Vilnius University Faculty of Law

Department of Private Law

Mariia Afanasieva,
II study year, International and European Law Programme Student

Master's Thesis

Text And Data Mining In Copyright Law

Supervisor: prof. dr. Ramūnas Birštonas

Reviewer: assist. Justinas Drakšas

Vilnius
2024

ABSTRACT AND KEY WORDS

This work analyses the copyright law frameworks governing text and data mining activities in the European Union and the United States. These legal regimes are critically evaluated and their efficiency is assessed through a comparative approach. The study then proposes legislative recommendations aimed at striking a fair balance between fostering innovation and safeguarding the interests of intellectual property rightholders.

Keywords: text and data mining, copyright law, artificial intelligence, right of reproduction, sui generis database right.

TABLE OF CONTENTS

INTRODUCTION

Our society is currently experiencing a surge in data generation. It is estimated that 181 zettabytes of data will be generated in 2025, compared to the 2 zettabytes produced in 2010.[1] To illustrate, storing 181 zettabytes of data on classic Blu-ray discs would require 7.24 trillion of those**.** When stacked, the discs would reach a height of over 1 million kilometres, which is more than twice the distance from the Earth to the Moon.[2]

Relevance

These exponentially growing, mounting amounts of information, call for it no longer to be labelled as simple data but 'Big Data,' containing great variety arriving in increasing volumes and with ever-higher velocity.[3] The rise of big data is practically justifiable, as it enables us to solve business problems that traditional data could not tackle.[4] Netflix debuted as a DVD rental-by-mail service and advanced to a giant corporation with the help of large quantities of data we created.[5] This data, however, its search, processing and tactical use is favourable not only towards businesses. It can fasten democratic participation, rights to information and bring invention. The "Panama Papers" scandal involved journalists searching huge amounts of financial data to reveal the case.[6] BlueDot AI, a tool trained on great corpora,[7] sent off early warnings before the Coronavirus outbreak.[8]

It is, thus, quite demonstrable that our dependence on big data only increases as the world evolves and presents new social dilemmas, technologies and industry challenges. In turn, said data depends on the techniques of crawling, scraping or mining which extract its value. Text and data mining (TDM) is a process of applying structure to unstructured texts

[1] Duarte, F. (2024). Amount of data created daily. Available at: https://explodingtopics.com/blog/data-generated-per-day (Accessed: 31 October 2024).

[2] Data generated with Perplexity (2024).

[3] For the primary conceptualization of big data's 3Vs (variety, volume, velocity), *see* Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety, at p. 2.

[4] More on the distinctions between data and big data, *see* Treehouse Tech Group (2021). Big Data vs. Traditional Data: What's the Difference? Available at: https://treehousetechgroup.com/big-data-vs-traditional-data-whats-the-difference/#:~:text=Ultimately%2C%20big%20data%20refers%20to (Accessed: 31 October 2024).

[5] VivekR (2023). How did Netflix use big data to transform their company and dominate the streaming industry? *Medium*. Available at: https://vivekjadhavr.medium.com/how-did-netflix-use-big-data-to-transform-their-company-and-dominate-the-streaming-industry-a93f90ae8dad (Accessed: 31 October 2024).

[6] Geiger, C., et al. (2018). Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?: Legal Analysis and Policy Recommendations. *IIC International Review of Intellectual Property and Competition Law*, *49*(7), 814–844. https://doi.org/10.1007/s40319-018-0722-2.

[7] The term refers to a set of texts under study, singular *corpus*.

[8] Niiler, E. (2020). An AI epidemiologist sent the first warnings of the coronavirus. *WIRED*. Available at: https://www.wired.com/story/ai-epidemiologist-wuhan-public-health-warnings (Accessed: 31 October 2024).

and employing statistical methods to discover new information and reveal patterns in the processed data.[9] TDM is a highly beneficial tool in the development of many heterogeneous environments but also enthralling in the eyes of legislators. For one, to specifically address TDM, in 2019 the European Union (EU) adopted the Directive on Copyright and Related Rights in the Digital Single Market (CDSMD).[10]

The above has caused EU-focused, as well as worldwide, discussions on the potential of TDM being a copyright infringement. Fairly so, as there is an ongoing tension between intellectual property protection and TDM activities. Even though fundamentals of copyright law dictate data as such is not protected but its creative form only,[11] legal issues still persist. Since TDM involves copying large quantities of content, some of it might actually be copyrighted and, any such copy, in whichever form, has the potential to infringe the right of reproduction. Thus, TDM may involve works covered by intellectual property protection, both copyrights or database sui generis rights and interfere with their protection, depending on the jurisdiction.

Aim and Object

The aim of this thesis is to analyze and compare the legal frameworks governing TDM under copyright law, with a focus on the EU and United States (US) approaches; critically comment on the efficiency of assessed regimes and explore the need for legislative change to balance innovation and the interests of rightholders. Regarding delimitations, the scope of the research is limited to EU and US jurisdictions, with a bigger stress on the former. Additionally, privacy and data protection aspects, commonly arising from personal data scrapings, are an independent topic and not part of this work.

Tasks and Novelty

The study in this thesis sets forth the following tasks:

- Provide an overview of text and data mining and its implications for copyright law.

[9] Dickson, E., et al. (2018). Data Mining Research with In-copyright and Use-limited Text Datasets: Preliminary Findings from a Systematic Literature Review and Stakeholder Interviews. *International Journal of Digital Curation*, *13*(1), 183–194. https://doi.org/10.2218/ijdc.v13i1.620.

[10] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (CDSMD) [2019] OJ L 130, 17.5.2019, p. 92–125.

[11] Art. 2(1), Berne Convention for the Protection of Literary and Artistic Works. (1886). Available at: https://www.wipo.int/wipolex/en/text/283698. Also, in *SAS Institute Inc. v World Programming Ltd.,* [CJEU], Nr. C 406/10, [02.05.2012]. EU:C:2012:259, para. 9, the CJEU recognized that although "computer programs are protected by copyright as literary works within the meaning of the Berne Convention ... *the ideas and principles* which underlie any element of a computer program, including those which underlie its interfaces, are not protected by copyright."

- Examine the EU copyright framework's treatment of TDM, with emphasis on existing exceptions and limitations to TDM activities.
- Analyze the US approach to TDM under the fair use doctrine.
- Compare and contrast the EU and US approaches to regulating TDM activities.
- Evaluate should legal frameworks strike a fair balance between fostering innovation through TDM and safeguarding the rights and interests of copyright holders.
- Assess the necessity for legislative reforms to achieve an optimal regulation of TDM under the copyright law.

The novelty of this work lies in the specific research intersection between emerging Generative Artificial Intelligence (GenAI) technologies, TDM practices for Large Language Model (LLM) training, and the ways copyright law restricts or permits those under different regimes.

Research Methods and Main Sources

The intended ways and methods to carry out the research tasks include:

- A qualitative method, used to examine relevant provisions in legal texts and academic articles, relevant to TDM under copyright law.
- Case studies of the Court of Justice of the European Union (CJEU/Court) and US courts practice attending the admissibility of TDM, copyright infringements and related issues.
- A comparative method, applied to conclude the preferability, advantages and (or) disadvantages of researched legal regimes.

The main sources used in this work consist of legal doctrine, literature and court practice, stressing the EU and US legal acts:

- Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market (CDSMD).
- Directive (EU) 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society (InfoSoc).
- Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act).
- Directive (EU) 96/9/EC on the Legal Protection of Databases (Database Directive).
- Copyright Act of 1976, 17 U.S.C. §§ 101 et seq. (as amended up to Public Law No. 117-81), United States of America (Copyright Act).

1. AN INSIGHT INTO TEXT AND DATA MINING

The purpose of this chapter is to, foremost, provide clarity on the correlating terms addressing crawling, scraping and mining activities. Second, to describe TDM, each step forming its procedure and illustrate why those raise legal concerns. Lastly, given that TDM is involved in the training of AI models, to delve into basic copyright issues of GenAI.

From the EU perspective, 'TDM' is a concept introduced in 2019 by CDSMD that has since been fostered as an umbrella term: "'text and data mining' means any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations."[12] On a bigger scale, text and (or) data mining terminology has been around since the 1990s.[13] For instance, the US courts and scholars had spoken on it long before the EU's input.[14] On the contrary, 'data scraping' is used to describe a plethora of Internet-based data retrieval methodologies, manual or automatic, occurring through screen-scraping, web-scraping or web-crawling.[15] Hence, scraping itself does not necessarily entail automatic study or later processing of the retrieved information, so it can be characterized as the preliminary step in the TDM process. That is, researchers frequently define scraping with emphasis on data being *prepared* for further quantitative or qualitative analysis.[16] Said scrapings usually involve personal data collection and get questioned on the lawfulness of such conduct under the General Data Protection Regulation, as opposed to copyright issues, usually raised with TDM. Thus, it is important to separate and not confuse the mentioned processes.

Legal issues arising in TDM procedure

TDM can occur through different techniques, have case-specific details and end goals. Nevertheless, for the most part, there are common steps as described henceforth. Academic sources conceptualize the steps of TDM research in various breadths. For example, one outlook is that typical TDM contains the following: (1) Access to content; (2) Extraction

[12] Art. 2(2) CDSMD. Previously referred to as 'data mining', e.g. *see* Cerquitelli, T., et al. (2017). Transparent Data Mining for Big and Small Data. http://www.springer.com/series/11970.

[13] First appeared in 1989 under the name Knowledge Discovery in Databases (KDD), also known in French as *Extraction de Connaissances à partir des Données* (EDC). https://books.openedition.org/oep/1745.

[14] For instance, in *Authors Guild, Inc. v. HathiTrust,* No. 11 CV-4351 (HB), 2012 WL 4808939 (S.D.N.Y. Oct. 10, 2012), at 14, Judge Baer held that the defendants were entitled to summary judgment on their fair use defense because, in part, their use was "transformative" in the sense that search capabilities of the HathiTrust database have already given rise to new methods of academic inquiry, such as *text mining*.

[15] Campbell, F. (2019). Data Scraping – Considering the Privacy Issues. *Fieldfisher*. Available at: https://www.fieldfisher.com/en/services/privacy-security-and-information/privacy-security-and-information-law-blog/data-scraping-considering-the-privacy-issues (Accessed: 2 November 2024).

[16] E.g. Krotov, V., Tennyson, M., (2018). Research Note: Scraping Financial Data from the Web Using the R Language. *Journal of Emerging Technologies in Accounting*, Vol. 15, No. 1, 169-181.

and/or copying of content; (3) Mining of text and/or data and knowledge discovery.[17] Similarly, often cited by others,[18] the scheme by Matthew Sag includes: (1) Access (either physical or digital); (2) Extraction (copying); (3) Mining (analytical processing, internal verification, and external validation); and (4) Use.[19] From a technical perspective, the workflow of a typical data mining application contains subsequent phases: (1) Data collection; (2) Data preprocessing (includes feature extraction and data cleaning); and (3) Analytical processing.[20] For the purposes of this work, the initial scheme seems most favourable.

The first stage – access to content – immediately puts forth legal considerations of obtaining *lawful* access to the respective work. To clarify, only digital works are contemplated here since accessing works in print would rather be a matter of property or contract law, and not copyright. There are two general ways lawful access to content can exist – said content is freely accessible (e.g. public domain) or access to it has been granted (via approved permission, licence, etc.). Though the latter does not necessarily entitle one to perform TDM on such data or text.[21] Some related aspects should be considered here. Is the content hosted on a medium presenting 'Terms of Service' (ToS) restrictions? And does it deploy some technical preventions? Meta Platforms explicitly prohibit attempting access and accessing or collecting data from their Products using automated means without prior permission; so do they reserve all rights against text and data mining.[22] The New York Times and LinkedIn supplement the same with actual technical measures, including the use of CAPTCHA, monitoring unusual access patterns that suggest automated data collection, blocking IP addresses that exhibit scraping behavior or using JavaScript to hinder automated tools. On the other hand, such platforms as Wikipedia, Project Gutenberg or Internet Archive are key sources allowing TDM of legally low-risk works for nearly every modern AI model.[23] Is TDM performed under the Arts. 3 and 4 CDSMD exceptions?

[17] Rosati, E. (2018). The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market, at p. 4.

[18] For example, Fernández-Molina, J. C., & de la Rosa, F. E. (2024). Copyright and Text and Data Mining: Is the Current Legislation Sufficient and Adequate? *Portal*, *24*(3), https://doi.org/10.1353/pla.2024.a931775, at p. 654.

[19] Sag, M. (2019). The New Legal Landscape for Text Mining and Machine. *Learning. Journal of the Copyright Society of the USA* 66, 2., at p. 34.

[20] Aggarwal, C. C. (2015). *Data Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14142-8.

[21] Rosati (n 17) 5.

[22] Facebook Terms of Service (12 January 2024). Available at: https://www.facebook.com/terms.php (Accessed: 2 November 2024).

[23] Levendowski, A. (2018). How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem. 93:579 *WASH. L. REV.*, at p. 610 noting that due to 'the friction copyright law causes for accessing certain works, many AI creators turn to easily available, legally low-risk works to serve as training data for AI systems. Data derived from these works are often demonstrably biased … *low-friction data*.'

Discussed in detail in Part II of this work, recently established Arts. 3 and 4 provide exceptions to copyright and database protection for reproductions and extractions made for the purposes of scientific research by research organisations and cultural heritage institutions;[24] or made by any type of beneficiaries, granted such use has not been expressly reserved by rightholders in an appropriate manner.[25] To highlight, these privileged uses for TDM also require lawful access as a condition to exercise the exceptions.[26] Thus, when the access step in TDM is performed, it is vital to consider (a) whether the content is freely accessible or could copying for the purpose of TDM otherwise qualify as fair use (under the US law)[27] or fall under the existing exceptions (under the EU law); (b) that for content which is not freely accessible prior permission or a licence may be required; (c) whether the ToS prohibit scraping/TDM activities or the medium itself can mechanically rebuff those.

Extraction or copying can be defined as the propaedeutic TDM phase. Even though copying for non-expressive uses in general, and TDM research in particular, is fair under American copyright law, the EU approach is more restrictive. Namely, if TDM involves extraction or copying in scenarios that do not satisfy any exceptions, both copyright and the sui generis right might come into consideration. To bear in mind, however, that not all TDM practices perform one of the exclusive rights of the author and not all extraction acts are necessarily subject to the control of the rightholder. For instance, acts of copying would not require permission if the content being temporarily copied satisfies the exemption conditions under Art. 5(1) InfoSoc Directive.[28] As held by CJEU on a couple of occasions, "a use should be considered lawful where it is authorised by the right holder or where it is not restricted by the applicable legislation."[29] Moreover, in this sense, the work might not be actually copied but rather some information about it extracted – e.g. hyperlinks, usage

[24] Art. 3 CDSMD.

[25] Art. 4 CDSMD.

[26] Geiger, C., Jütte, B. J. (2024). Copyright as an Access Right: Concretizing Positive Obligations for Rightholders to Ensure the Exercise of User Rights. *GRUR International*, *73*(11), 1019–1035. https://doi.org/10.1093/grurint/ikae130 elaborating *positive obligations* must be imposed on *rightholders*, specifically to grant users access to works and protected subject matter, to foster and promote creativity and innovation.

[27] 17 U.S.C. § 107 of the Copyright Act of 1976.

[28] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L167/10 (InfoSoc Directive).

[29] *Stichting Brein v. Jack Frederik Wullems,* [CJEU], Nr. C-527/15, [26.04.2017]. EU:C:2017:300, para. 65, with reference to 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.

logs or compiled data.[30] The *LAION* case is a prominent example.[31] In it, a photographer sued the nonprofit organization LAION for copyright infringement after LAION included his image, which was uploaded to a stock photo site with restrictions against automated use, in a dataset of nearly six billion image-text pairs intended for AI training, arguing that the limitations to copyright did not apply. However, LAION is a dataset not containing copies of works but extracted text descriptions and image links, treated as a 'communication to the public', thus, the Hamburg Regional Court has recently held the reproduction at issue by LAION was covered by the exception for TDM for scientific purposes. Based on this, some argue that most trained AI models do not contain copies of the dataset, but highly compressed information in a latent space, meaning after the model has been trained, the copied data is not needed anymore, so its nature is temporary.

Mining of text or data and knowledge discovery, factually, is the core of TDM and the ultimate purpose the previous steps work for. The mining procedure is as follows: 1) pre-processing, which includes removal of unnecessary or unwanted information, dealing with tables, figures, formulas, and the normalization of text/data; 2) extraction, in which there is tokenization, identification of synonyms, text transformation, identification of equivalence classes; and, finally, 3) identification of patterns and events extraction.[32] Analytical processing is not considered a copyright-relevant action, as it does not substantially reproduce the copyrighted material in a copy, nor does it publicly perform or display the work.[33] Yet with respect to knowledge discovery and use, it can be argued these parts correlate with what is traditionally the output stage in Machine Learning, i.e. the idea of AI-generated outputs or 'works.' Legal questions then arise about whether said outputs are protected by copyright, if they are derivative works, or infringe on third-party works used in training, etc.[34] However, answers to said reservation are limited within the scope of this work.

Copyright issues of GenAI

Since Autumn 2022, it is practically impossible to avoid GenAI discussions from any taken viewpoint – computer science, finance, ethics, law, education, gaming and so forth. Of

[30] Examples of similar metadata – 'the musical attributes of songs by Billie Eilish, or the song data from every Taylor Swift album'. *See* Guadamuz, A. (2024). A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs. *GRUR International*, *73*(2), at p. 11.

[31] *Kneschke v. LAION,* [Hamburg Regional Court], Nr. 310 O 227/23, [27.09.2024].

[32] Rosati (n 17) 10.

[33] Sag (n 19) 43.

[34] Quintais, J. P. (2024). Generative AI, Copyright and the AI Act (v.2), at p. 4.

course, society knew AI for many decades before that, but the release of ChatGPT, DALL·E 2, Midjourney, Stable Diffusion[35] marked a pivotal moment in public awareness about what GenAI can do. Essentially, the most noteworthy aspect of GenAI is its capability in content creation and the profound understanding of context. For instance, with cultural indications, GenAI understands "The Great Gatsby" is a reference to F. Scott Fitzgerald's novel and its themes of the American Dream while traditional AI may not recognize the significance or background of the reference. Same with ambiguous terms, when facing the word "Bank", GenAI can discern from context whether it refers to a financial institution or the side of a river. Not to mention, GenAI can create an oil painting in the style of Renoir of a cat playing poker.[36] As elucidated by Rosenberg, "these GenAI tools can create new pieces of content that are original and awe-inspiring."[37] Tense interchanges between TDM, Machine Learning and law, specifically, copyright are inevitable. As outlined, LLMs are trained on millions of digital objects, and those might be copyright protected. Said training, which is essentially the second step in TDM or the *input* stage, entails copying the underlying texts, images, sounds and, thus, gives rise to potential legal implications, namely, – copyright can be an obstacle for TDM, henceforth, GenAI development. Even under the supple American regulation, there are contexts where the process of creating GenAI may transition from fair use to infringement when these LLMs "memorize" the training data instead of merely "learning" from it.

In 2024, the European creative community presented a joint letter to Members of the European Parliament on the impact of AI.[38] The addressors make several points on the darker aspect of GenAI, stating that "all generative AI models in existence today have been trained in full opacity on enormous amounts of copyright-protected content and personal data which have been scraped and copied from the internet, without any authorisation nor any remuneration for the creators we represent." In criticizing the EU's legal framework, they express that even though rules do exist, they are either not yet applied, not enforced or ignored by generative AI models. And the previously discussed Art. 4 CDSMD opt-out right has not helped any members reserve their rights in an efficient manner. Interestingly

[35] Systems based on Large Language Models (LLMs), also known as Foundation Models.

[36] Image prompted by Michael D. Murray.

[37] Rosenberg, L. (2022), Generative AI: The technology of the year for 2022, *Big Think*. Available at: https://bigthink.com/the-present/generative-ai-technology-of-year-2022/. (Accessed: 5 November 2024).

[38] Joint letter to Members of the European Parliament on the impact of Artificial Intelligence on the European creative community. (23 July 2024). Available at: https://europeanwriterscouncil.eu/wp-content/uploads/2024/07/Joint-letter-to-Members-of-the-European-Parliament-on-the-impact-of-Artificial-Intelligence-on-the-European-creative-community.pdf.

so, tools like ‘Spawning’ already exist and are unprecedentedly widely used.[39] Spawning appeared as soon as GenAI boomed, in December 2022, and has helped thousands of individual artists and organizations remove 78 million artworks from AI training. In this context, it should be mentioned that the number of lawsuits against AI is only escalating. An anonymous source ‘Chat GPT Is Eating the World’ keeps track of copyright lawsuits in the US, and as of 30 August 2024, there was about 30 of those.[40] The most talked about, *Getty Images (US), Inc. v. Stability AI Inc.* and *Andersen v. Stability AI Ltd.,* involve Stability AI and Midjourney being accused of, among other things, massive copyright infringement.[41] Even though the legal implications and standpoints of opposing parties seem obvious, these cases are not just black and white. As stipulated by Sag, on the surface, the allegations appear quite credible, as it occurs evident that the Machine Learning models central to these cases were trained on thousands of copyrighted photos from Getty Images and millions of works by creators like Sarah Andersen, all without obtaining permission. However, for the plaintiffs to succeed, they must demonstrate that a long history of fair use rulings supporting similar forms of non-expressive use by copyright-dependent technologies were either wrongly decided or do not pertain to generative AI.[42]

Sub-conclusion

To sum up, this chapter has delineated that the steps involved in the TDM procedure give rise to potential legal issues. In particular, when TDM is performed, it must be questioned whether lawful access to a respective work/content has been obtained and if TDM involves extraction or copying that satisfies any exceptions or falls under the fair use doctrine. Furthermore, as examined, copyright issues of GenAI are highly relevant and complex since modern LLMs are trained on millions of digital items, which may be copyright protected Notably, the legal questions raised thus far are subject to further, more profound, analysis.

[39] Spawning opts out 78 million artworks from AI training. (2023). Available at: https://spawning.substack.com/p/spawning-opts-out-78-million-artworks (Accessed: 7 November 2024).

[40] Available at: https://chatgptiseatingtheworld.com/2024/08/27/master-list-of-lawsuits-v-ai-chatgpt-openai-microsoft-meta-midjourney-other-ai-cos/. (Accessed: 7 November 2024).

[41] *Getty Images (US), Inc. v. Stability AI Inc*., No. 1:23-cv00135-UNA (D. Del. Feb. 3, 2023); *Andersen v. Stability AI Ltd.,* No. 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023).

[42] Sag, M. (2023). Copyright Safety For Generative AI. In HOUS. L. REV (Vol. 295), at p. 301.

2. COPYRIGHT LAW FRAMEWORKS ADDRESSING TEXT AND DATA MINING

Digitalisation has altered the way we interact with vast amounts of content by creating new avenues for cooperation. The shift from analogue to digital made it convenient, even effortless, to engage with information accessed online; to gather, process and make use of it in creative and innovative ways. In particular, emerging technologies, such as AI, have led to unexpected possibilities for working with copyrighted material. Reasonably, this also spurred the reaction of the legal frameworks. Thus, both parties, users of digital works and rightholders, have their rights and constraints under copyright law, an essential framework regulating the exploitation of information in the digital environment. However, compared to the seemingly easy way to unwarily access online data, navigating legal rules that determine the conditions for it has become more complex and difficult. Hence, this chapter explores the EU and US copyright law frameworks addressing emerging information uses – TDM activities – and provides a comparative analysis of those with notes on possible amendments.

2.1. EU Copyright Law

The most important regulation on TDM in the EU thus far is the already-mentioned 2019 Copyright in the Digital Single Market Directive. The basic history behind its implementation is the rise of digital technologies, accompanied by the float of legal uncertainty in copyright norms. As mentioned in Recital 5 CDSMD: "In the fields of research, innovation, education and preservation of cultural heritage, digital technologies permit new types of uses that are not clearly covered by the existing Union rules on exceptions and limitations. In addition, the optional nature of exceptions or limitations provided for in Directives 96/9/EC, 2001/29/EC and 2009/24/EC in those fields could negatively impact the functioning of the internal market." Subject to further analysis, new mandatory exceptions and limitations were introduced by CDSMD ('Title II'), along with measures to ensure that their exercise is respected.

Stemming from the above Recital, CDSMD references other substantial EU *acquis* on TDM. In particular, for the purposes of TDM, Member States are to provide exceptions or limitations to the sui generis rights provided in Art. 5(a), 7(1) Databases Directive; and reproduction rights contained in Art. 2 InfoSoc Directive; Art. 4(1)(a), (b) Software

Directive.[43] Notably, the EU interpretation of the 'three-step test' presented in international copyright law by Art. 9(2) Berne Convention, Art. 13 TRIPS Agreement,[44] has significantly shaped its approach when establishing the *acquis* on exceptions. As argued by Geiger and Jütte,[45] the understanding of the three-step test in the EU is overly restrictive, as opposed to the arguably more flexible US fair use doctrine. It should also be highlighted that the patchwork of EU Directives has harmonised the copyright *acquis* 'vertically,' and one of the few – the InfoSoc Directive – has applied a 'horizontal' approach to harmonisation of only certain aspects of copyright (primarily rights), leaving Member States considerable discretion with regard to exceptions.

2.1.1. Protectable Subject Matter and Exclusive Rights

As previously mentioned, TDM can potentially affect two intellectual property regimes – copyright law and sui generis database law. It is thus essential to analyze protected subject matter under these regimes and the exclusive rights conferred upon rightholders.

First, let us consider the basis for the protection of works in general, followed by the specifics for databases. On a general scale and moderately harmonised reaction of national frameworks, what is protected by copyright, has been entrenched by the Berne Convention: the rights of authors in their literary and artistic works,[46] or their collections.[47] Hence, as established that in its second phase, TDM involves extraction or copying of text/data, the exclusive right of reproduction may come into infringement. Stipulated by Art. 2(a) InfoSoc Directive: "The Member States shall provide for the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part: (a) for authors of their works." Therefore, the new Directive, CDSMD, establishes exceptions or limitations with regard to *reproductions and extractions* made for the purposes of TDM.

It is important to weigh the likelihood of the mined content being protected under copyright law. A conclusion surfacing from the InfoSoc Directive analysis and the CJEU

[43] Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Software Directive) [2009] OJ L 111, 5.5.2009, p. 16–22. The sentence rephrases Art. 3(1) CDSMD. Listed items are contained in Art. 3(1) and Art. 4(1) CDSMD.

[44] "TRIPS Agreement (as amended on 23 January 2017)". *World Trade Organization*. Available at: https://www.wto.org/english/docs_e/legal_e/31bis_trips_01_e.htm.

[45] Geiger and Jütte (n 26) 1022.

[46] Art. 1 Berne Convention.

[47] Art. 2(5) Berne Convention.

practice (heavily carried by the *Infopaq* case),[48] suggests that the right of reproduction must be interpreted broadly and the standard for originality is rather low. Taking an approach favourable towards rightholders, Recital 21 of the InfoSoc Directive provides that: "This Directive should define the scope of the acts covered by the reproduction right with regard to the different beneficiaries. […] A broad definition of these acts is needed to ensure legal certainty within the internal market;" and links this to Art. 2 which offers expressions such as "direct or indirect", "temporary or permanent", "by any means" and "in any form".[49] In support of this, the *Infopaq* case judgement ruled that even the extraction from a newspaper, consisting of (11) eleven words, falls within the scope of the author's exclusive right of Art. 2 InfoSoc Directive to authorize or prohibit the making of (partial) reproductions if that part constitutes the author's own intellectual creation.[50] Followed by a more recent 2020 case example, even foldable bicycles constitute expressing the author's creative ability in an original manner.[51] Nonetheless, as stipulated in the preceding section, not all materials employed for TDM purposes attain the necessary threshold of intellectual creation to merit copyright protection, nor do all pertinent activities inherently involve full-on copying or extraction methods as their components. The CJEU did not specify the amount of required creativity, but instead delineated the boundaries of originality through an exclusionary approach. It held that where choices are dictated by technical function, rules or constraints, the author is not able to express his creative ability in an original manner by making free and creative choices.[52] As represented by the *LAION* case, when software merely "crawls" through text or the data, without copying the entire work, but instead extracting only minimal elements like individual words, such activity does not constitute copying in terms of copyright, so it does not require the consent of the rightholder and no exception is needed.[53]

Second, the other type of 'work' – a database – is defined by the Database Directive as "a collection of independent works, data or other materials arranged in a systematic or

[48] *Infopaq International A/S v. Danske Dagblades Forening*, [CJEU], Nr. C-5/08, [16.07.2009]. EU:C:2009:465.

[49] Rosati, E. (2024). Infringing AI: Liability for AI-generated outputs under international, EU, and UK copyright law. *European Journal of Risk Regulation.,* at p. 8.; *Austro-Mechana*, C-433/20, EU:C:2022:217, at paras 16-18.

[50] Supra n 48, at §§34-36.

[51] *Brompton Bicycle*, [CJEU], Nr. C-833/18, [11.06.2020]. EU:C:2020:461. at para. 38: "… where that product is an original work resulting from intellectual creation, in that, through that shape, its author expresses his creative ability in an original manner by making free and creative choices in such a way that that shape reflects his personality…"

[52] Caspers M. et al. (2016). D3.3 Baseline Report of Policies and Barriers of TDM in Europe (*FutureTDM).*

[53] Meeûs d'Argenteuil, J. et al. (2014). European Commission: Directorate-General for the Internal Market and Services. *Study on the legal framework of text and data mining (TDM)*, Publications Office. https://data.europa.eu/doi/10.2780/1475, at p. 31.

methodical way and individually accessible by electronic or other means." Therefore, for databases, the data within is technically irrelevant, and copyright may exist in the collection of either data, independent of any copyright existing in the contents themselves.[54] So, "databases which, by reason of the *selection or arrangement* of their contents, constitute the author's own intellectual creation shall be protected as such by copyright."[55] Sui generis rights then come into play when there has been "qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database," referring to original or *non-original* databases.[56] The exclusive rights under sui generis protection therefore include extraction (i.e. reproduction) and re-utilisation (i.e. distribution or communication to the public). However, it should be mentioned that the CDSMD presented mandatory exceptions which exempt specifically acts of *extraction* for the sui generis database right. Over the years, this special EU intellectual property right has received a sizeable amount of backlash with the primary argument presented: exclusive rights protecting data in both original and non-original qualifying databases *de facto* protect simple data in certain cases.[57] Furthermore, the sui generis right has also been revisited back and forth in the CJEU practice. In the 2021 *CV-Online v Melons* case, the CJEU redefined the sui generis right.[58] Therein, the defendant developed its method to explore the database – a job adverts website – and provided deep links without using the plaintiff's search function. In its ruling, the Court established a sui generis right infringement test, which indicates that if there is no significant detriment to the database maker's investment, there is no infringement. As such, an extraction or re-utilization of the content is not sufficient enough to conclude that there was an infringement. It can be questioned whether the CJEU practice on sui generis rights brings legal clarity and/or is practically reliable, given that a case-by-case analysis is still principal. Whether the Court confirms or overrules this doctrine will be tracked in the future

It is apparent that, for example, the act of copying during TDM may infringe upon the right of reproduction; however, one must consider the potential implications for other

[54] Caspers and others (n 52) 19.

[55] Art. 3(1) Database Directive.

[56] Art. 7(1) Database Directive.

[57] Margoni, T., and Kretschmer, M. (2022). A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology. GRUR International, 71(8), 685–701. https://doi.org/10.1093/grurint/ikac054, at p. 699.

[58] *CV-Online Latvia v. Melons*, [CJEU], Nr. C-762/19, [03.06.2021]. EU:C:2021:434.

exclusive rights. In principle, the processing of protected works can also involve such exclusive rights as distribution, adaptation or communication to the public. Scholars hold differing views on this matter; however, the most logical reading of CDSMD suggests that TDM exceptions do not cover acts of communication to the public, distribution of TDM results, etc. In essence, this rationale aligns with the nature of the TDM process, which culminates at the knowledge discovery stage – with mining at its crux, any subsequent steps are not TDM anymore. Consequently, the scope of TDM exceptions is limited to acts of reproduction (for copyright subject matter) and extraction (for the sui generis database right) only.

The EU approach on protectable subject matter and exclusive rights arising in the course of TDM is evolving and changing even at the time of this work's composure. Until recently, academics and industries have theorized whether TDM under the new CDSMD is possibly applicable for GenAI training. The 2024 AI Act answered these presumptions affirmatively: "The development and training of such [General-purpose AI] models require access to vast amounts of text, images, videos and other data. Text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content, which may be protected by copyright and related rights."[59] *LAION* case is the first EU-level judicial benchmark on AI training/TDM and copyright, which has established that reproductions of works (including plaintiff's photograph) performed by the defendant (LAION database) were covered by the TDM exception for scientific purposes:"[the] creation of a dataset as the basis for training AI systems can be considered scientific research as it is a fundamental step for future knowledge generation".[60] Simultaneously, both the CDSMD and the new AI Act seem to have a cumulative effect of pilling up regulations without addressing the fundamental approach of the EU legislator. Whether this strategy is prospective and shielding for rightholders or deeply flawed is left for later analysis.

Overall, the ultimate question of whether TDM could be considered a copyright or database sui generis right infringement within the EU framework can be answered in two ways. From the standpoint of legal positivism, that is, the aforementioned provisions of EU Directives, the acts of extraction or reproduction as part of the TDM process are copyright-relevant and constitute an infringement of exclusive rights, unless exceptions apply. Ergo, under the copyright regime, the subject matter would be literary and artistic works,

[59] Recital 105 AI Act.

[60] *Kneschke v. LAION* (n 31) section 3 (a), page 18.

protected by the exclusive right of reproduction; and concerning the sui generis regime, the subject matter would be original or non-original databases protected by the exclusive right of extraction. Furthermore, a compelling point made by Rosati is that if no copyright-relevant act was undertaken, then the provisions introduced by legislatures over the past several years – ranging from Japan, to Singapore, from the UK to other EU Member States – would have been unnecessary, if not altogether misleading. The fact that legislatures have deemed such exceptions necessary indicates that TDM activities entail the doing of copyright-relevant acts.[61] Another battling viewpoint is that foundational concepts of copyright law protect the original expression of ideas, not ideas themselves, nor mere facts or data, so "TDM should not be considered a copyright infringement, but a matter external to copyright's scope."[62]

2.1.2. Text and Data Mining Exceptions and Limitations

Generally speaking, the concept of exceptions and limitations[63] is primarily regulated at the EU level by the InfoSoc Directive. An exhaustive list of (20) twenty optional limitations to the otherwise exclusive rights of reproduction, communication to the public, and distribution is provided in Art. 5(2)–(4). Furthermore, the Software and Database Directives incorporate several optional and mandatory limitations.[64] The Orphan Works Directive introduces a mandatory limitation for specific uses of orphan works, which may be subject to fair compensation.[65] The Marrakesh Treaty Directive establishes a mandatory limitation for certain uses of accessible format copies, benefiting individuals who are blind, visually impaired, or otherwise print-disabled, with limited options for implementing compensation schemes.[66]

[61] Rosati (n 49) 6.

[62] Margoni and Kretschmer (n 57) 700. *See* also Sag (n 19) 31: "The ECJ's approach to copyright infringement implies that whatever the smallest identifiable quanta of creativity or authorship in a work might be, that quanta should be protected from reproduction as though it were a separate work and not just a small part of some larger work."

[63] To note, the terms "exceptions" and "limitations" are used interchangeably. As explained by Quintais, the EU legislature adopts said language as a compromise between different legal traditions. *See* Quintais, J. Copyright in the Age of Online Access: Alternative Compensation Systems in EU Law (Alphen aan den Rijn: Kluwer Law International, 2017), pp. 191-197. Also, according to the WTO Panel, the notions of 'exceptions' and 'limitations' overlap in part in the sense that an *exception* refers to a derogation from an exclusive right provided under national legislation in some respect, while a *limitation* refers to a reduction of such a right to a certain extent. *See* Wymeersch, P. (2023). EU Copyright Exceptions and Limitations and the Three-Step Test: One Step Forward, Two Steps Back. *GRUR International*, *72*(7), at p. 633.

[64] Art. 6(1), 2(d) Database Directive; Art. 5(1)-(3), Art. 6 Software Directive.

[65] Art. 6 Directive 2012/28/EU of the European Parliament and of the Council of 25 October 2012 on certain permitted uses of orphan works (Orphan Works Directive).

[66] Art. 3 Directive (EU) 2017/1564 of the European Parliament and of the Council of 13 September 2017 on certain permitted uses of certain works and other subject matter protected by copyright and related rights for the benefit of persons who are blind, visually impaired or otherwise print-disabled and amending Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society (Marrakesh Treaty Directive).

The list of EU acquis on exceptions and limitations is thus supplemented by Title II of the new CDSMD. Notably, compared to the InfoSoc Directive regime, these are *mandatory* exceptions, which, with the exception of Art. 4 CDSMD, cannot be overridden by contract.[67] Some argue that a normative interpretation of the reproduction right limits its application to exploitative uses of the work as a work. Hence, "such [non-exploitative] use as a work does not exist in the case of TDM."[68] Nevertheless, EU lawmakers determined that an explicit TDM exception was necessary to ensure legal certainty; as provided by Recital 8 CDSMD: "Such [TDM] technologies benefit universities and other research organisations, […]. However, in the Union, such organisations and institutions are confronted with legal uncertainty as to the extent to which they can perform text and data mining of content." Since the TDM exception or limitations set in Arts. 3 and 4 CDSMD are at the core of this research for the EU part, it is of value to provide their brief cross-acquis analysis.

Text and data mining for the purposes of scientific research (Art. 3 CDSMD)

As remarked earlier, Art. 3 CDSMD furnishes an exception for acts of TDM for the purposes of scientific research by research organisations and cultural heritage institutions, concerning works or subject matter to which they have lawful access. The legal uncertainty is thus "[…] addressed by providing for a mandatory exception for universities and other research organisations, as well as for cultural heritage institutions, to the exclusive right of reproduction and to the right to prevent extraction from a database."[69]

Various limitations are imposed on this particular TDM exception. Users of works are defined by CDSMD as follows: (1) "ʹresearch organisationʹ means a university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research; (3) ʹcultural heritage institutionʹ means a publicly accessible library or museum, an archive or a film or audio heritage institution." Interestingly so, this list excludes, for example, public broadcasting organizations and commercial research institutes, from the scope of Art. 3, but they might still find solace in Art. 4 CDSMD.[70]

[67] Art. 7(1) CDSMD.

[68] Ducato, R., Strowel, A. (2019). Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility." IIC International Review of Intellectual Property and Competition Law, 50(6), 649–684. https://doi.org/10.1007/s40319-019-00833-w, at p. 667.

[69] Recital 11 CDSMD.

[70] Hugenholtz, P. B. (2019). The New Copyright Directive: Text and Data Mining (Articles 3 and 4). *Kluwer Copyright Blog*. Available at: https://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ (Accessed: 15 November 2024).

Importantly, with respect to research organisations, it is indicated that respectful activities should be performed on a non-for-profit basis or pursuant of a public interest mission recognised by the Member State. At the same time, the cited Recital 11 CDSMD articulates that "research organisations should also benefit from such an exception when their research activities are carried out in the framework of public-private partnerships."

A compulsory requirement is that the TDM of works or other subject matter may only be conducted when lawful access is secured.[71] The latter "should be understood as covering access to content based on an open access policy or through contractual arrangements between rightholders and research organisations or cultural heritage institutions, such as subscriptions, or through other lawful means [also covering access to content that is freely available online, e.g. public domain]."[72] Parts 2 and 3 of Art. 3 CDSMD strike to provide somewhat of a balance between the interests of users of works and the rightholders. Since research organisations and cultural heritage institutions could in certain TDM cases, for example, to verify scientific research results, need to retain copies, those should be stored in a secure environment.[73] Simultaneously, "rightholders shall be allowed to apply measures (for instance, through IP address validation or user authentication) to ensure the security and integrity of the networks and databases" given a potentially high number of access requests to, and downloads of, their works or other subject matter.[74] Considering the minimal potential damage (if any) that rightholders might incur from this exemption, Member States are not required to implement financial compensation measures.[75] This is relevant to the issue of private ordering and fair remuneration of authors. Once again, compared to Art. 4 CDSMD, in the discussed case of Art. 3 CDSMD, it is not possible to opt-out from TDM, nor is contractual overridability allowed: "Any contractual provision contrary to the exceptions provided for in Articles 3, 5 and 6 shall be unenforceable." Lastly, since the InfoSoc Directive also establishes an analogous but optional exception in Art. 5(3)(a) (concerning use for the sole purpose of illustration for teaching or scientific research), Art. 3 CDSMD will have to be articulated with it.[76]

[71] Art. 3(1) CDSMD.

[72] Recital 14 CDSMD.

[73] Art. 3(2), Recital 15 CDSMD. However, "uses for the purpose of scientific research, *other than text and data mining*, such as scientific peer review and joint research, should remain covered, where applicable, by the exception or limitation provided for in Article 5(3)(a) of Directive 2001/29/EC."

[74] Art. 3(3), Recital 16 CDSMD.

[75] Recital 17 CDSMD.

[76] Recital 10 CDSMD.

Exception or limitation for text and data mining (Art. 4 CDSMD)

Following the examined exception or limitation for the purposes of scientific research, Art. 4 CDSMD provides clarity and, as such, addresses acts that may not meet the conditions of the temporary and transient copy exception in Art. 5(1) InfoSoc Directive, meaning extractions of lawfully accessed works or subject matter for the purposes of TDM, made by any type of beneficiaries, beyond just research organisations or cultural heritage institutions.

In justifying the supplement of Art. 4 to the CDSMD, Recital 18 mentions that "In addition to their significance in the context of scientific research, text and data mining techniques are widely used both by private and public entities to analyse large amounts of data in different areas of life and for various purposes, including for government services, complex business decisions and the development of new applications or technologies." Reproductions and extractions made in such cases may be retained for as long as is necessary for the purposes of TDM.[77] Fundamentally distancing itself from Art. 3, this broader TDM exception can be opted-out or overridden by the rightholders: "[it] shall apply on condition that the use of works and other subject matter has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online,"[78] including metadata and terms and conditions of a website or a service.[79] Thus, if the content has been made publicly available online, it should only be considered appropriate to reserve those rights by the use of machine-readable means. In other cases, it can be appropriate to reserve the rights by other means, such as contractual agreements or a unilateral declaration. In essence, Art. 4 CDSMD allows rightholders to effectively prevent TDM for commercial uses by incorporating machine-readable metadata, such as robot.txt files, into their online content. One supported conclusion is that this gives rightholders considerable control of licensing or even entirely prohibiting TDM and, as such, CDSMD might legitimize a derivative market for TDM/AI.[80]

Speaking of opt-outs, the *LAION* case has recently shed some light on their sufficiency. To be precise, the German Court did not make a final decision on whether a natural language opt-out, as opposed to a standardized machine-readable format like the

[77] Art. 4(2) CDSMD.

[78] Art. 4(3) CDSMD.

[79] Recital 18 CDSMD.

[80] Hugenholtz (n 70).

Robot Exclusion Protocol, met the criteria of being sufficiently 'machine-readable' to satisfy the Art. 4(3) CDSMD exception for TDM. In this case, the plaintiff reserved his rights though Bigstock.com's Terms of Service (ToS): "YOU MAY NOT […] Use automated programs, applets, bots or the like to access the Bigstock.com website or any content thereon for any purpose, including, by way of example only, downloading Content, indexing, scraping or caching any content on the website."[81] However, the Court expressly leaned towards agreeing with Mr. Kneschke that the natural language opt-out would indeed suffice[82] because it clearly excluded the use of bots 'for any purpose.'[83] This reading of machine-readability in opt-outs by the Court can cause confusion. First, the standardized machine-readable format (XML, JSON, CSV) is adopted as a requirement by Art. 4 CDSMD to make it easily understood by automated web-clawless when, simply put, they are allowed to scrape data or is said activity reserved. However, this is complicated when there is a [non-readable] digital plain text present, like in this case contained within the ToS. At the same time, the plaintiff's argument here seems also valid, namely that requiring the use of specific formats is undesirable because most authors do not have the technical knowledge to effectively protect their works from being crawled.[84] Second, when it comes to the language itself – "YOU MAY NOT" – does this imply that all types of general statements (for example, "all rights reserved") are to be interpreted as a reservation of rights under Art. 4(3) CDSMD? Unfortunately, the Court did not expressly explain these observations. Furthermore, it concluded the importance of case-by-case assessments with reference to Art. 53(1)(c) AI Act. The latter institutes that "providers of general-purpose AI models shall (c) put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790." As stated by the Court, "these "state-of-the-art technologies" unambiguously include AI applications that are capable of recognising the content of text written in *natural language*," discussed in detail further.[85]

[81] Part III(18) of Bigstock ToS available at: https://www.bigstockphoto.com/usage.html.

[82] Beckmann, C. et al. (2024). First court decision on text and data mining copyright exceptions: Kneschke v LAION e.V. *Bristows*. Available at: https://inquisitiveminds.bristows.com/post/102jmd3/first-court-decision-on-text-and-data-mining-copyright-exceptions-kneschke-v-lai. (Accessed: 16 November 2024).

[83] Keller, P. (2024). Machine readable or not? – notes on the hearing in LAION e.v. vs Kneschke. *Kluwer Copyright Blog*. Available at: https://copyrightblog.kluweriplaw.com/2024/07/22/machine-readable-or-not-notes-on-the-hearing-in-laion-e-v-vs-kneschke/. (Accessed: 16 November 2024).

[84] ibid.

[85] *Kneschke v. LAION* (n 31) section 2 b) (4), at p. 16.

Interim Conclusion

To sum up, Art. 3 CDSMD establishes a mandatory exception that allows research organizations and cultural heritage institutions to reproduce and extract content for TDM in the context of scientific research. This exception applies to both copyrighted works and databases protected under the sui generis database right. In contrast, Art. 4 CDSMD offers a broader scope, permitting any type of user to engage in various uses of the content. However, it includes a significant provision that allows rightholders to expressly reserve their rights, enabling them to override this exception through an 'opt-out' or 'contract-out' mechanism.

Apart from evident differences, described exceptions for TDM carry vital overlapping traits. First, the three-step test should be mentioned. It pertains to the new TDM exceptions according to Art. 7(2) CDSMD: "Article 5(5) of Directive 2001/29/EC shall apply to the exceptions and limitations provided for under this Title." Thus, said exceptions and limitations "shall only be applied in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightholder."[86] As stated in Recital 6 CDSMD, the Directive makes an effort to pass the exceptions and limitations in compliance with the three-step test. This is evidenced by, for example, the same opt-out option because it is given to rightsholders to balance their interests and those of the user.[87] In the *Mircom* case, the CJEU noted that adhering to the fair balance principle helps fulfil the goal of EU copyright legislation to provide strong protection for rightholders and ensure they receive appropriate compensation for the use of their copyrighted works or other protected material.[88] As previously noted, the three-step test has an overly stringent understanding EU-wide. Yet considerable academia suggests that these criteria are more flexible than conventionally portrayed, potentially allowing for greater latitude in implementing copyright reforms that promote accessibility.[89]

[86] Art. 5(5) InfoSoc Directive.

[87] Guadamuz (n 30) 120.

[88] *Mircom*, [CJEU], Nr. C-597/19, [17.06.2021]. EU:C:2021:492, at para 58. More on the three-step test *see* Rosati, E. (2023). No step-free copyright exceptions: The role of the three-step in defining permitted uses of protected content (including TDM for AI-training purposes). *Stockholm Faculty of Law*, 1-22.

[89] E.g. *see* Geiger, C. et al. (2014) The Three Step Test Revisited: How to Use the Test's Flexibility in National Copyright Law. *American University International Law Review 581.*

Another intersecting matter is the technological protection measures (TPMs) anti-circumvention provisions that would also apply to all new exceptions.[90] TPMs are defined in Art. 6(3) InfoSoc Directive as "any technology, device or component that, in the normal course of its operation, is designed to prevent or restrict acts, in respect of works or other subject-matter, which are not authorised by the rightholder […]." Those can be encryption, password protection, IP address validation or user authentication. For example, Spotify is a prominent example of a website that deploys TPMs, *inter alia,* through a Digital Rights Management system by preventing direct downloading of music tracks, limiting offline listening to premium subscribers and encrypting audio streams. With respect to TDM activities, measures to protect the security and integrity of networks and databases might allow rightholders to block access for researchers trying to conduct TDM. To note, Recital 16 CDSMD emphasizes "those measures should remain proportionate to the risks involved, and should not exceed what is necessary to pursue the objective of ensuring the security and integrity of the system and should not undermine the effective application of the exception." Nonetheless, the application of anti-circumvention provisions might trample over users' privileged uses.[91] As such, in trying to achieve proportionality, EU acquis (with stress on Art. 6(4) InfoSoc Directive) "creates an obligation to provide the means to exercise a limitation, [yet] this obligation is imposed on rights owners and does not give users any authority to perform the act of circumvention themselves."[92] Therefore, users of works find themselves in a tough spot when it comes to overcoming TPMs – compared to the prohibition of contractual overridability outlined in Art. 3 CDSMD, there is no equivalent rule in the Directive safeguarding the enjoyment of the exception when it comes to technological overridability.

2.1.3. The AI Act vs EU Copyright Law Dynamics

The current subchapter goes further into the discussion of GenAI within EU copyright law and, importantly, the new AI Act. As speculated in Part I of this work concerning the TDM procedure, the GPAI Model[93] lifecycle encompasses multiple stages in which copyright considerations may come into play as well. Therefore, the input (or training) stage, model itself and output stage (generated or assisted by AI) are considered. At length, it is then

[90] Art. 7(2) CDSMD: "[…] The first, third and fifth subparagraphs of Article 6(4) of Directive 2001/29/EC shall apply to Articles 3 to 6 of this Directive."

[91] Geiger et al. (2019). Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU, at p. 35.

[92] Guibault et al. (2007). Study on the Implementation and Effect in Member States' Laws of Directive 2001/29/EC on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society, at p. 106.

[93] GP(general-purpose)AI and the difference between a model and a system according to the AI Act are explained further.

examined how the AI Act responds to TDM activities used in GenAI training, rights and limitations imposed on the concerned parties, and the Regulation's interrelation with CDSMD.

To begin with, by now it is quite evident that TDM plays a pivotal role in driving advancements and innovations in AI. Some AI companies, like Midjourney, refrain from disclosing how they train their GenAI models,[94] but it is universal that every GenAI product mines large datasets of content – sometimes tens of datasets. For example, reportedly in the training of GPT-3, OpenAI crawled datasets created by others, with 16% of the training data coming from independent sources – Books1 and Books – comprising 12 billion and 55 billion tokens, respectively.[95] These large numbers suggest how practically unreachable it is to individually process each dataset and encode these models. Given that GenAI models are generated through an automated process of training, and their outputs are then stored in a latent space, they are also known to have a *transformer architecture.* If you type "Hello, how are", the model used in your phone's keyboard suggests words such as "you", or "your" as the next word. However, these words usually do not link to anything with a meaning. Transformers, on the other hand, keep track of the context, so they have the power to generate an entire coherent text.[96] The resulting AI model thus comprises two key components: 1) a run file defining the model's architecture and functionality and 2) a substantially larger file containing parameters/weights,[97] or as described by Martin Andreson, "the ultimate 'gold' that emerges after weeks or even months of training a system."[98]

The input stage, *inter alia*, in terms of this work, is integral, as it directly relates to TDM. The CDSMD broadcasts this by providing a very broad definition of TDM in Art. 2(2), essentially presenting "a tool able to analyse autonomously or semi-autonomously vast amounts of data," which covers most areas of AI/Machine Learning.[99] The mined datasets can have diverse compositions, encompassing full works (images,

[94] Rose, J. (2022) Inside Midjourney, The Generative Art AI That Rivals DALL-E. Vice. Available at: https://www.vice.com/en/article/inside-midjourney-the-generative-art-ai-that-rivals-dall-e/. (Accessed: 18 November 2024).
[95] Guadamuz (n 30) 112.
[96] Amanatullah. (2023). Transformer Architecture explained. *Medium*. Available at: https://medium.com/@amanatulla1606/transformer-architecture-explained-2c49e2257b4c. (Accessed: 18 November 2024).
[97] Sousa e Silva, N. (2024). Are AI models' weights protected databases? Kluwer Copyright Blog. Available at: https://copyrightblog.kluweriplaw.com/2024/01/18/are-ai-models-weights-protected-databases/. (Accessed: 18 November 2024).
[98] Andreson, M. (2023). Weights in Machine Learning. *Metaphysic*. Available at: https://blog.metaphysic.ai/weights-in-machine-learning/. (Accessed: 18 November 2024).
[99] Margoni and Kretschmer (n 57) 687.

videos, texts, and music) or only links to data (such as the LAION dataset) from which a copy might be stored permanently or temporarily. Regarding the resulting AI model, the primary legal issue arising pertains to whether the model weights can be classified as protected databases. In short, it is mostly agreed that there is sufficient capacity to classify the contents of a model weights file as a database.[100] This is supported by the extensive interpretation of a database established in the *Verlag Esterbauer* case. Therein, the Court stated that "the autonomous informative value of material which has been extracted from a collection must be assessed in the light of the value of the information not for a typical user of the collection concerned, but for each third party interested by the extracted material."[101] Therefore, it was determined a topographical map can qualify as a protected database. Given that AI model weights carry value for several parties, it can be speculated that these can also be protected databases and, thus, enjoy protection from extraction or re-utilization of the whole or a substantial part of the database under the Database Directive. At the same time, regarding what constitutes a database, it is questionable whether model weights have a systematic or methodical arrangement to be considered a database. Therefore, a case-by-case analysis would likely be required. From the output perspective, key legal questions encompass 1) the copyright eligibility of AI-generated outputs; 2) granted copyrightable – the applicability of copyright exceptions therein (e.g., for caricature, criticism, caricature, etc. under Art. 5(3) InfoSoc Directive); 3) otherwise, if AI-generated outputs can potentially classify as derivative works; 4) and can they infringe on third-party materials used in the training process, or as otherwise prompted by Rosati, if AI-generated output, which in technical terms is a *'plagiaristic output,'* might be regarded as an actionable reproduction and who would be prima facie liable for said acts of reproduction: user/developer/provider.[102] Although it is clear that the question of outputs (and models) is vital, as well as heavily overlooked compared to other issues linked to the development and use of LLMs, since this study has TDM as its objective, i.e. the input phase *per se*, all output-related concerns are not examined in depth but only those concerning the training data.[103]

[100] Sousa e Silva (n 97).

[101] *Freistaat Bayern v. Verlag Esterbauer GmbH*, [CJEU], Nr. C-490/14, [29.10.2015], EU:C:2015:735.

[102] Rosati (n 49) 5.

[103] Meaning, for example, the issue of copyrightability of AI-generated outputs is quite complex and requires a separate study. However, further analysis will show that from the US perspective, drawing a line between inputs and outputs is essential to identify what kind of use (expressive/non-expressive) was performed on the copyrighted material and thus if it was fair use.

It is worth emphasizing that a substantial body of academia and other experts support the claim that the role of inputs/outputs is heavily misjudged. This essentially stems from the idea of a 'latent space.' For the sake of simplicity, latent space is basically a compressed representation of the original data where each dimension corresponds to a specific feature or characteristic.[104] This can be understood through a plain example of organizing books in a library. Through some techniques – in this case, simplification – instead of describing each book by its full content, it is represented with a few key features like genre, length, and reading level. "Harry Potter" then might be fantasy, medium-length, young adult and so forth. Therefore, it is claimed that inputs are not copied, only the accumulated representation of items is extracted, and the process of generating outputs relies on the statistics that a trained model contains because the trained data *per se* is no longer required.[105] While this reasoning appears logical, it cannot be asserted that this methodology applies to all TDM examples, at the very least. Especially AI-wise, when dealing with enormous datasets used to train LLMs, copyright-protected works can realistically slip (or be put intentionally) into the entry and then reproduced. After all, if no copyright-relevant act was undertaken, the adoption of TDM-specific cross-national exceptions would not be sensible. The 'pro-TDM' arguments presented by Guadamuz, Margoni, Kretschmer and others appear to originate from two primary factors. First, the language chosen by legislators is flawed and almost inviting for criticism – "text and data mining" – referring not to copyrightable works but simple text and data that are *by default* not protected. Second, restrictions on TDM simultaneously impede the development of unbiased AI models, which are trained on the 'high-friction' copyrighted content and not the freely accessible public domain. This aspect pertains to the concept of fair balance which is further explored in Part III of this study.

The AI Act

The upcoming segment explores the newly implemented AI Act and how its provisions interact with previously examined EU copyright legislation. It also identifies potential ambiguities and overlooked aspects of this regulation that impact the EU acquis.

The AI Act is a highly complex piece of legislation, containing 68 definitions, 113 articles, 13 annexes and 180 recitals. Moreover, it can be characterized as an example of

[104] AI Maverick (2023). A Comprehensive Guide to Latent Space. Available at: https://samanemami.medium.com/a-comprehensive-guide-to-latent-space-9ae7f72bdb2f. (Accessed: 18 November 2024).
[105] Guadamuz (n 30) 115.

the so-called “regulatory brutality” drift[106] with (1) severe penalties up to EUR 35 000 000 or, if the offender is an undertaking, up to 7 % of its total worldwide annual turnover for the preceding financial year, whichever is higher;[107] (2) an extraterritoriality effect: the “policies” obligation should apply even if the relevant TDM takes place outside the EU;[108] (3) the new supervision established at national and EU level, including the EU AI Office,[109] the EU AI Board,[110] an advisory forum and a scientific panel of independent experts.[111]

The definition of 'AI' within the AI Regulation warrants some scrutiny. Following Art. 3(1) AI Act, “'AI system' means a machine-based system that is designed to operate with varying levels of *autonomy* and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, *infers*, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.” Foremost, although this definition is fairly broad, it also implies specific characteristics (e.g., autonomy and inferences) that would, for instance, exclude “simple automations, formulas, static software or totally deterministic programming (if x, then y).”[112] Thus, based on Recital 12, “the definition should be based on key characteristics of AI systems that distinguish it from simpler traditional software systems or programming approaches and should not cover systems that are based on the rules defined solely by natural persons to automatically execute operations.” Secondly, the AI Act is quite tangled due to its intertwining references to models and systems. To clarify, most copyright obligations are imposed specifically on *GPAI model providers* (parties placing the object on the market), not on GPAI/AI systems. The Recital 97 AI Act clarifies this distinction: “Although AI models are essential components of AI systems, they do not constitute AI systems on their own. AI models require the addition of further components, such as for example a user interface, to become AI systems. AI models are typically integrated into and form part of AI systems.”[113] Hence, GPAI models are part of (GP)AI systems, basically their “engine.” To illustrate, models are generative pre-trained transformers or GPTs, while the systems are ChatGPT, Midjourney, Dall-E, or Firefly.[114]

[106] V. Papakonstantinou/Paul De Hert (2022). The Regulation of Digital Technologies in the EU: The law-making phenomena of "act-ification", "GDPR mimesis" and "EU law brutality". *Technology and Regulation*, 48-60.
[107] Art. 99(3) AI Act.
[108] Recital 106 AI Act.
[109] Art. 64 AI Act.
[110] Art. 65 AI Act.
[111] Arts. 67, 68 AI Act. Regarding the respective national competent authorities – Art. 70 AI Act.
[112] Sousa e Silva, N. (2024). The Artificial Intelligence Act: Critical Overview, at p. 10.
[113] Recital 97 AI Act.
[114] Quintais (n 34) 8.

While it may appear that those 'usual systems,' i.e., the widely used AI tools, exert greater influence due to their direct impact on users, the approach of the EU regulator towards targeting models is arguably justifiable. That is, GPAI models enable versatile content generation in formats such as text, audio, images, and video, easily adapting to various tasks, which underlines their transformative nature. Furthermore, a single model can create multiple systems with different applications, purposes, and operational methods. To paraphrase, GPAI models have the potential to exert a significantly wider impact and reach than systems.

As mentioned before, the AI Act has now settled for good the controversy whether Arts. 3 and 4 CDSMD intended to include TDM exceptions to activities related to the development of GenAI. The affirmative answer can be traced to the Recital 105 AI Act, which states that large generative AI models present opportunities but also challenges to artists, authors, and other creators and the way their creative content is created, distributed, used and consumed. In particular, in AI training, TDM techniques may be used extensively on protected content. If TDM is performed, the authorisation of the rightsholder should be obtained, unless relevant copyright exceptions and limitations under CDSMD apply. Notably, however, the AI Act recognizes the relevance of TDM for AI training, but it does not indicate that TDM is synonymous with AI training or that everything in-between TDM and AI training is covered by Arts. 3 and 4 CDSMD.[115] This “in-between” includes ‘upstream players’ like the LAION dataset, Common Craw[116] or ‘downstream players’ such as AI systems providers or deployers.[117] When these upstream players perform TDM they are not subject to the AI Act’s copyright-relevant obligations but just to CDSMD requirements. However, it is not as black&white with the downstream players, necessitating future elucidation from the CJEU, since a model and a system can possibly constitute a single entity.

From the copyright standpoint, the most significant provisions regarding GPAI models are found in Chapter V of the AI Act. Therein, the provision in Art. 53(1)(c) and (d) should be given substantial weight. Art. 53(1)(c) establishes extraterritoriality on compliance with Art. 4(3) CDSMD, regardless of the geographical location where the acts

[115] Rosati (n 49) 7.

[116] A website for web scraping, *see* at: https://commoncrawl.org.

[117] Quintais, J. P. (2024). Copyright, the AI Act and extraterritoriality. *Kluwer Copyright Blog*. Available at: https://copyrightblog.kluweriplaw.com/2024/11/28/copyright-the-ai-act-and-extraterritoriality/. (Accessed: 19 November 2024).

of extraction and reproduction for TDM purposes occur, if the AI model is made available within the EU. As such, "Providers of general-purpose AI models shall (c) put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790." As clarified in Recital 106, this regulation is set forth to ensure a level playing field among providers of general-purpose AI models. An interesting discussion has been raised about the 'non-retroactivity' of this regulation. While some commentators[118] conclude that the obligation under Art. 53(1)(c) will *not* apply to AI models trained outside the EU and made available in the EU *before* the entry into force of the AI Regulation, and others[119] take an opposing stance, the reality is somewhere in-between. In principle – yes, non-retroactivity suggests that GPAI models introduced to the EU market before the AI Act's implementation are exempt from its regulatory requirements. However, if a GPAI model, available on the EU market before the AI Act, undergoes some modifications/updates/fine-tuning after the Act's release, its provisions (including copyright-related) could apply to those effects. Also, new GPAI models will be required to comply with the Regulation from 2 August 2025; and those placed on the market before 2 August 2025 only from 2 August 2027.[120] The extraterritoriality reach of Art. 53(1)(c) and Recital 106 AI Act does not extend to the relevant TDM activities conducted to *pre-train* and *train* the GPAI model outside the EU. However, to agree with Quintais,[121] in instances where TDM activities, particularly web scraping, have a discernible connection to EU territory, model providers should be obligated to adhere to EU copyright law, including compliance with the opt-out provision. The second obligation on transparency is put in Art. 53(1)(d) AI Act for providers of GPAI models to draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office. Recital 107 of the Act essentially explains the ample relevance this provision has in terms of the TDM exception of Art. 4 CDSMD. Accordingly, the mentioned summary should be generally comprehensive in its scope to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, 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. While the EU AI Office

[118] E.g. *see* Peukert, A. (2024). Copyright in the Artificial Intelligence Act – A primer. 73(6) GRUR.

[119] E.g. *see* Rosati (n 49).

[120] Arts. 111(3) and 113(b) AI Act.

[121] Quintais (n 117).

takes shape and provides the promised template, some sources[122] have already established a blueprint for the transparency template. Its content requirements are quite promising, *inter alia*, for TDM's transparency, including (1) the overall size of the training data; (2) a list of all used data sets and sources in training the model; and, importantly, within this part information on the sources from which data was obtained (scraped from the internet, copyright-protected content licensed from rightholders or third-party intermediaries, obtained from proprietary databases, etc.); (3) data diversity compliance (important to ensure unbiased models); and (4) description of (pre-)processing steps applied. Altogether, the above can help assess if relevant TDM activities occurred – or have sufficient connections – within EU territory, and whether in terms of Art. 4 CDSMD the sources were accessed lawfully, the reservations of rightholders through opt-outs/contract-outs were respected or, essentially, copyright infringement took place.

As evidenced by now, the EU copyright acquis, particularly the CDSMD, exhibits a close interrelationship with the AI Act. Through regulation of GPAI models, the AI Act also addresses the TDM exception of Art. 4 CDSMD. At the same time, whether this interface is rational in terms of private vs public law division can be questioned. As pointed out by Peukert, "Art. 53(1)(c) and (d) AI Act merge two different types of laws."[123] Although copyrights and related rights are inherently private in nature, the AI Act's approach to addressing copyright-related challenges diverges from conventional copyright law methodologies. Mainly, with public interest as an objective. To exemplify, Recital 3 explains untrustworthy AI systems should be prevented by laying down obligations and guaranteeing the uniform protection of overriding reasons of *public interest* and rights of persons. As such, AI may generate risks and cause harm to *public interests* (Recital 5), therefore, a high level of protection of *public interests* should be established (Recital 7). Furthermore, the establishment of public supervisory bodies, like the EU AI Office, to oversee compliance with obligations imposed on GPAI model providers, supports the approach of the AI Act towards public regulatory oversight, rather than individual copyright enforcement mechanisms. In enumerating the areas to which the AI Regulation is not applicable, Art. 2 omits any reference to Union copyright law. However, stemming from Recital 108, "This Regulation does not affect the enforcement of copyright rules as provided for under Union law." Meaning, that when a GPAI model provider fails to comply

[122] Warso, Z. et al. (2024) Blueprint of the Template for the Summary of Content Used to Train General-Purpose AI Models (Article 53(1)d AIA) – v.2.0. *Open Future Foundation.* Available at: https://openfuture.eu/blog/sufficiently-detailed-summary-v2-0-of-the-blueprint-for-gpai-training-data/. (Accessed: 20 November 2024).
[123] Peukert (n 118) 4.

with the copyright obligations of the AI Act, this does not simultaneously mean a copyright infringement under CDSMD. However, the Act's obligations are complementary to CDSMD and might result in a "regulatory spill-over" between public law and private law.[124] Simply put, if a GPAI model provider chooses to comply with the AI Act-relevant copyright obligations, it can be expected to also comply with Art. 4 CDSMD lawful access (transparency requirement) and opt-out (policies requirement) obligations. Time will show how this evolving dynamic is interpreted in the future jurisprudence from CJEU.

Sub-conclusion

In summary, the CDSMD has represented a pivotal revolution in EU copyright law on TDM by introducing mandatory exceptions and limitations to respective rights in Title II. As such, Art. 3 CDSMD established a mandatory exception that allows research organizations and cultural heritage institutions to reproduce and extract lawfully accessed content for TDM in the context of scientific research. Art. 4 CDSMD broadens this scope, permitting any type of beneficiary to engage in TDM, unless rightholders expressly reserve their rights to this. These provisions reflect the recognition by the EU regulator that TDM potentially impacts two intellectual property regimes—copyright law and sui generis database law—and their respective subject matter: literary and artistic works protected by the exclusive right of reproduction, and original or non-original databases protected by the exclusive right of extraction. Within EU discussions, an opposing viewpoint argues that this approach is faulty as copyright law only safeguards the original expression of ideas, excluding ideas, facts, or data, and thus TDM falls outside copyright infringements. Nonetheless, the newly enacted AI Act further reinforces the EU stance on TDM, establishing that during AI training, TDM techniques may be used extensively on protected content. The AI Act is specifically complementary to Art. 4 CDSMD, emphasizing transparency and compliance with EU copyright law for GPAI models. It imposes specific obligations on providers of GPAI models in Art. 53(1)(c) and (d), including the implementation of policies to comply with copyright law and related rights, particularly in relation to the opt-outs under Art. 4(3) CDSMD, as well as requiring detailed summaries of the content used to train AI models. The potential strengths and shortcomings of the EU approach to regulating TDM within copyright law can be further explored by comparing it with the US regime, discussed in the following section.

[124] Quintais (n 34) 21.

2.2. US Copyright Law

This chapter examines the US approach to TDM activities within the framework of copyright law, specifically the fair use doctrine contained in Section 107 of the Copyright Act. In this sense, it investigates the extensive US case law surrounding TDM/AI, or to be exact, copy-reliant technology, which has significantly shaped the current legal landscape.

To preface, the US approach to regulating activities of copy-reliant technology, whether considered independently or as a component of Machine Learning, diverges from the EU view on things but is not entirely alien to it. That is, the performance of copying or reproduction acts during TDM-like processes is uncontested to be an infringing activity in the US doctrine.[125] However, since the US regime does not deploy any EU-like TDM exceptions, in the absence of express or implied permission, said infringement or the absence thereof fully relies on the fair use doctrine.

2.2.1. Extent of Text and Data Mining Under the US Fair Use

The US Constitution provides in the intellectual property clause that copyright's purpose is "To promote the Progress of Science and useful Arts, […]."[126] Certainly, this statement does not negate authors' legitimate interests but instead aims to establish a fair equilibrium between the rights of creators and society's competing needs for accessible information, arts, innovation, etc. This 'good for all' attitude towards copyright law treats it principally as a form of public law. In other words, the advancement of *overall public* welfare is prioritized as a normative objective, serving as the driving force behind the promotion of private interests, including those of both authors and users. Such discussion of copyright's fundamentals is particularly important from the US perspective, as it provides a basis for the fair use doctrine and the recognition of TDM's non-expressive nature within it. As such, it seems the EU approach towards discussed issues is more direct or even 'point-blank.' It focuses on the interest of rightholders; the threshold for creativity is rather low, and the

[125] For instance, in Authors Guild v. Google, Inc., No. 13-4829-cv (2d Cir. Oct. 16, 2015), even though found to fall under fair use, it was not disputed that Google performing reproductions of third-party books to then mine Google Books was a *prima facie* infringing activity. Nonetheless, as discussed, TDM does not equal AI training and the latter complicates things. Several academic sources deny that AI training models at any point make even a single copy of the training data. That is, they 'learn' from it and then store this knowledge in a latent space. E.g., Murray takes a very firm approach towards advocating this in Murray, M. D. (2023). Generative AI Art: Copyright Infringement and Fair Use. *SMU Science and Technology Law Review*, *26*(2), 259. https://doi.org/10.25172/smustlr.26.2.4. Essentially, this is a battle of opinions. The approach taken by Murray might be technically true, indeed. But if one nudges the curve a bit, it is also shallow and legally unforeseeable as there is always an element of unsureness. This unsureness particularly rises when AI-generated outputs are direct copies of inputs, which has happened in some cases before and generally represents the issue of *AI memorization,* incompatible with non-expressive use, as discussed in depth below.

[126] Art. I, § 8, cl. 8 U.S. Constitution.

right of reproduction is interpreted very broadly, so even the smallest identifiable quanta of creativity will be protected from reproduction. Therefore, the question of how exactly this quanta is manipulated by the user and whether this use is somehow 'fulfilling' is not discussed. Conversely, this is at the crux of the US approach. In part, it is also highly related to the idea-expression dichotomy. Even though this distinction is followed by most copyright systems, within the US, it is also constitutionalized as one of copyright's "built-in First Amendment accommodations."[127] In *Harper & Row*, the Supreme Court established that copyright's idea/expression dichotomy "strike[s] a definitional balance between the First Amendment and the Copyright Act by permitting free communication of facts while still protecting an author's expression."[128] An idea endorsed by American scholars like Sag[129] and Carroll[130] but criticized by European academics such as Rosati[131] implies that the aforementioned dichotomy confines copyright protection to the expressive elements of an author's work and therefore allows for *non-expressive* copying in the context of TDM and related processes without constituting infringement. This non-expressive copying in terms of training, functioning of algorithms or data analysis, describes the unintentional duplication of data and raw source material for purposes unconnected to producing, consuming, or distributing the expressive elements of the material.[132]

The fair use doctrine originated in the judiciary and was later formalized into a four-factor doctrine, codified in Section 107 of the Copyright Act of 1976. It states that "Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright." Furthermore, "In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include – (1) the purpose and character of the use (does it have a commercial nature or is for nonprofit educational purposes); (2) the nature of the copyrighted work; (3) the amount and substantiality of the

[127] Eldred v. Ashcroft, 537 U.S. 186, 220 (2003).
[128] Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 556 (1985).
[129] Sag (n 19).
[130] Carroll, M. W. (2019). Copyright and the Progress of Science: Why Text and Data Mining Is Lawful. 53 UC Davis Law Review 893, American University, WCL Research Paper No. 2020-15, Available at SSRN: https://ssrn.com/abstract=3531231.
[131] Rosati (n 49) 7: "[…] the claim that copyright would not restrict non-expressive uses of protected content appears erroneous [due to] the broad construction of the right of reproduction and the fact that specific exceptions to the right of reproduction to allow (at certain conditions) TDM have been adopted in multiple jurisdictions over the past several years."
[132] Murray (n 125) 275.

portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work. Henceforth, fair use copying is not simply excused; it falls entirely outside the realm of infringement, eliminating the need for additional licenses or justifications.[133] In *Campbell v. Acuff-Rose*, the Supreme Court noted that these four statutory factors should not be treated as a checklist or a scorecard, they are "to be explored and weighed together in light of copyright's purpose of promoting science and the arts."[134] The four factors individually come into play through different copy-reliant technology/TDM considerations but perhaps the most weighty one in terms of this work is the first factor – 'the purpose and character of the work.'

What the first factor represents is the idea of *transformative use*, widely assessed in several Supreme Court benchmark decisions. Commentators argue that there is no reason that the same decisions cannot apply to TDM activities and AI training models. Thus, copy-reliant technology processing can not only be non-expressive by nature but also further transformative, for example, when exploring AI-generated outputs. The aforementioned *Campbell v. Acuff-Rose* provided a foundation for the copyright fair use transformative test which lies at the heart of the doctrine. Inherently, in finding if there was transformative use, it should be assessed whether the copier's use adds something new, with a further purpose or different character, altering the copyrighted work with new expression, meaning or message. Importantly, in *Google v. Oracle,* the Supreme Court positively applied the test in the context of fair use copying of computer code for a new function and purpose within a different software application.[135] The most recent 2023 *Andy Warhol Foundation v. Goldsmith* case upheld the established doctrine.[136] Therein, the concerned object was Goldsmith's photograph of Prince from 1981. Based on it, Warhol created a series of silkscreen prints and illustrations that were challenged in court as transformative enough to apply for fair use. The court ruled against Warhol, explaining that even though the defendant's works did add new expressions to the original work, most derivative works add new expressions, meaning or message, etc. and to allow all such adaptations and alterations of original content to be transformative fair uses would "swallow the copyright owner's exclusive right to prepare derivative works." Therefore, the use was not transformative – it was easy to notice Goldsmith's depiction of Prince first, while Warhol's adaptation of the work was *not discernible*. This case could mean a lot in terms of TDM, if the question is

[133] Sag (n 19) 14.
[134] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994).
[135] Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1196 (2021).
[136] Andy Warhol Found. for the Visual Arts, Inc., 598 U.S. 508 (2023).

posed likewise. For example, let us consider a hypothetical case scenario where a certain tech company develops an LLM by training it on a vast corpus of literary works, including the entire Harry Potter series by J.K. Rowling. The company does not seek permission from Rowling or her publishers for this use because, according to developers, it does not reproduce the books verbatim but uses the information to generate new ideas, identify patterns, trends and correlations, i.e. the use is transformative. Nonetheless, when prompted to describe the beginning of Harry Potter and the Sorcerer's Stone, said AI tool *inter alia* regurgitates original pieces of the book that are easily discernable in AI's adaptation of the work, as such 'overshadowing' the transformation AI made. Moreover, the AI tool most definitely in this way communicates the original expression. Given the Supreme Court's reasoning in *Andy Warhol Foundation v. Goldsmith* and the parallels to this AI scenario, a ruling in favour of the AI company would be highly improbable.[137] Even more so, this sample is highly probable, being based on real-life instances retrieved by researchers. For instance, when prompted in a certain way,[138] ChatGPT regurgitated the first 3 pages of Harry Potter and the Sorcerer's Stone *verbatim*.[139]

Therefore, the above was used to show that common predictions in favour of AI systems against authors (like in the ongoing suit *Andersen v. Stability AI Ltd.*) might not be as easy to follow in line with the transformative test under fair use. But, of course, this is a specific example and its plausibility, much like the claims made by the plaintiffs in *Andersen v. Stability AI Ltd.* and *Getty Images (US), Inc. v. Stability AI Inc.*, must endure scrutiny under a long history of fair use rulings that support similar types of non-expressive use by technologies reliant on copying.[140] To exemplify, the *Kelly v. Arriba Soft Corp*. case involved search engines that crawl the web to scrape copyrighted images and make copies of those.[141] This activity was found to be transformative fair use because the crawler Arriba Soft Corp. (the defendant) performed a transformation of a sort: it downloaded full-size

[137] On a related note, in *Warner Bros. Entertainment Inc. v. RDR Books*, the district court ruled that the Harry Potter Lexicon, a comprehensive reference guide to the Harry Potter world, infringed the copyright of the original series. The court determined that while the Lexicon's intended use was transformative, its implementation fell short of being truly transformative due to excessive direct copying from Rowling's works.

[138] This element is important because it further extends to the process of AI memorization. In this case, researchers on purpose used one of the most popular books on the Top 100 all time best sellers list. *See* Henderson, P. et al. (2023) Foundation Models and Fair Use. *Stanford Law and Economics*. Paper No. 584.

[139] Henderson and others (n 138) 8.

[140] For instance, copying without permission was recognized as fair use in the context of software reverse engineering in *Sega Enters. v. Accolade, Inc*., 977 F.2d 1510, 1514 (9th Cir. 1992); *Sony Computer Ent. v. Connectix Corp*., 203 F.3d 596, 608 (9th Cir. 2000); plagiarism detection software in *A.V. ex rel. Vanderhye v. iParadigms*, LLC, 562 F.3d 630, 644–45 (4th Cir. 2009); and the digitization of millions of library books facilitating meta-analysis and indexing in *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 100–01 (2d Cir. 2014); *Authors Guild v. Google, Inc.*, 804 F.3d 202, 225 (2d Cir. 2015). *See* Sag (n 42) 304.

[141] *Kelly v. Arriba Soft Corp*., 336 F.3d 811 (9th Cir. 2003).

images to its servers, created thumbnail versions, deleted the originals, and then only displayed the thumbnails in search results. Google's Image Search operates similarly, transforming visual images into a “pointer directing a user to a source of information,” so it “provides social benefit by incorporating an original work into a new work, namely, an electronic reference tool,”[142] rather than reproducing and utilizing images as aesthetic objects for viewing and consumption, meaning *expressive* use.[143]

2.2.2. Likely Copyright Infringement Cases: AI Memorization Impact

AI memorization instances put the fair use arguments, relied upon by model developers/providers, on edge. Specifically, this is an issue with GenAI. As such, training a machine learning model on copyrighted data may be deemed fair use if the model does not directly generate content, but the analysis becomes more complex for foundation models used in *generative* applications since they are designed to produce content, which might as well be similar to copyrighted inputs. A legally-reliant definition of memorization is constructed by Cooper et al., by which they mean a process “when an exact or near-exact copy of a piece of training data can be reconstructed by examining the model through any means.”[144] This is not only incompatible with the non-expressive use but also suggests that the LLMs used in generative AI might need an overall different treatment by their developers and legislators because of their potential for memorization.[145]

To point out a prominent example, in The New York Times copyright complaint against OpenAI and Microsoft, the former argues that OpenAI's GPT models have memorized its articles. Similar to the hypothetical Harry Potter scenario, when given a snippet from a New York Times article, ChatGPT would generate a lengthy continuation that includes copied passages from the article. It is fair to assume alike cases will grow over time since it has now become easy to manipulate GenAI. For example, one study[146] identified 350,000 of the most frequently duplicated images in the training dataset for Stable Diffusion and created 500 new images using prompts that were identical to the original images. This particular example has an important element – ‘frequently duplicated

[142] *Perfect 10, Inc. V. Google Inc*. 508 F.3d 1146 (9th Cir. 2007).
[143] Murray (n 125) 278.
[144] Cooper, A. et al. (2024). The Files are in the Computer: On Copyright, Memorization, and Generative AI. Cornell Legal Studies Research Paper Forthcoming, Chicago-Kent Law Review, Forthcoming, Available at SSRN: https://ssrn.com/abstract=4803118.
[145] Sag (n 42) 302.
[146] Carlini, N. et al. (2023). Extracting Training Data from Diffusion Models. Available at: http://arxiv.org/abs/2301.13188.

images' – which suggests that there are preconditions for 'effective' memorization. For text-to-image models, Sag defines those as (1) the number of duplicates of a work, (2) image association with unique text descriptions and (3) the ratio of model size to training data.[147] In this context, a frequently discussed topic is the known Snoopy problem,[148] Pikachu Paradox,[149] or Italian plumber problem.[150] The search shows that prompting GenAI to make reproductions of famous characters like Snoopy easily provokes copyright infringement. Because Snoopy is such a famous character, it is probable the metadata used for model training contains a high volume of images, descriptions, and so forth mentionings of Snoopy. Hence, the presence of potentially numerous duplicates may compel the model to memorize this character to such an extent that the resulting 'Snoopy output' images, along with the recognized strong copyrightability of Snoopy as a character, will most likely be infringing. The more known, precise and repeated the description of a character is, the easier it is to generate, in terms of copyright, a substantially similar work. However, a case-by-case assessment is still crucial because seemingly every AI-generated work can carry different effects. For example, Guadamuz's AI-generated picture of Mario and Pikachu[151] has equal chances to fall under the transformative fair use or maybe even parody.

GenAI developers are well aware of the memorization issue because it does not happen randomly, but rather, it is programmed into a model during the training stage. Consequently, reducing memorization and thus the possibility of copyright infringement, at least towards most scandalous items, is possible. After the Harry Potter verbatim citations with ChatGPT, it was observed that OpenAI has made some modifications, and even tried to hide that ChatGPT was trained on copyrighted books.[152] At the crux of technology and law, Sag proposes some practices for copyright safety in GenAI that can be a useful guide. Some of those include: (1) not training LLMs on duplicates; (2) considering the larger an LLM is, so is the likelihood of memorization; (3) filtering model outputs; (4) keeping

[147] Sag (n 42) 296.

[148] ibid 327.

[149] Guadamuz, A. (2024). Snoopy, Mario, Pikachu, and reproduction in generative AI. *TechnoLlama*. Available at: https://www.technollama.co.uk/snoopy-mario-pikachu-and-reproduction-in-generative-ai. (Accessed: 1 December 2024).

[150] Lee, T., Grimmelmann, J. (2024). Why The New York Times might win its copyright lawsuit against OpenAI. Available at: https://www.understandingai.org/p/the-ai-community-needs-to-take-copyright. (Accessed: 1 December 2024).

[151] Guadamuz (n 149).

[152] *See* the Reddit post at: https://www.reddit.com/r/books/comments/15xilfs/openai_now_tries_to_hide_that_chatgpt_was_trained/.

detailed records on the obtained copyrighted works; (5) restricting the open-sourcing of LLMs that pose a significant risk of copyright.[153]

Sub-Conclusion

To conclude, under the US copyright approach, TDM activities, which represent a procedure inherently reliant on acts of copying/reproduction, are not regarded to be infringing. The realm of infringement is overmined by the fair use doctrine. As such, copyright-infringing activities performed as part of a TDM process constitute a foremost non-expressive or even further transformative use of input material. Therefore, within the US legal landscape, courts have already established that reproduction activities related to TDM do not constitute copyright infringement. Nevertheless, the applicability of this reasoning, bolstered by an extensive body of case law, to GenAI models exhibiting memorization tendencies remains a subject of uncertainty. As such, the development of GenAI models may transition from fair use to copyright infringement due to LLM's tendency to extensively memorize training data instead of merely extracting knowledge from it, subsequently engaging in an expressive display of original copyrighted works.

2.3. Comparative Analysis of the EU and US Approaches

Based on the provided research, it can be inferred that the copyright protection approach against TDM activities/AI model training is *rights-oriented* within the EU law and *industry-oriented* under US law. One factor contributing to this is the current lack of a global consensus on copyright matters pertaining to Gen(AI) model training. As a result, nations are reverting to the framework for limitations and exceptions set forth by the Berne Convention (Art. 9(2)), which stipulated regarding the right of reproduction that "it shall be a matter for legislation in the countries of the Union to permit the reproduction of such works in certain special cases provided that such reproduction does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author" (three-step test). As such, this discretion provides significant legislative freedom but also introduces notable challenges, particularly the lack of consensus on the rules governing copyright limitations and exceptions across different countries.[154] Of course, just stating that the EU is rights-oriented and the US industry-

[153] Sag (n 42).

[154] Wu, H. et al. (2024). Copyright protection during the training stage of generative AI: Industry-oriented U.S. law, rights-oriented EU law, and fair remuneration rights for generative AI training under the UN's international governance regime for AI. *Computer Law and Security Review*, 55, at p. 2.

oriented does not explain the complexities within both systems. That is, both regimes have setbacks but also reciprocal advantages through their contrasting approaches. This sub-chapter will take a look at those through a comparative lens and with that basis, the forthcoming Part III will examine a pathway regime to balance TDM/AI model training (innovation) with copyright protection (interests of rightholders).

In no particular hierarchy, it can be summarized that the EU regime in question exhibits such notable strengths: (1) through the establishment of TDM exceptions, the CDSMD firmly protects the rights of original creators; (2) the AI Act provides a foundation for transparency of GPAI model training that can avoid copyright infringement cases; (3) as such, the approaches taken by the CDSMD and AI Act strike to balance legitimate interests involved in using copyright-protected works for TDM/AI model training. On the contrary, possible drawbacks are: (1) the CDSMD fails to accommodate the vast data demands for GenAI model training; (2) this causes consecutive issues of developing biased/discriminating GenAI models; (3) the CDSMD promotes an EU copyright law that favours private ordering over public policy,[155] e.g., through the possibility of contractual overridability of TDM resulting in over-licensing; (4) the EU regulation of TDM makes prior authorization the default and a right to access (and use) the exception, going against copyright's social function;[156] (5) overall, the established TDM exceptions can contribute to EU market loss as GenAI model providers seek for more lenient legal landscapes such as the one provided under the US law. Respectively, the US regulatory framework demonstrates the following advantages: (1) the fair use doctrine allows for copy-reliant operations (such as TDM) to use the content without copying the exact way someone expressed it, which facilitates technological innovation while preventing direct copying which produces substantial similarities to original works;[157] (2) therefore, the US copyright regime towards TDM is deemed to be more flexible than EU's TDM exceptions; and (3) as such, it also provides greater scope for training AI and machine learning models than others. Some shortcomings of this legal landscape can be presented as follows: (1) the US fair use doctrine does not adequately consider the economic compensation for copyright holders contributing to innovation; (2) the latter might cause exacerbating tensions within the GenAI industry and also result in the known 'tragedy of the commons,'[158] i.e. the over-

[155] Quintais, J.P. (2019). The New Copyright in the Digital Single Market Directive: A Critical Look. European Intellectual Property Review 2020(1). Available at: https://ssrn.com/abstract=3424770.
[156] Geiger and Jütte (n 26).
[157] Wu and others (n 154) 15.
[158] Conceptualized by Garrett Hardin; means the overuse that occurs when resources are freely available, leading to their destruction.

dissemination & destruction of artistic and scientific knowledge; (3) securing legal protection through fair use of copyrighted materials for TDM/AI model training necessitates navigating a complex and unpredictable framework delineated by four factors – hence, causing legal uncertainty;[159] (4) moreover, fair use loses its grounds when GenAI models memorize training data in a manner incompatible with non-expressive use; (5) lastly, several proposals of best practices for copyright safety in GenAI to be promulgated by the US Copyright Office,[160] indicate that the US regime would significantly benefit from a domestic EU-like AI Act that would facilitate a fairer and more transparent GenAI ecosystem.[161]

Sub-Conclusion

In conclusion, notwithstanding their disparities, both regulatory frameworks could potentially enhance their efficacy by incorporating the aforementioned beneficial attributes of each other. Determining a superior system proves challenging, as they fundamentally prioritize distinct objectives – the EU copyright acquis adopts a rights-oriented approach to TDM and AI model training regulation, whereas the US framework is industry-centric. A shared concern for both jurisdictions is the urgent need for a regulatory equilibrium that balances the fostering of innovation with the safeguarding of the interests of rightholders. This topic is explored in wider depth in the subsequent section of this research.

[159] Fernandes, P. M. (2024). AI Training and Copyright: Should Intellectual Property Law Allow Machines to Learn. *Bioethica*, *10*(2), 8–21. https://doi.org/10.12681/bioeth.39041, at p. 18.

[160] Sag (n 42) 188; ibid 16.

[161] To note, a similar bill was introduced in the US – "Generative AI Copyright Disclosure Act of 2024." *See* available at: https://www.congress.gov/bill/118th-congress/house-bill/7913. For example, it mandates a detailed summary of all copyrighted works used in GenAI systems, imposing a minimum civil penalty of $5,000 for failure to comply.

3. FAIR BALANCE BETWEEN INNOVATION AND INTERESTS OF RIGHTHOLDERS

The imperative to strike a balance between fostering (AI) innovation and safeguarding intellectual property rights has been extensively deliberated in academic literature as well as via official discourse. United Nations (UN) and the World Intellectual Property Organization (WIPO) have been expressing their concerns about GenAI model training and copyright. In particular, in its 2024 Resolution "Enhancing international cooperation on capacity-building of artificial intelligence," the UN General Assembly emphasized, "that Member States should enjoy equal opportunities in the design, development, deployment, decommissioning and use of artificial intelligence, while respecting intellectual property rights and promoting innovation."[162] In its Report "Governing AI for Humanity," the UN Advisory Body on AI noted that most experts *inter alia* were concerned about the risk of AI violations of intellectual property rights – for instance, "profiting from protected intellectual assets without compensating the rights holder."[163]

Within the context of this study on TDM, among the multitude of concerns surrounding AI, the most pertinent issue is biased AI models. As astutely noted by the Obama White House Paper on the future of AI, it is important to focus on AI being produced by and for diverse populations; and "Doing so helps to avoid the negative consequences of narrowly focused AI development, including the risk of biases in developing algorithms, by taking advantage of a broader spectrum of experience, backgrounds, and opinions."[164] Crucially in terms of regulating mining of content, "AI needs good data. If the data is incomplete or biased, AI can exacerbate problems of bias."[165] Over the years, numerous instances of bias in AI have been documented and analyzed. According to a 2022 study by the University of Southern California's Information Sciences Institute, bias was detected in up to 38.6% of 'facts' used by AI.[166] The referred facts, what we call common knowledge, were found to be not fair as they were contributed by ordinary people to the ConceptNET database which is essentially like Wikipedia. Therefore, AI models trained on this common knowledge also exhibited the same biases. Further to this discussion, as argued by

[162] UN. Resolution Adopted by the General Assembly on 1 July 2024: 78/311. Enhancing International Cooperation on Capacity-Building of Artificial Intelligence.

[163] UN (2024). Governing AI for Humanity.

[164] Office Of The President. (2016). Preparing For The Future Of Artificial Intelligence., at p. 28.

[165] ibid 30.

[166] Gruet, M. (2024). That's Just Common Sense. USC researchers find bias in up to 38.6% of "facts" used by AI - USC Viterbi | School of Engineering. Available at: https://viterbischool.usc.edu/news/2022/05/thats-just-common-sense-usc-researchers-find-bias-in-up-to-38-6-of-facts-used-by-ai/. (Accessed: 5 December 2024).

Crawford, the predominantly homogeneous community of AI developers, which skews toward white males, also contributes significantly to the presence of bias.[167] For example, one research showed that in a Google image search for chief executive officer (CEO), 11 percent of the people depicted were women, compared with 27 percent of U.S. CEOs who are women.[168] Following many other sexist, racist, societal AI biases, the researchers have been producing tools for uncovering bias in AI models.[169] However, as Levendowski observes, just as code and culture play significant roles in how AI agents learn about and act in the world, so too do the laws that govern them; in this sense, copyright law exerts the most significant influence on AI bias.[170]

Restrictive regimes, such as the EU, which employs only exceptions to allow TDM, might contribute to biased research and technology the most. Essentially, this is one of the strongest criticisms towards CDSMD, and as pointed out by Margoni and Kretschmer, by establishing regulatory frameworks and defining ownership of key technological components, we influence the trajectory and societal implications of emerging technologies for the foreseeable future.[171] At the same time, allowing too much reshaping of the established safeguards of intellectual property rights against TDM/AI model training does not seem to fix the dilemma. Tilting the scales on either side through overregulation and constraints has historically proved to be the wrong way to go legislature-wise. It seems that at the crux of altercations between GenAI model developers and authors of copyrighted works is a need for mutual considerations.[172] GenAI model developers are primarily focused on producing well-functioning models, the basic value of which increases significantly when their outputs are high-quality, accurate, and frictionless. To note, TDM performed for scientific purposes has the same goals. Developers thus aim to harvest inputs from both public domain sources and copyright-protected works and to do so quickly, that is, without unnecessary constraints. For example, they would seek to avoid the limitations posed by the CDSMD, such as confirming if rightsholders have reserved the right to make reproductions for TDM/AI model training and examining the ToS for such opt-outs, which

[167] Crawford, K. (2016). Artificial Intelligence's White Guy Problem, N.Y. TIMES. Available at: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html. (Accessed: 5 December 2024).

[168] Langston, J. (2015). Who's a CEO? Google image results can shift gender biases. UW News. Available at: https://www.washington.edu/news/2015/04/09/whos-a-ceo-google-image-results-can-shift-gender-biases/. (Accessed: 7 December 2024).

[169] E.g. *see* Text Embedding Models Contain Bias. Here's Why That Matters. (2018). Available at: https://developers.googleblog.com/en/text-embedding-models-contain-bias-heres-why-that-matters/. (Accessed: 7 December 2024).

[170] Levendowski (n 23) 18.

[171] Margoni and Kretschmer (n 57)

[172] As in the doctrine of contacts in common law; a need for mutual bargain.

by the way may not even be machine-readable; or dealing with possible technological overridability imposed by TMPs. Most importantly, it is certainly not beneficial for AI developers to avoid said restraints by entering into multiple licensing deals to gain access to desired content. However, consider the positioning of rightholders, who might enjoy the extra protection, but their claiming of control over the 'parasitic' exploitation of the works is also financially motivated. The issue of author remuneration rights in this context is true for both the EU and the US regime. To exemplify, Recital 17 CDSMD provides that "Member States should, therefore, *not provide* for compensation for rightholders as regards uses under the text and data mining exceptions introduced by this Directive." Therefore, as some have fairly noted after CDSMD's implementation, "the policymakers will have to contend with angry rightsholders that see their works used in Machine Learning without equitable remuneration."[173] From the US perspective, it is still debated whether GenAI training contradicts the fourth factor in the fair use doctrine, that is if AI-generated outputs harm the commercial market for copyrighted works.

To rectify the above, a feasible consensus could be the proposed by Geiger and Iaia statutory license for TDM/machine learning purposes[174] or Senftleben's AI system "levy."[175] These proposals aim to foster an appealing environment for AI while preserving the essential role of human authors. Although they could face difficulties being fully implemented, at least some considerations will help to achieve a fairer balance. Given the essential role GenAI training plays nowadays for human beings, the proposal for a machine learning statutory license stems from such basic rights as freedom of expression and information;[176] freedom of the arts and sciences;[177] the right freely to participate in the cultural life of the community, etc.[178] Therefore, the described instruments aim to establish a revenue-sharing framework wherein AI developers split their earnings with the authors for utilizing their intellectual works in algorithmic training. Similar is the concept of an AI system levy. It would transform AI content revenue into human content revenue by granting a neighbouring right – in the form of a remuneration claim – in favour of human authors.[179]

[173] Guadamuz (n 30) 18.

[174] Geiger, C. and Iaia, V. (2023). The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI. *Computer Law & Security Review*, vol 52.

[175] Senftleben, M. (2022). A Tax on Machines for the Purpose of Giving a Bounty to the Dethroned Human Author – Towards an AI Levy for the Substitution of Human Literary and Artistic Works.

[176] Art. 11 EU Charter of Fundamental Rights.

[177] Art. 13 EU Charter of Fundamental Rights.

[178] Art. 27(1) Universal Declaration of Human Rights.

[179] Senftleben (n 175) 3.

Sub-Conclusion

In summary, this chapter has established that achieving a fair equilibrium between promoting (AI) innovation and preserving intellectual property rights is of paramount importance. Specifically, the prevalence of bias in AI models, as evidenced by numerous studies, underscores the critical need to address this issue. Besides the technical causes, copyright law emerges as a dominant factor in shaping said bias. Therefore, some proposals include an author-focused remuneration right approach to optimize the current US and EU copyright law regimes towards TDM/AI model training. These include possible statutory licenses for machine learning purposes and AI system levies, which aim to establish revenue-sharing mechanisms between AI developers and authors. Ultimately, such proposals can help foster an environment that promotes innovation while simultaneously preserving the essential role of human authors and should be taken into consideration by national legislators.

CONCLUSIONS AND PROPOSALS

1. Over the last few decades, TDM practices have been actively subjected to legal scrutiny and placed in regulatory frameworks within the copyright law domain. As an automated computational analysis, TDM is used to reveal patterns, trends, correlations, and discover new information or technology to benefit journalism, science, healthcare, education, environmentalism, etc. But since TDM is performed on a vast corpus of inter alia copyright-protected content, the exclusive right of reproduction of works or other subject matter can be infringed in this process.
2. More recently, rapid technological developments, such as the GenAI surge, have catalyzed the legislative shift toward the adoption of exceptions to respective rights for TDM acts. This is rationalized by the need to achieve a fair balance between the rights and interests of authors and other rightholders, on the one hand, and of users on the other.
3. Under the current EU legal framework, TDM-related reproductions of works are considered a copyright infringement. However, TDM can be exempt: Arts. 3 and 4 CDSMD provide mandatory TDM exceptions to the right of reproduction and the sui generis database right. The distinction is that Art. 3 permits TDM of works to which there is lawful access done by research and cultural institutions for non-profit scientific research, while Art. 4 allows TDM of lawfully accessed works for other purposes by any beneficiary, unless rightholders expressly reserve their rights (opt-out). Moreover, both exceptions can be technically overridden by rightholders using TPMs to restrict content access. Therefore, the introduced TDM exceptions cannot be said to be fully mandatory.
4. The AI Act reinforces the CDSMD's regulation of TDM, which is extensively used in GPAI model training on protected content. It imposes obligations on providers of GPAI models in Art. 53(1)(c), (d) to implement policies to comply with copyright law and related rights, particularly respecting the opt-outs under Art. 4(3) CDSMD and provide detailed summaries of the content used to train the GPAI models. These provisions can help avoid copyright infringement cases during AI training.
5. Within current US law and jurisprudence, TDM and other scrapings of copyright-protected content performed by copy-reliant technologies have been recognized as falling under the fair use doctrine, outlined in Section 107 of the 1976 Copyright Act. TDM falls outside the realm of infringement because it constitutes a non-expressive use, or can perform a transformative use of input material. Based on

established jurisprudence, it is probable that US courts will soon extend fair use recognition to AI training involving the use of copyrighted materials.

6. However, fair use arguments may no longer apply when AI-generated outputs mirror the copyrighted input material. This occurs when AI models memorize training data instead of merely extracting knowledge from it, resulting in the expressive use of original copyrighted works. Thus, copyright infringement is plausible in such cases.
7. From a comparative point, which legal regime is more appealing depends on the affected party. The EU copyright acquis adopts a rights-oriented approach to regulating TDM and AI model training, whereas the US framework is industry-centric. Both regimes exemplify their advantages and shortcomings.
8. The EU's CDSMD firmly safeguards original creators' rights and paired with the AI Act’s restrictions on GPAI model providers, it deters copyright infringement cases. The shortcoming of the same is CDSMD's failure to meet the extensive data needs for TDM/AI model training. Because of the rightsholders’ leverage to contractually/technologically override TDM exceptions, users of works are coerced to enter into licensing deals. Moreover, imposed restrictions generally slow down the European market for research and innovation. A legislative suggestion could be to change CDSMD’s provisions to prohibit the contractual/technological overridability of TDM exceptions when the user has lawful access to a work. In this way, it will be more feasible for users to benefit from the aforementioned exceptions.
9. The US fair use doctrine permits copy-reliant activities, such as TDM, without constituting copyright infringement, thereby promoting technological innovation. Simultaneously, rightholders do not enjoy the required safeguards against the use of their works in alike processes. Since the fair use doctrine loses its grounds in cases of AI memorization, the US legal regime could benefit from implementing a domestic AI Act similar to the EU's, which would enhance transparency in AI model training and assist in tracking scrapped copyrighted works.
10. Legislative changes are needed in both systems to balance the innovation and interests of rightholders. A viable solution could involve introducing a statutory license for machine learning or an AI system levy, which would create a revenue-sharing framework where AI developers (or TDM researchers) compensate authors for reproducing their intellectual property. This approach would secure authors’ remuneration rights in exchange for non-infringing use of their copyrighted works.

LIST OF REFERENCES

Regulatory Legal Acts

1. Charter of Fundamental Rights of the European Union [2012] (Fundamental Rights Charter) OJ C 326, 26.10.2012, p. 391–407.

2. Berne Convention for the Protection of Literary and Artistic Works 1886 (Berne Convention).

3. UN General Assembly, Universal Declaration of Human Rights, 217 A (III), 10 December 1948.

4. 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) (AI Act).

5. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (Database Directive) [1996] OJ L 77, 27.3.1996, p. 20–28.

6. Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (InfoSoc Directive) [2001] OJ L 167, 22.6.2001, p. 10–19.

7. Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Software Directive) [2009] OJ L 111, 5.5.2009, p. 16–22.

8. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (CDSMD) [2019] OJ L 130, 17.5.2019, p. 92–125.

9. Copyright Act of 1976, 17 U.S.C. § 107.

Special Literature

10. Aggarwal, C. C. (2015). *Data Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14142-8.

11. Carlini, N. et al. (2023). Extracting Training Data from Diffusion Models. Available at: http://arxiv.org/abs/2301.13188.

12. Carroll, M. W. (2019). Copyright and the Progress of Science: Why Text and Data Mining Is Lawful. 53 *UC Davis Law Review* 893, American University, WCL

Research Paper No. 2020-15, Available at SSRN: https://ssrn.com/abstract=3531231.

13. Caspers M. et al. (2016). D3.3 Baseline Report of Policies and Barriers of TDM in Europe (*FutureTDM).*

14. Cerquitelli, T. et al. (2017). Transparent Data Mining for Big and Small Data. http://www.springer.com/series/11970.

15. Cooper, A. et al. (2024). The Files are in the Computer: On Copyright, Memorization, and Generative AI. Cornell Legal Studies Research Paper Forthcoming, Chicago-Kent Law Review, Forthcoming, Available at SSRN: https://ssrn.com/abstract=4803118.

16. Dickson, E. et al. (2018). Data Mining Research with In-copyright and Use-limited Text Datasets: Preliminary Findings from a Systematic Literature Review and Stakeholder Interviews. *International Journal of Digital Curation*, *13*(1), 183–194. https://doi.org/10.2218/ijdc.v13i1.620.

17. Ducato, R., Strowel, A. (2019). Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility." IIC International Review of Intellectual Property and Competition Law, 50(6), 649–684. https://doi.org/10.1007/s40319-019-00833-w.

18. Fernandes, P. M. (2024). AI Training and Copyright: Should Intellectual Property Law Allow Machines to Learn. *Bioethica*, *10*(2), 8–21. https://doi.org/10.12681/bioeth.39041.

19. Fernández-Molina, J. C., de la Rosa, F. E. (2024). Copyright and Text and Data Mining: Is the Current Legislation Sufficient and Adequate? Portal, 24(3). https://doi.org/10.1353/pla.2024.a931775.

20. Geiger et al. (2019). Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU.

21. Geiger, C. and Iaia, V. (2023). The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI. *Computer Law & Security Review*, vol 52.

22. Geiger, C. et al. (2014) The Three Step Test Revisited: How to Use the Test's Flexibility in National Copyright Law. *American University International Law Review 581.*

23. Geiger, C. et al. (2018). Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?: Legal Analysis and Policy Recommendations. *IIC International Review of Intellectual Property and Competition Law*, *49*(7), 814–844. https://doi.org/10.1007/s40319-018-0722-2.

24. Geiger, C., Iaia, V. (2023). The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI. *Computer Law & Security Review*, vol 52.

25. Geiger, C., Jütte, B. J. (2024). Copyright as an Access Right: Concretizing Positive Obligations for Rightholders to Ensure the Exercise of User Rights. *GRUR International*, *73*(11), 1019–1035. https://doi.org/10.1093/grurint/ikae130.

26. Guadamuz, A. (2024). A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs. *GRUR International*, *73*(2).

27. Guibault et al. (2007). Study on the Implementation and Effect in Member States' Laws of Directive 2001/29/EC on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society.

28. Henderson, P. et al. (2023) Foundation Models and Fair Use. *Stanford Law and Economics*. Paper No. 584.

29. Krotov, V., Tennyson, M., (2018). Research Note: Scraping Financial Data from the Web Using the R Language. *Journal of Emerging Technologies in Accounting*, Vol. 15, No. 1, 169-181.

30. Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety, Technical report, META Group.

31. Levendowski, A. (2018). How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem. 93:579 *WASH. L. REV*.

32. Margoni, T., Kretschmer, M. (2022). A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology. *GRUR International*, 71(8), 685–701. https://doi.org/10.1093/grurint/ikac054.

33. Meeûs d'Argenteuil, J. et al. (2014). European Commission: Directorate-General for the Internal Market and Services. *Study on the legal framework of text and data mining (TDM)*, Publications Office. https://data.europa.eu/doi/10.2780/1475.

34. Murray, M. D. (2023). Generative AI Art: Copyright Infringement and Fair Use. *SMU Science and Technology Law Review*, 26(2).

35. Peukert, A. (2024). Copyright in the Artificial Intelligence Act – A primer. 73(6) GRUR.

36. Quintais, J. P. (2024). Generative AI, Copyright and the AI Act (v.2).

37. Quintais, J.P. (2019). The New Copyright in the Digital Single Market Directive: A Critical Look. European Intellectual Property Review 2020(1). Available at: https://ssrn.com/abstract=3424770.

38. Rosati, E. (2018). The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market.

39. Rosati, E. (2023). No step-free copyright exceptions: The role of the three-step in defining permitted uses of protected content (including TDM for AI-training purposes). *Stockholm Faculty of Law*, 1-22.

40. Rosati, E. (2024). Infringing AI: Liability for AI-generated outputs under international, EU, and UK copyright law. *European Journal of Risk Regulation.*

41. Sag, M. (2019). The New Legal Landscape for Text Mining and Machine. *Learning. Journal of the Copyright Society of the USA* 66, 2.

42. Sag, M. (2023). Copyright Safety For Generative AI. In HOUS. L. REV (Vol. 295).

43. Senftleben, M. (2022). A Tax on Machines for the Purpose of Giving a Bounty to the Dethroned Human Author – Towards an AI Levy for the Substitution of Human Literary and Artistic Works. SSRN Electronic Journal. 10.2139/ssrn.4123309.

44. Sousa e Silva, N. (2024). The Artificial Intelligence Act: Critical Overview.

45. V. Papakonstantinou/Paul De Hert (2022). The Regulation of Digital Technologies in the EU: The law-making phenomena of "act-ification", "GDPR mimesis" and "EU law brutality". *Technology and Regulation*, 48-60.

46. Wu, H. et al. (2024). Copyright protection during the training stage of generative AI: Industry-oriented U.S. law, rights-oriented EU law, and fair remuneration rights for generative AI training under the UN’s international governance regime for AI. *Computer Law and Security Review*, 55.

47. Wymeersch, P. (2023). EU Copyright Exceptions and Limitations and the Three-Step Test: One Step Forward, Two Steps Back. *GRUR International*, *72*(7).

Court Jurisprudence

48. *Infopaq International A/S v. Danske Dagblades Forening*, [CJEU], Nr. C-5/08, [16.07.2009]. EU:C:2009:465.

49. *SAS Institute Inc. v. World Programming Ltd.,* [CJEU], Nr. C 406/10, [02.05.2012]. EU:C:2012:259.

50. *Freistaat Bayern v. Verlag Esterbauer GmbH*, [CJEU], Nr. C-490/14, [29.10.2015], EU:C:2015:735.

51. *Stichting Brein v. Jack Frederik Wullems*, [CJEU], Nr. C-527/15, [26.04.2017]. EU:C:2017:300.

52. *Brompton Bicycle,* [CJEU], Nr. C-833/18, [11.06.2020]. EU:C:2020:461.

53. *CV-Online Latvia v. Melons*, [CJEU], Nr. C-762/19, [03.06.2021]. EU:C:2021:434.

54. *Mircom*, [CJEU], Nr. C-597/19, [17.06.2021]. EU:C:2021:492.

55. *Kneschke v. LAION,* [Hamburg Regional Court], Nr. 310 O 227/23, [27.09.2024].

56. *Harper & Row, Publishers, Inc. v. Nation Enters*., 471 U.S. 539, 556 (1985).

57. *Campbell v. Acuff-Rose Music, Inc.,* 510 U.S. 569 (1994).

58. *Eldred v. Ashcroft*, 537 U.S. 186, 220 (2003).

59. *Kelly v. Arriba Soft Corp*., 336 F.3d 811 (9th Cir. 2003).

60. *Perfect 10, Inc. V. Google Inc*. 508 F.3d 1146 (9th Cir. 2007).

61. *Authors Guild, Inc. v. HathiTrust,* No. 11 CV-4351 (HB), 2012 WL 4808939 (S.D.N.Y. Oct. 10, 2012).

62. *Google LLC v. Oracle Am., Inc*., 141 S. Ct. 1183, 1196 (2021).

63. *Andersen v. Stability AI Ltd.,* No. 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023).

64. *Getty Images (US), Inc. v. Stability AI Inc*., No. 1:23-cv00135-UNA (D. Del. Feb. 3, 2023).

65. *Andy Warhol Found. for the Visual Arts, Inc.*, 598 U.S. 508 (2023).

Other Sources

66. AI Maverick (2023). A Comprehensive Guide to Latent Space. Available at: https://samanemami.medium.com/a-comprehensive-guide-to-latent-space-9ae7f72bdb2f. (Accessed: 18 November 2024).

67. Amanatullah. (2023). Transformer Architecture explained. *Medium*. Available at: https://medium.com/@amanatulla1606/transformer-architecture-explained-2c49e2257b4c. (Accessed: 18 November 2024).

68. Andreson, M. (2023). Weights in Machine Learning. *Metaphysic*. Available at: https://blog.metaphysic.ai/weights-in-machine-learning/. (Accessed: 18 November 2024).

69. Beckmann, C. et al. (2024). First court decision on text and data mining copyright exceptions: Kneschke v LAION e.V. *Bristows*. Available at: https://inquisitiveminds.bristows.com/post/102jmd3/first-court-decision-on-text-and-data-mining-copyright-exceptions-kneschke-v-lai. (Accessed: 16 November 2024).

70. Campbell, F. (2019). Data Scraping – Considering the Privacy Issues. *Fieldfisher*. Available at: https://www.fieldfisher.com/en/services/privacy-security-and-information/privacy-security-and-information-law-blog/data-scraping-considering-the-privacy-issues (Accessed: 2 November 2024).

71. Crawford, K. (2016). Artificial Intelligence's White Guy Problem, N.Y. TIMES. Available at: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html. (Accessed: 5 December 2024).

72. Duarte, F. (2024). Amount of data created daily. Available at: https://explodingtopics.com/blog/data-generated-per-day (Accessed: 31 October 2024).

73. Gruet, M. (2024). That's Just Common Sense. USC researchers find bias in up to 38.6% of "facts" used by AI - USC Viterbi | School of Engineering. Available at: https://viterbischool.usc.edu/news/2022/05/thats-just-common-sense-usc-researchers-find-bias-in-up-to-38-6-of-facts-used-by-ai/. (Accessed: 5 December 2024).

74. Guadamuz, A. (2024). Snoopy, Mario, Pikachu, and reproduction in generative AI. *TechnoLlama*. Available at: https://www.technollama.co.uk/snoopy-mario-pikachu-and-reproduction-in-generative-ai. (Accessed: 1 December 2024).

75. Hugenholtz, P. B. (2019). The New Copyright Directive: Text and Data Mining (Articles 3 and 4). *Kluwer Copyright Blog*. Available at: https://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ (Accessed: 15 November 2024).

76. Joint letter to Members of the European Parliament on the impact of Artificial Intelligence on the European creative community. (23 July 2024). Available at: https://europeanwriterscouncil.eu/wp-content/uploads/2024/07/Joint-letter-to-Members-of-the-European-Parliament-on-the-impact-of-Artificial-Intelligence-on-the-European-creative-community.pdf.

77. Keller, P. (2024). Machine readable or not? – notes on the hearing in LAION e.v. vs Kneschke. *Kluwer Copyright Blog*. Available at: https://copyrightblog.kluweriplaw.com/2024/07/22/machine-readable-or-not-notes-on-the-hearing-in-laion-e-v-vs-kneschke/. (Accessed: 16 November 2024).

78. Langston, J. (2015). Who's a CEO? Google image results can shift gender biases. UW News. Available at: https://www.washington.edu/news/2015/04/09/whos-a-ceo-google-image-results-can-shift-gender-biases/. (Accessed: 7 December 2024).

79. Lee, T., Grimmelmann, J. (2024). Why The New York Times might win its copyright lawsuit against OpenAI. Available at: https://www.understandingai.org/p/the-ai-community-needs-to-take-copyright. (Accessed: 1 December 2024).

80. Niiler, E. (2020). An AI epidemiologist sent the first warnings of the coronavirus. *WIRED*. Available at: https://www.wired.com/story/ai-epidemiologist-wuhan-public-health-warnings (Accessed: 31 October 2024).

81. Quintais, J. P. (2024). Copyright, the AI Act and extraterritoriality. *Kluwer Copyright Blog*. Available at: https://copyrightblog.kluweriplaw.com/2024/11/28/copyright-the-ai-act-and-extraterritoriality/. (Accessed: 19 November 2024).

82. Rose, J. (2022) Inside Midjourney, The Generative Art AI That Rivals DALL-E. Vice. Available at: https://www.vice.com/en/article/inside-midjourney-the-generative-art-ai-that-rivals-dall-e/. (Accessed: 18 November 2024).

83. Rosenberg, L. (2022), Generative AI: The technology of the year for 2022, *Big Think*. Available at: https://bigthink.com/the-present/generative-ai-technology-of-year-2022/. (Accessed: 5 November 2024).

84. Sousa e Silva, N. (2024). Are AI models' weights protected databases? Kluwer Copyright Blog. Available at: https://copyrightblog.kluweriplaw.com/2024/01/18/are-ai-models-weights-protected-databases/. (Accessed: 18 November 2024).

85. Spawning opts out 78 million artworks from AI training. (2023). Available at: https://spawning.substack.com/p/spawning-opts-out-78-million-artworks (Accessed: 7 November 2024).

86. Text Embedding Models Contain Bias. Here's Why That Matters. (2018). Available at: https://developers.googleblog.com/en/text-embedding-models-contain-bias-heres-why-that-matters/. (Accessed: 7 December 2024).

87. Treehouse Tech Group (2021). Big Data vs. Traditional Data: What's the Difference? Available at: https://treehousetechgroup.com/big-data-vs-traditional-data-whats-the-difference/#:~:text=Ultimately%2C%20big%20data%20refers%20to (Accessed: 31 October 2024).

88. VivekR (2023). How did Netflix use big data to transform their company and dominate the streaming industry? *Medium*. Available at: https://vivekjadhavr.medium.com/how-did-netflix-use-big-data-to-transform-their-company-and-dominate-the-streaming-industry-a93f90ae8dad (Accessed: 31 October 2024).

89. Warso, Z. et al. (2024) Blueprint of the Template for the Summary of Content Used to Train General-Purpose AI Models (Article 53(1)d AIA) – v.2.0. *Open Future Foundation*. Available at: https://openfuture.eu/blog/sufficiently-detailed-summary-v2-0-of-the-blueprint-for-gpai-training-data/. (Accessed: 20 November 2024).

SUMMARY

Text And Data Mining In Copyright Law

Mariia Afanasieva

This master's thesis provides an analysis of the EU and US legal frameworks governing text and data mining (TDM) activities under copyright law. Because TDM is essential for Generative Artificial Intelligence (AI) model training, this subject is also explored through the copyright law prism. The research critically evaluates the efficiency of both regimes, focusing on the EU's TDM mandatory exceptions to the right of reproduction, the sui generis database right and the US fair use doctrine.

The EU legal framework considers TDM-related reproductions of protected works as copyright infringement. The scope of TDM admissibility is limited to a narrow exception for specific purposes and broader exception for other purposes. However, their mandatory nature is undermined by established possibilities for rightholders to contractually reserve their rights against TDM (or AI model training) and impose technical protection measures that restrict access to their works. US copyright doctrine and case law recognize TDM and similar copy-reliant activities as fair use. As such, those are considered to be non-expressive or transformative uses of original works. At the same time, AI-generated outputs that closely resemble copyrighted input material due to memorization rather than just knowledge extraction may constitute copyright infringement, as they represent expressive use of original works.

Through a comparative approach, this work recognizes that said legal regimes have their advantages and drawbacks, but at the crux for both is the legislative need to achieve a fair balance between innovation and safeguarding the interests of rightholders. A possible solution could be introducing a statutory license for machine learning or an AI system levy. Those would establish a revenue-sharing mechanism in which AI developers or TDM researchers provide compensation to authors for reproducing their intellectual property.

Lastly, as technological innovation accelerates, the legal challenges explored in this thesis are likely to intensify and grow more intricate. Consequently, legislators must respond swiftly with more open and adaptive regulatory frameworks to effectively maintain a balanced copyright law regime.