strategy.ec.europa.eu/en/policies/protection-databases [last access 01 July 2024].
9. European Commission, "Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining: Report from the Expert Group" (2014) (Publications Office of the European Union) 51. [last access 1 April 2024].
10. European Copyright Society, "General Opinion on the EU Copyright Reform Package" (2017) page 5 Available at: https://europeancopyrightsociety.org/wp-content/uploads/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf
11. European Parliament, Artificial Intelligence Act Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))1 https://artificialintelligenceact.eu/wp-content/uploads/2023/06/AIA-%E2%80%93-IMCO-LIBE-Draft-Compromise-Amendments-14-June-2023.pdf

12. European Parliament, "EU AI Act: first regulation on artificial intelligence" (last updated 18 June 2024). https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [last access 7 July 2024].
13. European Union, "Copyright and related rights in the Digital Single Market, Summary of Directive (EU) 2019/790 on copyright in the Digital Single Market" (2019), Eur-Lex. https://eur-lex.europa.eu/EN/legal-content/summary/copyright-and-related-rights-in-the-digital-single-market.html [last access 9 July 2024.]
14. Laurent Le Meur, "TDM Reservation Protocol (TDMRep). Final Community Group Report" (2024), W3C Community and Business Groups. https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/ [last access 09 July 2024].
15. Leistner Matthias and Antoine Lucie, "IPR and the Use of Open Data and Data Sharing Initiatives by Public and Private Actors" (2022), Study commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs at the request of the Committee on Legal Affairs) 50 Available at: https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2022)732266 [last access 7 July 2024];
16. Rosati Eleonora, The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market: Technical Aspects, (2018), Briefing requested by the JURI committee, Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament, page 2.
17. Tambiama Madiega, "Modernisation Of European Copyright Rules: Directive On Copyright In The Digital Single Market" (2019), European Parliament, Legislative Train. https://www.europarl.europa.eu/legislative-train/package-better-access-to-digital-goods-services/file-jd-directive-on-copyright-in-the-digital-single-market text updated 20 June 2024. [last access 09 July 2024].
18. Written by Martin R.F. Senftleben, "Study on EU copyright and related rights and access to and reuse of data" (2022), European Commission. Available at https://archivio.unicas.it/media/7514527/study-on-eu-copyright-and-related-rights-and-access-KI0822205ENN.pdf [last access 07 July 2024].
19. 76/76/EEC: Convention for the European patent for the common market (Community Patent Convention) (OJ L 17 26.01.1976, p. 1, CELEX: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:41975A3490)

Reports

1. Caspers M and Guibault L, "Deliverable D3.3 Baseline Report of Policies and Barriers of TDM in Europe" (2016), FutureTDM. [last access 1 April 2024].
2. Committee on Internal Market and Consumer Protection (IMCO), "Activity Report 2019-2024", https://www.europarl.europa.eu/cmsdata/283430/IMCO%20Activity%20Report%20-%202019-2024.pdf
3. Dahlstedt Palle, "Big Data and Creativity" (2019), European Review, Department of Computer Science and Engineering, European Review, Volume 27, Issue 3, page 411-439.
4. Iain Hargreaves, "Digital Opportunity - Review of Intellectual Property and Growth" (2011), Department for Business, Innovation & Skills, Independent Report 11/968 47 page 5, https://www.gov.uk/government/publications/digital-opportunity-review-of-intellectual-property-and-growth [last access 1 April 2024].
5. LIBER Europe, "Europe's TDM Exception for Research: Will It Be Undermined By Technical Blocking From Publishers?" (10 March 2020) https://libereurope.eu/article/tdm-technical-protection-measures/ [last access 1 April 2024].
6. Max Planck Institute for Innovation and Competition, "Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules: PART B Exceptions and Limitations: Chapter 1 Text and Data Mining" (2017), 4. [last access 31 March 2024].
7. World Intellectual Property Organization (WIPO) "Generative Artificial Intelligence. Patent Landscape Report" (2024), Geneva. https://doi.org/10.34667/tind.49740 [last access 19 July 2024].

Books

1. Carine Bernault, "Droit d'auteur et open access" (2016) Larcier, No. 198.
2. Dusollier Séverine, "*The Protection of Technological Measures: Much Ado About Nothing or Silent Remodeling of Copyright?*" (2014).
3. E Alpaydın, "*Introduction to Machine Learning*" (2004), Cambridge, MA, MIT Press.
4. Europäische Kommission Generaldirektion Kommunikationsnetze, Inhalte und Technologien Copyright, Norbert Maier, Federico De Michiel, Viola Peter, María del Carmen Calatrava Moreno, Chiara Pancotti, Francesca Monaco, Maurice Shellekens, Inge Graef, Nadya Purtova, Andreas Wiebe, European Commission. Directorate General for Communications Networks,

Content and Technology, Technopolis, Centro studi industria leggera, Tilburg University, "Study to Support an Impact Assessment for the Review of the Database Directive Final Report" (2022), Publications Office of the European Union, 2022,1.

5. Gary Miner et. al., "*Practical Text Mining and Statistical Analysis for Non-Structured Text Data*" (2012), First Edition, Academic Press.
6. Jiawei Han et. al., "*Data Mining – Concepts and Techniques*" (2012), Third Edition, Elsevier Inc.
7. Julia Luc, « *L'Intelligence artificielle n'existe pas* » (First Éditions, 2019).
8. Jurafsky Daniel and Martin James H., "*Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*" (2008), Prentice Hall, Upper Saddle River, New Jersey 07458, Second Edition.
9. Rochelle Cooper Dreyfuss and Jane C Ginsburg (eds), *Intellectual Property at the Edge* (Cambridge University Press).
10. Senftleben Martin, "*Study on EU Copyright and Related Rights and Access to and Reuse of Data*" (2022), European Commission.

Articles

1. Adamopoulou E. and Moussiades L., "An Overview of Chatbot Technology" (2020) Artificial Intelligence Applications and Innovations.
2. Ahmadi Sina, "A Comprehensive Study on Integration of Big Data and AI in Financial Industry and its Effect on Present and Future Opportunities" (2024), International Journal of Current Science Research and Review, 07 (01).
3. A. Alemanno, G. Helleringer, Geneviève and A.-L. Sibony, «Brève introduction à l'analyse comportementale du droit», D. (2016), p. 911
4. Badillo et al., "An Introduction to Machine Learning" (2020), Clinical Pharmacology & Therapeutics, 107. 10.1002/cpt.1796.
5. Bender Emily M. and Koller Alexander, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data" (2020), In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198, Online. Association for Computational Linguistics.
6. Blessing Elisha et al., "Utilizing AI and Data Analytics to Derive Insights from Large Datasets, Aiding in Decision-Making Processes" (2023).

7. Bottis Maria and others, “Text and Data Mining in the EU Acquis Communautaire Tinkering with TDM & Digital Legal Deposit” (2019) 12, Erasmus Law Review, 180.
8. Brown, Tom & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared & Dhariwal, Prafulla & Neelakantan, Arvind & Shyam, Pranav & Sastry, Girish & Askell, Amanda & Agarwal, Sandhini & Herbert-Voss, Ariel & Krueger, Gretchen & Henighan, Tom & Child, Rewon & Ramesh, Aditya & Ziegler, Daniel & Wu, Jeffrey & Winter, Clemens & Amodei, Dario, “Language Models are Few-Shot Learners” (2020).
9. Cano Jenn, “The V’s of Big Data: Velocity, Volume, Value, Variety, and Veracity” (2014).
10. Casellati Alvise Maria, “The Evolution of Article 6.4 of the European Information Society Copyright Directive” (2001), 24 Columbia - VLA Journal of Law & the Arts 369, 374.
11. Christensen Kristina, “A European Solution for Text and Data Mining in the Development of Creative Artificial Intelligence: A Specific Focus on Articles 3 and 4 of the Digital Single Market Directive” (2021), Stockholm Intellectual Property Law Review, Volume 4, Issue 2. https://stockholmiplawreview.com/wp-content/uploads/2022/01/TA-European-solution-for-Text-and-Data-Mining_ryck_IP_nr-2_2021_A4.pdf
12. Clark, Jonathan, “Text Mining and Scholarly Publishing” (2012), Publishing Research Consortium, page 5-6. Available at: https://www.stmassoc.org/2012_01_01_PRC_Clark_Text_Mining_and_Scholarly_Publishing.pdf [last access 7 January 2024].
13. Closa Carlos and Fossum John Erik, “Multi-Level Governance and the European Union: Analysing Variations in Effectiveness” (2004), in European Integration online Papers (EIoP), Vol. 8 (2004) N° 16.
14. Colangelo Giuseppe, “European Proposal for a Data Act-A First Assessment” (2022). Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4199565
15. Dale R., “The return of the chatbots” (2016) Natural Language Engineering.
16. De Michele Roberta and Furini Marco, “IoT Healthcare: Benefits, Issues and Challenges” (2019), GoodTechs ‘19: EAI International Conference on Smart Objects and Technologies for Social Good. Available at: https://dl.acm.org/doi/10.1145/3342428.3342693
17. Derclaye Estelle and Husovec Martin, “Why the sui generis database clause in the Data Act is counter-productive and how to improve it?” (2022). Available at SSRN: https://ssrn.com/abstract=4052390 [last access 16 June 2024].
18. Diksha Khurana & Aditya Koli & Kiran Khatter & Sukhdev Singh, “Natural language processing: state of the art, current trends and challenges” (2022), Springer.

19. Dusollier Séverine, “Tipping the Scale in Favor of the Right Holders: The European Anti Circumvention Provisions” (2003), Eberhard Becker and others (eds), Digital Rights Management: Technological, Economic, Legal and Political Aspects, Springer 2003, 465–466. https://doi.org/10.1007/10941270_29, also see note 163, page 253.
20. Fenwick Mark et al., “Generative AI: Rights and Liabilities,” (2023), Journal of European Tort Law.
21. Floridi Luciano, “AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models” (2023), Philosophy & Technology. 36. 10.1007/s13347-023-00621-y.
22. Gasser Urs, “Legal Frameworks and Technological Protection of Digital Content: Moving Forward Towards a Best Practice Model” (2006), Berkman Klein Center for Internet & Society [last access 1 April 2024].
23. Geiger Christophe, Frosio Giancarlo, and Bulayenko Oleksandr, “The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects” (2018), Directorate General For Internal Policies, Policy Department For Citizens Rights and Constitutional Affairs. https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf
24. Geiger Christophe, Iaia Vincenzo, (Luiss Guido Carli University) “Generative AI, Digital Constitutionalism and Copyright: Towards a Statutory Remuneration Right grounded in Fundamental Rights—Part 1” (17 October 2023), Kluwer Copyright Blog. Available at https://copyrightblog.kluweriplaw.com/2023/10/17/generative-ai-digital-constitutionalism-and-copyright-towards-a-statutory-remuneration-right-grounded-in-fundamental-rights-part-1/ [last access 19 July 2024].
25. Gemini Team, Google, “Gemini: A Family of Highly Capable Multimodal Models” Google Deep Mind. Available at: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf
26. Gillen Martina and Sutter Gavin, “DRMS and Anti-Circumvention: Tipping the Scales of the Copyright Bargain?” (2006) 20 International Review of Law, Computers & Technology 287, 291.
27. Giannopoulou Alexandra, “Proposed Directive on Copyright in the Digital Single Market: A Missed Opportunity?” (2018), Alexander von Humbolt Institute for Internet und Gesellschaft: Digital Society Blog. https://www.hiig.de/en/proposed-directive-on-copyright-in-the-digital-single-market-a-missed-opportunity/ [last access 19 July 2024].

28. Guido Noto La Diega and Derclaye Estelle, "Opening Up Big Data for Sustainability: What Role for Database Rights in the Fourth Industrial Revolution?" (2022), Ole-Andreas Rognstad, Taina Pihlajarinne and Jukka Mähönen (eds), Promoting Sustainable Innovation and the Circular Economy: Legal and Economic Aspects, 26. Available at SSRN: https://ssrn.com/abstract=4220534 [last accessed 7 July 2024];
29. Hacker Philipp et al., "Regulating ChatGPT and Other Large Generative AI Models" (2023), in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1112–23 (New York, Association for Computing Machinery 2023) https://doi.org/10.1145/3593013.3594067 [last access 5 March 2024].
30. Hearst Marti A., "Untangling Text Data Mining. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics" (1999), College Park, Maryland, USA. Association for Computational Linguistics. pages 3–10. Available at: https://dl.acm.org/doi/10.3115/1034678.1034679
31. Heudin, Jean-Claude, "Intelligence artificielle et intelligence humaine" (2019), Futuribles. N° 428. 93. 10.3917/futur.428.0093
32. Hugenholtz P. Bernt, "Auteursrecht op informatie (1989), Kluwer, Deventer as discussed in Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, ''Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU', 6.
33. Hugenholtz P. Bernt, "The New Copyright Directive: Text and Data Mining (Articles 3 and 4)" (2019), Kluwer Copyright Blog. https://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ [last access 09 July 2024].
34. GPT-4 System Card, OpenAI, (March 23, 2023) https://cdn.openai.com/papers/gpt-4-system-card.pdf
35. Jianyang Deng and Yijia Lin, 'The Benefits and Challenges of ChatGPT: An Overview' (2022) 2 FCIS 81, 82.
36. Jondet Nicolas, "The Text and Data Mining Exception in the Proposal for a Directive on Copyright: Why the European Union Needs to Go Further than the Laws of Member States" (2018), 67 Propriétés Intellectuelles 25, 19.
37. Johnson KB et al. "Precision Medicine, AI, and the Future of Personalized Health Care" (2021), Clin Transl Sci.;14(1):86-93. doi: 10.1111/cts.12884.
38. Keller Paul and Warso Zuzanna, "Defining Best Practices for Opting Out of ML Training" (2023), Open Future Policy Brief #5. https://openfuture.eu/wp-content/uploads/2023/09/Best-_practices_for_optout_ML_training.pdf

39. Kuschel Linda and Dolling Jasmin, "Access to Research Data and EU Copyright Law" (2022), 13 JIPITEC 247. Available at https://www.jipitec.eu/archive/issues/jipitec-13-3-2022/5558
40. Lokesh Kumar et. al., "Text Mining: Concepts, process and application" (2013), Journal of Global Research in Computer Science. Volume 4, Issue 3, page 37. Available at https://www.researchgate.net/publication/277160258_TEXT_MINING_CONCEPTS_PROCESS_AND_APPLICATIONS
41. Manteghi Maryna, "Overcoming Barriers to Text and Data Mining in the Era of ChatGPT: The Proposed Data Act as a Game-Changer" (2024), GRUR International, Volume 73, Issue 1, Pages 34–44, Available at: https://doi.org/10.1093/grurint/ikad098
42. Margoni Thomas and Kretschmer Martin, "A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Machine Learning" (2022), 19 European Intellectual Property Review 397.
43. Mariani Marcello and Dwivedi Yogesh Kumar, "Generative Artificial Intelligence in Innovation Management: A Preview of Future Research Developments" (2024), Journal of Business Research, Volume 175.
44. MA Lemley and B Casey, "Fair Learning" (2021). Law Review, 743.
45. Meyer et al. "ChatGPT and large language models in academia: opportunities and challenges", (2023), BioData Mining. 16. 10.1186/s13040-023-00339-9.
46. Özkent Yasemin, "Social Media Usage to Share Information in Communication Journals: An Analysis of Social Media Activity and Article Citations" (2022), PLoS ONE 17(2): e0263725. https://doi.org/10.1371/journal.pone.0263725
47. Ozkayagan Hande, "Discover Insights from the Copyright Office Hours on Artificial Intelligence" (2023), Europeana Pro. https://pro.europeana.eu/post/discover-insights-from-the-copyright-office-hours-on-artificial-intelligence [last access 09 July 2024].
48. Quintais João Pedro and Schwemer Sebastian Felix, "The Interplay Between the Digital Services Act and Sector Regulation: How Special Is Copyright?" (2022), European Journal of Risk Regulation, 13(2), 191–217. doi:10.1017/err.2022.1
49. Portnoff AY and Soupizet, "Artificial intelligence: opportunities and risks. Futuribles" (2018) doi: 10.3917/futur.426.0005
50. Rossana Ducato and Strowel, Alain M., "Ensuring Text and Data Mining: Remaining Issues With the EU Copyright Exceptions and Possible Ways Out" (2021), European Intellectual Property Review, 322, 333. Available at SSRN: https://ssrn.com/abstract=3829858
51. Rossana Ducato and Alain Strowel, "Limitations to Text and Data Mining and Consumer Empowerment. Making the Case for a Right to "Machine Legibility" (19 March 2019), Kluwer

Copyright Blog http://copyrightblog.kluweriplaw.com/2019/03/19/limitations-to-text-and-data-mining-and-consumerempowerment-making-the-case-for-a-right-to-machine-legibility/ [last access 1 April 2024].

52. Rosati Eleonora - Adrian Aronsson-Storrier, "Contractual override and the new exceptions in the Copyright in the Digital Single Market Proposal" (2018).
53. Rosati Eleonora, "An EU Text and Data Mining Exception for the Few: Would it Make Sense?" (2018), Journal of Intellectual Property Law & Practice, 13 (6), 429-430.
54. Rosati Eleonora, "Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity" (2019), Asia Pacific Law Review, Volume 27, Issue 2, pages 198-217.
55. Rosati Eleonora, "No step-free copyright exceptions: The role of the three-step in defining permitted uses of protected content (including TDM for AI-training purposes)" (2024) European Intellectual Property Review https://www.diva-portal.org/smash/get/diva2:1825121/FULLTEXT01.pdf
56. Senftleben Martin, "Generative AI and Author Remuneration" (2023), International Review of Intellectual Property and Competition Law, 54 (2023), pp. 1535-1560. Available at http://dx.doi.org/10.2139/ssrn.4543794
57. Strowel Alain, "ChatGPT and Generative AI Tools: Theft of Intellectual Labor?" (2023), SprinkerLink IIC 54, 491–494 https://doi.org/10.1007/s40319-023-01321-y
58. Synodinou Tatiana Eleni, "Lawfulness for Users in European Copyright Law: Acquis and Perspectives" (2019) JIPITEC 20 para 1. https://www.jipitec.eu/archive/issues/jipitec-10-1-2019/4876/JIPITEC_10_1_2019_20_Synodinou
59. Tsaone Swaabow Thapelo et al., "Sasscal Websapi: A Web Scraping Application Programming Interface to Support Access to Sasscal's Weather Data" (2021), Data Science Journal, vol. 20, no. 1.
60. Vesala Juha, "Developing Artificial Intelligence-Based Content Creation: Are EU Copyright and Antitrust Law Fit for Purpose?" (2023), 54 International Review of Intellectual Property and Competition Law, 351.
61. Wang Z. et al., "Applications of Generative Adversarial Networks (GANs) in Radiotherapy: Narrative Review" (2022), Precision Cancer Medicine (PCM) A Journal Aiming to Deliver Long Lasting Blow to Cancer.
62. Wiederhold, Gio & McCarthy, John, "Arthur Samuel: Pioneer in Machine Learning" (1992), IBM Journal of Research and Development. 36. 329 - 331. 10.1147/rd.363.0329.

63. Ziaja Gina Maria, “The Text and Data Mining Opt-Out in Article 4(3) CDSMD: Adequate Veto Right for Rightholders or a Suffocating Blanket for European Artificial Intelligence Innovations?” (2024), Journal of Intellectual Property Law & Practice, 2024, Vol. 19, No. 5.

Miscellaneous Electronic Sources

1. Balázs Bodó, “The Science of Piracy, the Piracy of Science. Who Are the Science Pirates and Where Do They Come from: Part 1” (6 March 2019, Kluwer Copyright Blog. https://copyrightblog.kluweriplaw.com/2019/03/06/the-science-of-piracy-the-piracy-of-science-who-are-the-science-pirates-and-where-do-they-come-from-part-1/ [last access 1 April 2024].
2. CIPPIC the Samuelson-Glushko Canadian Internet Policy and Public Interest Clinic, “Text & Data Mining On Copyright Protected Works For Use By Generative AI | TDM versus Training: The difference between making and using a dataset” (2024). https://www.cippic.ca/articles/the-difference-between-tdm-and-training 25/04/2024 [last access 09 July 2024].
3. CMS Law Now, “An Overview of the Data Act” (Germany, 15 January 2024) https://cms-lawnow.com/en/ealerts/2024/01/an-overview-of-the-data-act [last access 19 July 2024].
4. Coraggio Giulio, “AI Act – What Is the Scope of the TDM Copyright Exception?” (April 2 2024) https://www.linkedin.com/pulse/ai-act-what-scope-tdm-copyright-exception-giulio-coraggio-yawcf/ [last access 01 June 2024].
5. Harrington Madelaine, Hansen Marty, Peets Lisa, Drake Marianna & Young Mark, “EU Parliament Adopts AI Act” (15 March 2024), Covington https://www.insideglobaltech.com/2024/03/15/eu-parliament-adopts-ai-act/ [last access 7 July 2024].
6. ECSA, Joint Statement on Generative Artificial Intelligence and the EU AI Act (25 April 2024) https://composeralliance.org/news/2024/4/joint-statement-on-generative-artifical-intelligence-and-the-eu-ai-act/ [last access 01 June 2024].
7. Edd Gent, “What Is Artificial Intelligence (AI)?” (14 April 2024), Live Science. https://www.livescience.com/technology/artificial-intelligence/what-is-artificial-intelligence-ai [last access 09 July 2024].
8. Emanuilov Ivo, Margoni Thomas (Ku Leuven Centre for IT & IP Law), “Memorisation in generative models and EU copyright law: an interdisciplinary view” (March 26t 2024), Kluwer

Copyright blog https://copyrightblog.kluweriplaw.com/2024/03/26/memorisation-in-generative-models-and-eu-copyright-law-an-interdisciplinary-view/ [last access 01 June 2024].

9. Fast Science, “Large language models (LLM) and NLP: A new era of AI and ML has begun” (2024) https://fastdatascience.com/generative-ai/llm-nlp/ [last access 10 July 2024].
10. Flynn and João Pedro Quintais, “Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for Action at International Level” (21 April 2020), Kluwer Copyright Blog, [last access 31 March 2024].
11. Foote Keith D., “A Brief History of Natural Language Processing, (2023), Dataversity https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/ [last access 3 February 2024].
12. Kabbay, Harcharan, "Artificial Neural Network Concepts and Examples" (2022). Theses. 402. https://irl.umsl.edu/thesis/402
13. Karathanasis Theodoros, “EU Copyright Directive: A ‘Nightmare’ For Generative AI Researchers And Developers?” (17 October 2023), AI Regulation. com. MIAI Grenoble Alpes. https://ai-regulation.com/eu-copyright-directive-a-nightmare-for-gai/ [last access 19 July 2024].
14. Keller Paul, “A first look at the copyright relevant parts in the final AI Act compromise” (Monday, December 11th, 2023), Kluwer Copyright Blog, (Institute for Information Law (IViR)) https://copyrightblog.kluweriplaw.com/2023/12/11/a-first-look-at-the-copyright-relevant-parts-in-the-final-ai-act-compromise/ [last access 19 July 2024].
15. Keller Paul, Research Librarians: New TDM exception can be undermined by technical blocking from publishers” (10 March 2020), Communia. https://communia-association.org/2020/03/10/reserach-librarians-new-tdm-exception-can-undermined-technical-blocking-publishers/ [last access 10 July 2024].
16. Kretschmer Martin, Eleni Synodinou Tatiana, Margoni Thomas, “The Paradox Of Lawful Access” (2024) European Copyright Society (ECS). Available at https://europeancopyrightsociety.org/wp-content/uploads/2024/06/kretschmer-synodinou-margoni.pdf
17. Kyriakidis, Harris, “Cyprus transposes the EU Copyright Directive in Cyprus law” (2022), Harris Kyriakides. https://www.harriskyriakides.law/insights/news/cyprus-transposes-the-eu-copyright-directive-in-cyprus-law [last access 09 July 2024].
18. Lefkowitz Melanie, Professor’s perceptron paved the way for AI – 60 years too soon” (2019), Cornell Chronicle https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon [last access 3 February 2024].

19. L'Institut des jurists d'entreprise, "The Eu Copyright Directive: The Three Most Controversial Provisions" (2020), Partnerblog. https://ibj.be/fr/news/partnerblog/the-eu-copyright-directive-the-three-most-controversial-provisions [last access 09 July 2024].
20. McKinsey and Company, "What Is Generative AI?" (2 April 2024). https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai [last access 17 July 2024].
21. OBVIOUS which produces art using artificial intelligence https://obvious-art.com/
22. Nobre Teresaand Mileszyk Natalia "Implementing the new EU protections against contractual and technological overrides of copyright exceptions" (9 December 2019), Communia https://communia-association.org/2019/12/09/implementing-new-eu-protections-contractual-technological-overrides-copyright-exceptions/ [last access 1 April 2024].
23. Open AI, "How ChatGPT and Our Language Models Are Developed". https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed [last access 19 July 2024].
24. Quintais J, "Generative AI, Copyright and the AI Act" (09 May 2023), Kluwer Copyright Blog. Available at https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/ [last access 18 July 2024].
25. Samuelson Pamela, "The EU's Controversial Digital Single Market Directive - Part II: Why the Proposed Mandatory Text- and Data-Mining Exception Is Too Restrictive" (12 July 2018), Kluwer Copyright Blog http://copyrightblog.kluweriplaw.com/2018/07/12/eus-controversial-digital-single-marketdirective-part-ii-proposed-mandatory-text-data-mining-exception-restrictive/ [last access 1 April 2024].
26. Schirru Luca, Margoni Thomas, "Arts 3 and 4 of the CDSM Directive as regulatory interfaces: shaping contractual practices in the commercial scientific publishing and stock images sectors" (22 August 2023) Kluwer Copyright Blog https://copyrightblog.kluweriplaw.com/2023/08/22/arts-3-and-4-of-the-cdsm-directive-as-regulatory-interfaces-shaping-contractual-practices-in-the-commercial-scientific-publishing-and-stock-images-sectors/
27. Schlackman Steve, "Who holds the Copyright in AI created art?" (2020), Artrepreneur. Available at: https://alj.artrepreneur.com/the-next-rembrandt-who-holds-the-copyright-in-computer-generated-art/ [last access 3 February 2024];
28. Tobin Sam, "Getty asks London court to stop UK sales of Stability AI system" (1 June 2023), Reuters https://www.reuters.com/technology/getty-asks-london-court-stop-uk-sales-stability-ai-system-2023-06-01/ [last access 9 March 2024].

29. University of Birmingham, Text and Data Mining (TDM). https://intranet.birmingham.ac.uk/as/libraryservices/library/copyright/text-and-data-mining/text-and-data-mining.aspx [last access 09 July 2024].
30. Varese Elena "Can generative artificial intelligence rely on the copyright text and data mining (TDM) exception for its training?" (2023), GamingTechLaw. https://www.gamingtechlaw.com/2023/02/artificial-intelligence-training-copyright-tdm-exception/ [last access 09 July 2024].
31. Vischer, Part 10, "Copyright and AI: Responsibility of Providers and Users" (19 March 2024). https://www.vischer.com/en/knowledge/blog/part-10-copyright-and-ai-responsibility-of-providers-and-users/ [last access 09 July 2024].
32. Wankhede Calvin, "What is Midjourney AI and how does it work?" (March 6, 2024). Available at: https://www.androidauthority.com/what-is-midjourney-3324590/ [last access 7 July 2024].
33. World Economic Forum, "Will Copyright Law Enable or Inhibit Generative AI?" (13 January 2024). https://www.weforum.org/agenda/2024/01/cracking-the-code-generative-ai-and-intellectual-property/ [last access 17 July 2024].
34. Würtenberger Gert, Protection of trade secrets and know-how in the European Union: the EU Trade Secrets Directive (EU) 2019/943, (20 August 2019) https://ip-iurisdictio.org/protection-of-trade-secrets-and-know-how-in-the-european-union-the-eu-trade-secrets-directive-eu-2019-943/