University of Malta

Master's Thesis

The Virtual Void - Copyright Challenges Faced in the Generation of AI Works

Author:
Stefani Marinova

Supervisor:
Dr. Ioannis Revolidis

A thesis submitted in partial fulfilment of the degree of
Master of Laws in European Business Law

Faculty of Laws, University of Malta

16 September 2024

Abstract

We live in an era where machines, having been fed with millions of pieces of human expression, are leveraging the extensive data they have collected to generate new, unconventional forms of art. While this has certainly been an entertaining endeavour for tech enthusiasts and a lucrative opportunity for AI developers, it has thus brought to the forefront important questions as to the implications for intellectual property rights in this creative process.

Over the years, text and data mining (hereinafter: TDM) has solidified its position as a building block for Artificial intelligence (hereinafter: AI) companies, particularly those specialising in creating machine-generated artworks. Nevertheless, 'the copyright black hole' associated with the generation of AI art, has sparked heated controversies over the use of copyrighted material for algorithmic training, often acquired, without proper authorisation. To address these concerns the European Union (hereinafter: EU) introduced two mandatory exceptions for TDM activities, as outlined in Articles 3 and 4 in the Directive on Copyright in the Digital Single Market (hereinafter: DSM Directive).

The thesis investigates how Article 4 DSM Directive navigates the legal complexities involved in the use of authorship work in TDM activities, specifically in instances where such materials are used without the explicit consent of rightholders. It begins by examining the essence of Article 4 DSM Directive and the underlying concept of TDM. The analysis then continues by critically assessing whether the copyright issues related to TDM, were adequately resolved within the said Article. Finally, the thesis concludes that, despite the inclusion of Article 4 DSM Directive, significant shortcomings persist. It concludes by proposing further initiatives to better align the EU legal framework with the evolving demands of AI technology.

Keywords: AI, copyright, input, TDM, DSM Directive

Table of contents

Table of Statues, Judgments and Treaties

Charter of Fundamental Rights of the European Union [2012] OJ C 326/391 (CFR)

Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market [2019] OJ L 130/92 (DSM Directive)

Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs [2009] OJ L111/16 (Software Directive)

Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market, COM(2016)593, 14 September 2016

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) [2024] OJ L2024/1689

Council of Europe legislation

Convention for the Protection of Human Rights and Fundamental Freedoms as Amended by Protocols No.11 and No.14 (adopted on 4 November 1950, entered into force 3 September 1953) CETS No. 005 (ECHR)

International Treaties

Agreement on Trade-Related Aspects of Intellectual Property Rights (adopted 15 April 1994, entered into force 1 January 1995) 1869 UNTS 299 (TRIPS)

Berne Convention for the Protection of Literary and Artistic Works (adopted 9 September 1886, entered into force 5 December 1887) 1161 UNTS 3 (Berne Convention)

WIPO Copyright Treaty (adopted 20 December 1996, entered into force 6 March 2002) 2186 UNTS 121 (WCT)

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AGI	Artificial General Intelligence
API	Application Programming Interface
IR	Information Retrieval
KDD	Knowledge Discovery in Databases
CNN	Convolutional Neural Networks
CLIP	Contrastive Language-Image Pretraining
LDM	Latent Diffusion Models
DM	Diffusion Models
IoT	Internet of Things
TDM	Text and Data Mining
ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
CJEU	Court of Justice of the European Union
EU	European Union
DSM Directive	Directive on Copyright in the Digital Single Market
ECHR	European Convention on Human Rights
WCT	WIPO Copyright Treaty
TRIPS	Agreement on Trade-Related Aspects of Intellectual Property Rights
Software Directive	Directive on the legal protection of computer programs
InfoSoc Directive	Directive on the harmonisation of certain aspects of copyright and related rights in the information society
AI Act	Artificial Intelligence Act
UK	United Kingdom
U.S.	United States

Acknowledgements:

I dedicate this thesis to my beloved mom and dad, who have never stopped believing in me. For their love, warmth and unwavering support.

I want to express my gratitude to all other important people in my life, who have walked this long path alongside me.

Special thanks to Dr. Ioannis Revolidis whose mentorship, expertise and constant guidance have been instrumental in the completion of this thesis.

CHAPTER 1

INTRODUCTION

The development of Artificial intelligence (hereinafter: AI) has achieved exceptional success in all walks of life and continues to prove as a vital tool, across many facets of everyday functioning. With its vast range of capabilities, AI has truly woven itself into the fabric of various industries, including healthcare, finance, transportation, and education. This integration has revolutionized how processes are optimized and services delivered, heralding a sea change in what computers are capable of doing. As these technologies evolve, they continue to refine how humans interact with machines, even in the context of art. This fosters new avenues of creativity by unlocking greater possibilities for exploration within the field of art.

AI is probably the defining technology of the last decade and may well continue to be so in the future.[1] It encompasses the fields of science and engineering and is focused on creating intelligent systems that closely replicate human-like intelligence and capabilities.[2] Best understood as a subset of computer science, AI intersects with disciplines such as data science, robotics, and cognitive computing. Essentially, AI serves as a catchall term for 'intelligent machines' that can solve problems, make or suggest decisions, and perform tasks that usually require human intervention.[3] It could be said that this is rooted in its capability to mimic cognitive functions such as perception, reasoning, and problem-solving, which allows machines to tackle a diverse array of tasks, ranging from simple automation to complex decision-making processes.

The concept of AI was born at a conference held at Dartmouth College in the summer of 1956, where John McCarthy coined the term 'artificial intelligence,' defining it as '*the science and engineering of making intelligent m*achines'.[4] In the late 1980s and early 1990s, the advent of machine learning (hereinafter: ML) enabled computers to analyse and learn from extensive amounts of data, improving their performance across various tasks. Deep learning

[1] Philip Boucher, 'Artificial intelligence:How does it work, why does it matter, and what can we do about it?' PE 641.547 - June 2020, I.
[2] David M. Townsend and Richard A.Hunt, 'Entrepreneurial action, creativity, & judgment in the age of artificial intelligence' (2019) 11 Journal of Business Venturing Insights 1, 2.
[3] Steve Blank, 'Artificial Intelligence and Machine Learning Explained' (*Gordian Knot Center for National Security Innovation*, 17 May 2022) 2 <https://steveblank.com/2022/05/17/artificial-intelligence-and-machine-learning-explained/> accessed 17 June 2024.
[4] Yougjun Xu, 'Artificial intelligence: A powerful paradigm for scientific research' (2021) 2(4) Cell Press 1, 2.

(hereinafter: DL), which is a subset of ML that employs multi-layered neural networks to learn from large datasets and make complex decisions, became a key area of AI research following a significant breakthrough in 2006 when Geoffrey Hinton and colleagues developed deeper neural networks and introduced techniques to address the gradient vanishing problem during training.[5] Since then, numerous neural network architectures and training methods have been developed to enhance the capabilities of DL, broadening its applications across diverse areas.[6]Nowadays, DL algorithms often surpass human performance in specific domains, particularly in computer vision and natural language processing (hereinafter: NLP).[7]

Although AI has been around for quite some time, there is still no universally accepted definition. Since its introduction, AI has generally been defined by its ability to exhibit intelligent behaviour. It is an umbrella term that covers a wide majority of technologies and applications, including various computational strategies, such as machine learning, cognitive computing and robotics.[8] Nevertheless, the European Commission Communication defined AI as '*systems that display intelligent behaviour by analysing their environment and taking action -with some degree of autonomy - to achieve specific goals*'[9]. *Strictu sensu*, AI refers to the imitation of human intelligence by computers, enabling machines to emulate the intricate skills that are inherently human.[10] It goes by no surprise that defining AI presents many obstacles. Although human intelligence has been long subject of research, there is no consensus on what it truly encompasses. This lack of clarity also makes it strenuous to define AI systems, especially since no one can predict with certainty what machines will be capable of in the future or how the field might evolve. Consequently, people tend to view human intelligence as an evolving concept, one that changes as technology advances and becomes more refined.

[5] Geoffrey E Hinton, Simon Osindero and Yee-Whye Teh, 'A fast learning algorithm deep belief nets' (2006) 18(7) Neural Computation 1527, 1527-1554.
[6] Maryssa Ahmad Ali Elfa and Mina Eshaq Tawfilis Dawood, 'Using Artificial Intelligence for enhacing Human Creativity' (2023) 2(2) Journal of Art, Design & Music 106, 108.
[7] Xu (n 4).
[8] Leonard Knoedler, 'Application possibilities of artificial intelligence in facial vascularized composite allotransplantation-a narrative review' (2023) 10 Frontiers 1, 2.
[9] European Commission, Communication on artificial intelligence for Europe, COM(2018)237, 25 April 2018, 1.
[10] Haroon Sheikh, Corien Prins and Erik Schrijvers, *Mission AI* (Springer 2023) 15.

Today, AI can be divided into two primary categories: narrow AI and general AI.[11] Narrow AI, which is what this thesis will be concentrating on and what we are familiar with today, consists of systems specifically designed to perform particular tasks, such as creating images.[12] While these systems are highly effective in their specialized areas, they do not possess the necessary capability to operate beyond their specific programming functions. In contrast, artificial general intelligence (hereinafter: AGI) is still largely a concept and seeks to perform any intellectual task that a human can, including those it has not been specifically trained for.[13]

1.1 General Appraisal

AI has been intertwined with conventional forms of art for serval decades, with its influence and application evolving outstandingly over time. The emergence of AI-driven creativity is predominantly propelled by the rising availability and accessibility of data.[14] These datasets, which form the backbone of AI training models, are often drawn from publicly available sources, which contain copyrighted materials that are frequently obtained without prior authorisation from the authors. Namely, the selection and scope of these data inputs play a crucial role in shaping how AI models learn and evolve. This practice, also known as text and data mining (hereinafter: TDM), raised substantial concerns regarding harvesting techniques and their interaction with owners' rights. Essentially, this unregulated use of protected works for AI training purposes introduced a myriad of complex legal and ethical challenges.

Copyright protection serves a dual purpose: it seeks to encourage creative expression, while simultaneously promoting public welfare through access to expressive works. In exchange for creating literary and artistic works, authors are granted exclusive rights, including the rights to reproduction, distribution, adaptation, communication to the public, and the preparation of derivative works, for a limited duration. However, concerns about copyright liability have the potential to stifle innovation, especially when AI systems interact with protected works during

[11] Tiing Ang and others, 'The rise of artificial intelligence: addressing the impact of large language models such as ChatGPT on scientific publications' (2023) 64(4) Singapore Medical Journal 219, 219.
[12] League of European Research Universities, 'Text and Data Mining: The need for change in Europe' (*Liber Europe*) <https://libereurope.eu/topic/text-data-mining/> accessed 1 June 2014.
[13] Anjan Chatterjee, 'Conceptual Analysis' (2022) 13 Frontiers 1, 2.
[14] European Commission, 'White Paper on Artificial Intelligence: a European Approach to Excellence and Trust', Brussels (19 February 2020) COM(2020) 65 final, 1.

the training process. Nonetheless, if AI models were restricted from analysing vast amounts of data through TDM, their ability to discover new patterns and insights would be inevitably hindered, potentially impeding also technological advancement.[15]

AI and copyright intersect when protected data is used to train machines to learn, reason and act as humans.[16] The input these systems rely on is not merely a technical aspect; rather, it is central to the ongoing debates regarding the permissible use of copyrighted materials during AI training processes and how far such usage may extend without violating the established legal rights of artists. Given the large sets of data often scrapped from the internet to train an AI system, there is a high likelihood that some of the inputted data may impinge upon the copyright granted to authors. Usually, AI systems tend to opt for copyrighted materials due to their high quality and valuable content which reflects the technical skills and trends, necessary for the generation of art.

The pressing legal issue revolving around whether AI systems can use copyrighted content, such as images, during the training phase without securing consent, has been articulated within a set of mandatory exceptions, introduced in the Directive on Copyright in the Digital Single Market (hereinafter: DSM Directive)[17]. The latter limitations permit the use of copyrighted materials without prior authorization, provided certain conditions are met. Although the European Union (hereinafter: the EU) tried to strike a balance with the introduction of the exceptions, the porous nature of copyright law, particularly as it relates to AI-generated art, continued igniting a range of intricate legal questions concerning the boundaries of intellectual property protection.

As evidence by the multiple lawsuits filed predominantly in the United States, the legal landscape in this area is still developing. One notable case, the outcome of which remains to be determined, is *Getty v Stability AI*[18], wherein Getty has asserted that the defendant

[15] Elena Rosati, 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspects PE, 604.942, 2.
[16] Jenny Quang, 'Does Training AI Violate Copyright Law' (2021) 36(4) Berkeley Technology Law Journal 1407, 1413.
[17] Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market [2019] OJ L 130/92 (DSM Directive).
[18] *Getty Images (US), Inc. v. Stability AI, Inc.*, 1:23-cv-00135, (D. Del.).

unlawfully copied over 12 million copyrighted images from its website without authorization to train the Stable Diffusion model. The outcome of the case will have profound implications for the future scope of intellectual property protection in the context of emerging technologies, particularly concerning the use of copyrighted images as input for training AI models. Other artists such as Greg Rutkowski, who is famous for employing classical painting styles for the design of his fantasy landscapes, chose not to pursue a legal battle in court.[19] Instead, he raised public awareness about the unauthorised use of his artwork without obtaining consent.[20] Respectively, this led to changes in Stable Diffusion 2.0, which removed the ability of the system to mimic specific artists' rights, although users later found workarounds to reintroduce them. [21] Other efforts to combat the unauthorized use of protected authorship content include the launch of the site 'Have I Been Trained,' which lets artists conduct an online query to see if their work is among the 5.8 billion images available in the dataset used to train models like Stable Diffusion.

Technology and copyright law have crashed into each other in a profound way, highlighting the urgent need to reconcile the rapid advancement of AI with the principles of intellectual property. As AI technologies constantly evolve and integrate into diverse creative aspects, it is imperative for legal frameworks to adapt and address the novel challenges posed by AI's use of copyrighted materials. Taking the right approach to tackle these alarming issues will not only shape the future of AI innovation but also determine the extent to which copyright protections can be harmonized with the demands of the technologically driven world.

1.2 Research Question

By incorporating Article 4 DSM Directive, the European Union intended to carve out a mandatory exception for TDM which both encourages innovation and provides more legal

[19] Melissa Heikkilä, 'This artist is dominating AI-generated art. And he's not happy about it. Greg Rutkowski is a more popular prompt than Picasso' (*MIT Technology Review,* 16 September 2022) <https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/> accessed 5 July 2024.

[20] Jose Antonio Lanz, 'Greg Rutkowski Was Removed From Stable Diffusion, But AI Artists Brought Him Back' (*Decrypt*, 30 July 2023) <https://decrypt.co/150575/greg-rutkowski-removed-from-stable-diffusion-but-brought-back-by-ai-artists> accessed 17 August 2024.

[21] Yubin Ma, 'The Ultimate Stable Diffusion LoRA Guide (Downloading, Usage, Training)' (*Aituts*, 24 April 2024) <https://aituts.com/stable-diffusion-lora/> accessed 1 August 2024.

certainty for harvesting activities.[22] This mandatory exception, allowed all kinds of beneficiaries to execute acts of '*reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining*'.[23] As of today, the effectiveness of this Article in reconciling the competing interests of AI developers and copyright holders remains a central and unresolved issue, even following the transposition of the DSM Directive by all Member States.

Therefore, in view of what was discussed in Subsection 1.1, this dissertation raises the following research question: *Does EU copyright law adequately accommodate TDM for the training of AI systems?* This question seeks to explore how Article 4 DSM Directive addresses the legal complexities surrounding the use of copyrighted works in TDM activities for the purpose of training AI algorithms, particularly in cases where protected materials are harvested without the explicit consent of rights holders. Specifically, this dissertation will focus on examining whether Article 4 DSM Directive successfully fills the lacuna in the law, as related to TDM activities, or whether further reforms are needed to align the EU legal framework with the evolving needs of AI technologies.

1.3 Methodology

To address the aforementioned research question, this thesis will employ a black-letter approach. The Black-letter approach aims to systemise, rectify, and clarify the law on a particular topic by applying a distinctive mode of analysis of authoritative texts that consist of primary and secondary sources.[24] This approach will be used in order to provide an interpretative tool and legal reasoning for evaluating the legal rules and suggesting recommendations for the development of the law.[25]

[22] Gina Maria Ziaja, 'The text and data mining opt-out in Article 4(3) CDSMD: Adequate veto right for rightholders or a suffocating blanket for European artificial intelligence innovations?' (2024) 19(5) Journal of Intellectual Property Law 453, 453.
[23] DSM Directive, art 4(1).
[24] Mike McConville and Wing Hing Chui (eds), *Research Methods for Law* (2nd Edition, Edinburgh University Press 2007) 4.
[25] Colin Bell, 'A Note on Participant Observation' (1969) 3(2) Sociology 417, 417.

The black-letter approach consists of a two-part process. First, the law needs to be analysed, and an objective reality established.[26] This requires locating and thoroughly scrutinising primary sources of the law, such as legislation and case law, in order to establish the nature and parameters of the law that form crux of the said methodology.[27] In particular, this thesis will build upon examining primary legal sources, such as International Treaties and Agreements, EU law, judgments rendered by the European Court of Justine (hereinafter: the CJEU), relevant case law from the US, with brief references to Human Rights Instruments and national laws, transposing EU legislation. The black-letter approach goes beyond merely describing the law, rather it requires a critical reading and an in-depth analysis that connects new findings to the existing body of law.[28] Further to the primary sources, secondary sources including, but not limited to, legal textbooks, academic journals, legal commentaries and opinions, white papers, recommendations, guidelines, other non-binding documents issued by the Commission, will be used to provide further context and interpretative insights. Although not legally binding, the latter sources provide guidance and clarify the practical enforcement of different legal areas.

The second step involves exploring how the established legal rules are applied in practical scenarios. The interpretative stage of the black-letter approach will apply an in-depth examination of the content and application of the legislation with the help of systematic and teleological methods of interpretation. The interpretation of the legislation seeks to examine the adequacy of the current TDM legal framework, as it relates to the training AI systems with copyrighted data. For the purpose of writing this thesis, the black-letter approach provides the most rigorous framework allowing not only to examine the law, as it is, but bring about new suggestions and perspectives on its interpretation, given the current state-of-the-art.

1.4 Research Limitations

This thesis will concentrate solely on the use of copyrighted works as inputs in the training of AI systems, without considering the legal implications tied to the generation of outputs by AI

[26] Terry Hutchinson, Researching and Writing in Law (3rd Edition, Thomson Reuters 2010) Researching 37.
[27] ibid.
[28] ibid 38.

models. While the production of outputs from AI systems presents its own unique set of copyright challenges, the author will refrain from exploring this area. Instead, it will focus on the potential infringements related to the use of copyrighted material by some of the most influential AI companies for TDM purposes, highlighting how these inputs may implicate the rights of original authors under the DSM Directive.

This dissertation will omit to provide a comprehensive analysis of U.S. law and the fair use doctrine, which is a distinctly U.S. legal concept. Although the fair use doctrine presents an affirmative defence to copyright infringements,[29] it significantly differs from the principles and limitations outlined in the DSM Directive. As such, any discussions or comparisons between the articles set out in the Directive and the said doctrine will be excluded from this analysis. References to the latter concept will be provided only for illustrative purposes. This serves to maintain a clear focus on the provisions and interpretations arising solely from the limitations of the DSM Directive. In addition to this, an extensive legal comparison with other similar exceptions and limitations to copyright, which exist outside of the scope of the DSM Directive, will be also excluded from the scope of this dissertation.

Another limitation present is the inherent lack of transparency present within the AI industry. This could be attributed to the fact that most providers refrain from disclosing information regarding their systems and practices. To mitigate this constraint, the thesis will focus on the most transparent providers so far in the industry. Therefore, the dissertation will provide an overview of the information that is currently available and accessible, and examine the development of the relevant case law, as they stand at this time.

It is also important to note that this thesis will deliberately avoid discussing thoroughly the national interpretations or transpositions of the DSM Directive within the laws of the Member States. Therefore, the legal analysis provided will be strictly confined to the text and objectives of the Directive itself. This approach ensures that the thesis remains focused on the rules and principles laid out in the Directive, rather than the potentially divergent interpretations adopted by Member States.

[29] Quang (n 16) 1435.

It is worth noting that this field of study is still evolving, and many of the relevant practices adopted by AI companies may change over time. While the available literature on this topic addresses only specific aspects of AI, the thesis sought to provide the most comprehensive and in-depth analysis as it is currently possible. With this in mind, the answers to the research question posed are based solely on the current state of technology and the law, considering that future developments may require different interpretations.

1.5 Outline

This thesis is divided into four chapters. Chapter One offers a general overview of the topic by outlining any foundational elements. It also provides information as to the research question, limitations and methodology that will be followed throughout.

Chapter Two proceeds with an explanation of what TDM is, providing a detailed description of how AI companies extract copyrighter material for model training. This will be followed by a deliberation on the process of text and data harvesting and the potential copyright violations associated with this activity. Further analysis will be provided as to how some of the leading AI companies mine art, prior to embedding the inputted data into their systems for training purposes. Although the legal landscape governing AI practices remains in flux, reference will be made to case law to illustrate the existing regulatory responses to these issues.

Chapter Three examines the law behind the operations involved in TDM. It begins by tracing the legal development that led to the adoption of the DSM Directive and, consequently continues by examining Articles 3 and 4 DSM Directive, which introduced exceptions for TDM in research and commercial settings. The discussion then moves on to assess specifically how Article 4 DSM Directive addresses copyright challenges, arising from the harvesting of text and data. This is followed by a general overview of the legal criteria for granting copyright protection. Lastly, the chapter will explore the conditions under which protected works can be legally used for TDM and the circumstances where such use is restricted, highlighting the legal limitations and boundaries imposed by the Directive.

Chapter Four serves as a concluding chapter, evaluating the efficacy of Article 4 DSM Directive in addressing the copyright issues pertaining to TDM, as discussed in the preceding chapters. Subsequently, an elaborate analysis will be provided as to the potential shortcomings inherent in the limitation. Finally, the chapter will propose potential solutions to rectify the identified problems and enhance the regulation of TDM activities in Europe.

CHAPTER 2

HOW AI PROVIDERS HARVEST AND MINE PROTECTED WORKS

AI has revolutionised not only the way works are generated but enabled the emergence of new groundbreaking forms of digital expression and interactive experiences. Although some may think that the concept of non-human-created art is a new tech advancement, its origins stretch back to ancient Greek mythology.[30] In the 1970s, efforts to create computer-generated art were already well-documented, with extensive discussions on the intersection of computers, AI, and art emerging well before the technology had fully developed.[31] Harold Cohen was the prominent artist and programmer behind the creation of AARON, known as the earliest example of AI software, codified to paint and draw. While AI creativity has long been considered in theory, major leaps in generative AI tools have only occurred recently. In the 21st century, AI experienced a significant pinnacle. This was fuelled by several factors of which the introduction of a more sophisticated class of algorithms, the availability of low-cost graphic processors capable of performing multiple calculations in milliseconds, as well as the availability of large, correctly annotated datasets, allowing for the more advanced learning in intelligent systems.[32] As a result, AI and ML applications now extend far beyond creating innovative art pieces. They also involve examining and categorizing different art styles, developing responsive installations, and harnessing cutting-edge data mining techniques for generating art.

Nevertheless, the bloom of AI art is mainly accelerated by the surged availability of data. AI algorithms, capable of creating digital art rely heavily both on the quality of data and the methods used to extract such value.[33] Due to the widespread access to massive, networked computing power and the exponentially enlarging digital datasets, the techniques employed for TDM are continuously advancing.[34] According to the 2016 European Commission's Impact

[30] Alex Shashkevich, 'Stanford researcher examines earliest concepts of artificial intelligence, robots in ancient myths' (*Stanford Report*, 28 February 2019) <https://news.stanford.edu/stories/2019/02/ancient-myths-reveal-early-fantasies-artificial-life> accessed 1 August 2024.
[31] Peter Kugel, 'Artificial Intelligence and Visual Art' (1981) 14(2) Leonardo 137, 137–139.
[32] Nikitad Patil, Shifa Patel and Sapna Lawsnd, 'Research on Artificial Intelligence and its Applications' (2023) 44(8) Journal of Advanced Zoology 229, 230.
[33] Matt Zwolenski and Lee Weatherill, 'The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things' (2020) 2(3) Australian Journal of Telecommunications and the Digital Economy 1, 1-9.
[34] Ian Hargreaves and others, 'Standardisation in the Area of Innovation and Technological Development, notably in the Field of Text and Data Mining' (*European Commission*, April 2014) 63 <https://www.researchgate.net/publication/265078277_Final_Report_from_the_Expert_Group_on_Standardisation_in_the_area_of_Innovation_and_technological_development_Notably_in_the_field_of_Text_and_Data_Mining> accessed 1 June 2024.

Assessment accompanying the Proposal for the DSM Directive, stakeholders widely agreed that TDM presents 'a nascent tool'.[35] Despite this, there was a clear consensus on its significant potential.[36] The decreased cost of data storage, the deployment of mobile devices and the Internet of Things (hereinafter: IoT) are some of the factors set to drive the ongoing exponential growth of data, underscoring their relevance in the overal development of AI algorithms creating art.[37]

2.1 What Is TDM?

Before discussing how systems extract copyrighted material, it is crucial to shed light on what TDM is. Strictly speaking, this process can be described as '*the selection and application of complex algorithms to the transformed alphanumerical dataset to gather hidden information*'.[38] This technique shall be viewed as an iterative process that connects unstructured data to a specific domain of interest. At the heart of it, TDM refers to the process of deriving large quantities of information from machine-readable materials. This involves analysing huge sets of unstructured text documents, extracting the relevant data, and recombining it to find patterns. As a matter of fact, TDM is not focused on replacing the original material; rather, it is tasked with extracting facts and data from extensive datasets. Essentially, through TDM, substantial amounts of text and data can be processed, extracted, and recombined which can lead to the discovery of new and useful information amongst enormous piles of irrelevant data -an action which goes far beyond the capacity of human minds.[39] The handiness of TDM expressed in the rich and diverse datasets it creates, is brought into effect through its application in the training of AI algorithms, ultimately aiming to advance AI-driven creativity.

[35] European Commission, Commission Staff Working Document - Impact Assessment on the modernisation of EU copyright rules, SWD(2016)301 final, 14 September 2016, 104.
[36] ibid.
[37] Sergey Filippov, 'Mapping Text and Data Mining in Academic and Research Communities in Europe' (2014) 16 The Lisbon Council 1, 3.
[38] Gonazalo Mariscal, Oscar Marbán and Covadonga Fernández, 'A Survey of Data Mining and Knowledge Discovery process Models and Methodologies' (2010) 25(2) KER 137, 137-166.
[39] Marti A. Hearst, 'Untangling Text and Data Mining' (1999) Proceedings of the 37th Annual meeting of the Association for Computational Linguistics 3, 3-10.

In essence, TDM shall be viewed not as a standalone field per se but rather as an interdisciplinary field that seamlessly integrates other areas, including information retrieval (hereinafter: IR), knowledge discovery in databases (hereinafter: KDD), information extraction, AI and NLP. It is important to draw a distinction between TDM and data mining, primarily because TDM analyses unstructured text data to extract useful information; whereas, data mining is primarily focused on discovering knowledge from structured databases.[40] For this reason, data mining was recognised as a subset within the larger process KDD.[41] In addition to the above, TDM shall be distinguished from information retrieval (hereinafter: IR) which refers to the retrieval of documents that are relevant to a user's information needs, and not the retrieval of information itself.[42] Although in IR, the user is entrusted with the task of selecting the desired information from documents, in TDM, the direct retrieval of information from documents is done by the machine.[43] In particular, it is the machine that also attempts to discover from documents new patterns of information, which is useful, non-trivial, and unknown from a previously action.[44]

2.2 Overview of the TDM Process

The TDM process contains three main stages. These include accessing the content, extraction and/or copying of content, as well as mining of text and data and knowledge discovery. While these stages provide a general framework for harvesting, the specific techniques and tools used can vary depending on the nature of the data and the objectives of the analysis, with the only common thing being the analysis and extraction of associations between concepts to identify patterns and relations.[45] The diversity of TDM techniques and practices makes it virtually impossible to provide a general and exhaustive description of how TDM operates.[46]

---

[40] Ramez Elmasri and Shamkant Navathe, *Fundamentals of Database Systems* (7th edition, Pearson 2016) 1069-1110.
[41] Mohammad Asim khan and Sharik Ahmad, 'An Emergence of AI in Data Mining and KDD: ANN its Strength & Weakness' (2020) 9(1) International Journal of Recent Technology and Engineering 380, 380.
[42] Hearst (n 39).
[43] Sanaa Alwidian, Hani Bani-Salameh and Alaa Alslaity, 'Text data mining: a proposed framework and future perspectives' (2015) 18(2) International Journal of Business Information Systems 127, 129.
[44] ibid.
[45] Eleonora Rosati, 'Copyright as an obstacle or an enabler? A European perspective on text and data mining and its role in the development of AI creativity' (2019) 27(2) Asia Pacific Law Review 198, 206.
[46] ibid.

Nevertheless, by simplifying the process, it becomes possible to discern three typical-though not always required - steps within TDM processes.

2.2.1 Access to Content

The first and propaedeutic step is accessing the content, in relation to which TDM is to be performed. Depending on the type of mining, which is to be done, access to the content can be in the form of text or data.[47] Usually at this stage, general, raw, processed, and target data are interconnected to each other. The content can be made available either open source or closed behind paid walls.[48] Nevertheless, even if the content could be found online, this does not necessarily imply that such is immune from copyright protection or that it can be freely mined by AI developers.

2.2.2 Extraction and/or Copying of Content

The second step is extracting and/or copying content which is where the legal issues arise. This may give result in copyright infringement unless the act is authorised under a license or a legal exception. When direct access is available, no extraction or additional downloads are required upon mining. This would entail no necessary need to reproduce the work in a separate location in order to commence the mining process; thereby, making it possible to avoid possible copyright infringement.[49] However, in cases where the mined work comes from sources which do not provide direct access, a copyright violation may be established. It is important to note that not all AI systems retain a complete or partial copy of their training data. Instead, some utilise the training data to generate algorithmic tokens that are subsequently processed by a multitude of artificial neurons that make probabilistic decisions.[50]

---

[47] Artha Drmawan, 'Text and data mining exceptions in the development of generative AI models: What the EU member states could learn from the Japanese "nonenjoyment" purposes?' (2023) 27 The Journal of World Intellectual Property 44, 47.
[48] Kalpana Tyagi, 'Copyright, text & data mining and the innovation dimension of generative AI' (2024) 19(7) Journal of Intellectual Law & Practice 557, 563.
[49] Jean-Paul Triaille, Jérôme de Meeûs d'Argenteuil and Amélie de Francquen, 'Study on the Legal Framework of Text and Data Mining (TDM)' European Union, March 2014, 47.
[50] David W. Opderbeck, 'Copyright in AI Training Data: A Human-Centered Approach' (2024) 76(4) Oklahoma Law Review 1, 9.

One of the most common methods for acquiring copyrighted images for TDM purposes is through web scraping. Web scrapping can be defined as software programs which are '*designed to automate the downloading and parsing of the data from the Web'.*[51] This technique interacts with web communication protocols to swiftly extract data from web pages.[52] The scabbed data could be then used for a multitude of reasons, including training AI systems which generate art. The process begins with sending an automated HTTP request to the targeted website server, which at a later point responds with an HTML document containing the requested content, including entire structures of the webpage, such as images.[53] Then the HTML document is analysed, with specific elements and data extracted, through a process known as parsing.[54] Web scraping tools, such as Scrapy or BeautifulSoup, are typical examples used to crawl and temporarily copy images into local storage for further analysis. Although this process eases a timely and relatively costly collection of data, it also raises significant copyright concerns, particularly when dealing with authorship material.

Another possibility for extracting copyrighted materials is through the implementation of Application Programming Interfaces (hereinafter: APIs) and data repositories. These two methods significantly enhance image extraction for AI companies by providing structured and efficient methods for acquiring large datasets. APIs allow any person with some programming skills to obtain huge volumes of data about the content of a platform.[55] It enables the targeted and systematic retrieval of images based on specific criteria, such as keywords and metadata, and supports bulk data retrieval through features like batch requests and pagination.[56] This approach streamlines the process of gathering extensive image collections. Data repositories, on the other hand, offer precompiled, annotated datasets that facilitate bulk downloading and include preprocessing and annotations, which save considerable time and resources.

[51] Jay M Patel, *Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale* (Apress 2020) xvii.
[52] Darren Ang, 'The Web Scraper's World of Copyright Exceptions and Contractual Overrides' (2022) 13 Singapore Law Review 1, 1-16.
[53] R Suganthi and others, 'Utilizing Web Scraping for Big Data: An Exploratory Analysis' (2023) 44(6) Journal of Propulsion Technology 6150, 6151.
[54] ibid 6152.
[55] Rebekah Tromble, 'Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age' (2021) 7(1) Sage Journals 1, 1-8.
[56] A Velez-Estevez and others, 'New trends in bibliometric APIs: A comparative analysis' (2023) 60(40) Information Processing and Management 1, 4.

These repositories provide ready-to-use datasets with labels and metadata essential for model training and benchmarking.[57] Together, APIs and repositories enable AI companies to efficiently collect and utilize high-quality image data, supporting the development and refinement of machine learning models that are focused on the creation of AI art.

Another key method for extracting content from images is Convolutional Neural Networks (hereinafter: CNNs), which are designed to extract high-level visual features from images, using convolutional filters, also known as kernels.[58] Once an image is fed into a CNN, the network applies a series of convolutional filters to detect patterns such as textures, edges, and shapes. These filters generate feature maps, which are abstracted representations of the visual characteristics of the image.[59] The feature extraction process copies the original image into memory, where it undergoes several convolutional and pooling operations. Once the relevant features are captured in a vector form, the original image is typically discarded, and only the feature embeddings are retained. This technical process entails the act of copying the image into memory, processing it, and extracting its usable data for further analysis.[60]

Additional methods for when AI companies need to extract or copy content is the purchase or access to commercial image datasets and archives which constitute precompiled collections of images that could be used in model training. Particularly, this method allows AI companies to acquire large-scale, high-quality datasets that have already been curated and often come with clear licensing agreements; thereby, reducing the legal risk of using copyrighted material without permission. Additionally, AI companies also collaborate with other organizations or institutions to access proprietary or exclusive image datasets. By leveraging these exclusive datasets, companies can obtain access to high-quality images that are often unavailable through public sources. This method will not only boost the performance of their models but also give them a competitive advantage in model development and deployment.

---

[57] Fred Prior and others, 'Open access image repositories: high-quality data to enable machine learning research' (2020) 75(1) Clinical Radiology 7, 7-12.
[58] Xia Zhao, 'A review of convolutional neural networks in computer vision' (2024) 57 Artificial Intelligence Review 1, 98.
[59] Rikiya Yamashita and others, 'Convolutional neural networks: an overview and application in radiology' (2018) 9 Insights Imaging 611, 628.
[60] D. R. Sarvamangala and Raghavendra V. Kulkarni, 'Convolutional neural networks in medical image understanding: a survey' (2022) 15(1) Evol Intell 1, 1-2.

2.2.3 Mining and Knowledge Discovery

After the extraction and/or copying is done, the last step involves the mining of the text and data and the knowledge discovery. Ultimately, this is the stage where potential documents are discovered and converted into a machine-readable format, enabling the extraction of structured data. This ensures that the model learns efficiently and that the values are at a suitable range. This stage typically begins with the pre-processing of relevant text and data. In many cases, the text is unstructured, and as such, the input data may be 'noisy', redundant or error-prone, *ergo* a necessary cleanup and preprocessing of the data may be required. Both cleaning and parsing are important steps, given that text is often created and stored in formats that require refinement in a computer-readable format. Following the initial processing, the extraction stage involves a few additional steps. Tokenisation relates to breaking the document into its constituent terms and eventually removing certain characters, such as punctuation in order to be used by the data mining algorithms, making the data aggregated and consolidated.[61] Synonyms are also identified, and this could be achieved through a multitude of methods, such as linguistic resources, data-driven statistical techniques, or any other rule-based approaches.[62] The subsequent step, text transformation, was predominantly employed by traditional AI models, which did not directly process raw texts. In this step, the text is converted into a structured representation, usually as a bag of words or in vector space, based on term frequency.[63] Finally, the process involves defining equivalence classes. This stage often requires assigning weights to certain classes to reflect their relative importance, which aids in distinguishing between them more effectively.

In this stage, the models also learn to associate specific words and phrases with visual elements. For example, if one uses the word 'sunset', this may trigger the use of warm colours like yellow and orange. Respectively, if one enters the word 'forest', this will automatically make a correlation with various shades of green to evoke the natural, lush environment typically associated with forests and incorporate different elements like tall trees, moss-covered ground, and dappled sunlight filtering through the leaves. This ability to map textual

[61] Rosati (n 45) 212.
[62] ibid.
[63] ibid.

inputs to visual outputs illustrates the practical application of knowledge discovery in transforming text data into meaningful visual information. This last step is essential to the realisation of the underlying goal of predictive TDM, which goes beyond mere information extraction and extends to knowledge discovery.

2.3 LAION and its Role Within TDM

With the introduction of generative AI tools in 2022-2024, the idea of what AI can do and could do in the future changed entirely.[64] LAION, a free and open-access database made by a small non-profit team of researchers in Germany, played a paramount role in this shift. Namely, it provided access to high volumes of image-text pars which allowed developers of AI art tools, to facilitate the creation of more advanced, generative art. Making use of internet crawling tools, LAION initially copied 400 million image-text pairs, which later expanded to 5 billion pairs, resulting in the creation of LAION-400M and respectively LAION-5B.[65] The critical process of associating text and images within LAION-5B is performed using Contrastive Language-Image Pretraining (hereinafter: CLIP). In simple terms, CLIP is a neural network model, developed by another big company OpenAI, which was designed to learn from image-text connections by processing data containing images and their respective descriptions. The goal behind it is to enable accurate image retrieval based on text queries and vice versa.[66]

Images in LAION are represented as matrices of pixel values.[67] Unlike humans, deep learning models perceive images differently. They discern only millions of pixel values for red, green and blue. These three colour components are usually represented as integer values between 0 and 255 (in an 8-bit per channel image).[68] Before being fed into models like CLIP, these pixel values are normalized to ensure the data is in a suitable range for the model to process effectively during training.

---

[64] Renana Peres and others, 'On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice' (2023) 40(2) International Journal of Research in Marketing 269, 269-275.
[65] Laura Burger, 'LAION: Image Data, AI, and Dispossession' (Master's Thesis, Utrecht University 2023) 20.
[66] ibid 21.
[67] Mattia Casini and others 'Current trends on the use of deep learning methods for image analysis in energy applications' (2024) 15 Energy and AI 1, 3.
[68] Anson Pinhero and others, 'Malware detection employed by visualization and deep neural network' (2021) 105 Computers & Security 1, 6.

To further enhance its usability, LAION continually processes larger datasets and incorporates a variety of filtering mechanisms. [69] One such feature is the aesthetic score filter, which allows users to create custom datasets by selecting images based on their aesthetic value. When speaking about training AI art-generating models, the higher the quality of the image, the better the result will be. Generally, the task of LAION is to put the dataset together and narrow it down by filtering out any watermarked and non-aesthetic images, such as logos.[70] A recent study on LAION's 12 million image dataset (LAION-Aesthetics v2 6+), revealed that that approximately half of the images were sourced from just 1,000 domains.[71] The ones appearing most often were Pinterest, followed by WordPress-hosted blogs and the print and poster-selling website Fine Art America.[72] The next most frequently used websites included stock photo libraries, social media platforms and e-commerce websites.[73] As an intermediary, between data preparation and implementation, LAOIN ensures that the datasets are properly curated and filtered before being used for model training.[74]

2.4 AI Companies Leveraging Datasets

Not many AI companies openly disclose the datasets they use for training their models. This lack of transparency is often tied to several strategic and legal concerns. First and foremost, the data used to train AI models represents a key asset and is inextricably tied up to the commercial interests of large tech companies. For this purpose, AI providers are acutely protective of this information to maintain a competitive advantage and safeguard proprietary knowledge. Revealing their data sources could expose valuable insights to competitors, diminishing their unique position in the market. Furthermore, any disclosure of datasets increases the risk of legal challenges, particularly because many AI models are trained on

[69] Schuhmann et al, "Laion-5B," 30.
[70] Heikkilä (n 19).
[71] Andy Baio, 'Exploring 12 Million of the 2.3 Billion Images.' (*WAXY*, 30 August 2022) <https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/> accessed 7 July 2024.
[72] ibid.
[73] ibid.
[74] Burger (n 65) 20.

publicly available but unlicensed content. This raises significant questions about intellectual property rights and the use of copyrighted material without explicit permission.

For instance, OpenAI, is one of the big market players, which has not fully revealed the specific dataset behind DALL-E 3 - the newer version of DALL-E. Named after Salvador Dali and Pixar's WALL-E,[75] this image generator leverages a neural network to interpret text and produce photorealistic images through the conversion of words into a set of features. It is indeed known that the model was trained on millions of text-image pairs from the internet, such as images with text captions, allowing users to produce synthetic digital images that are based on their text input, also known as prompts. OpenAI has been cautious about revealing exact details about the sources of its training data. However, judging by the quality and breadth of images it creates, it could be assumed that the training data is likely to include an enormous compilation of content, potentially even copyrighted one.
Similarly, Midjourney, another innovative AI tool that converts textual instructions into intricate digital images, has kept the details of its training datasets under wraps.[76] Midjourney not only generates static images but also offers the capability to create animations and videos from platform-generated images or user-uploaded content.[77]

One of the very few AI companies declaring publicly which dataset they make use of upon training their algorithms is Stability AI. The company has publicly disclosed their use of the LAION dataset, a decision that has drawn a lot of attention in light of the ongoing copyright concerns. One of their most popular tools is Stable Diffusion, a deep-learning-based model capable of producing detailed images from textual descriptions. This model, similar to OpenAI's DALL-E, uses a large dataset of text-image pairs to generate synthetic digital art based on user inputs. Specifically, it is built upon the works of Latent Diffusion Models

[75] Joe Foley, 'DALL-E 2 access is now open to everyone, and that's kind of terrifying' (*Creative Bloq*, 1 October 2022) <https://www.creativebloq.com/news/dalle2-access-open-to-all> accessed 18 July 2024.
[76] 'How Does Midjourney Work? Explaining the Generative AI Platform; (Design Dash, 6 March 2024) <https://designdash.com/2024/03/06/how-does-midjourney-work-explaining-the-generative-ai-platform/#:~:text=Journal%20Prompt&text=By%20leveraging%20advanced%20artificial%20intelligence,with%20unprecedented%20ease%20and%20flexibility> accessed 7 July 2024.
[77] Dena Magdy Hanna, 'The Use of Artificial Intelligence Art Generator "Midjourney" in Artistic and Advertising Creativity' (2023) 4(2) Journal of Design Science and Applied Arts 42, 44.

(hereinafter: LDMs) combined with insights from conditional Diffusion Models (hereinafter: DMs), enabling it to create rich and contextually relevant visuals.

However, the use of Stable Diffusion for creating art has not been without any controversy. By disclosing their reliance on LAION, Stability AI has brought into focus potential copyright issues, as many of these images may be protected by intellectual property laws. Namely, the growing concerns about the eventual copyright infringement deriving from the misuse of protected materials in the training of these models have already surfaced in a number of court actions. To further explicate the matter, the thesis will focus on Stability AI and the legal cases revolving around the latter issue due to the appreciable amount of available information about the company's operations and its eventual regulatory challenges.

.

2.5 Copyright Concerns Addressed in Court

It is hard to pin down where exactly the line between open data and proprietary content is drawn. While LAION's code is publicly available on GitHub, and anyone can browse through the 5 billion images collected, concerns over copyright violations loom large. The underpinning problems, related to copyright violations, are reflected in a number of ongoing legal cases, currently occupying the legally nebulous space. The question of copyright is foundational in the pending lawsuit between Getty Images a leading provider of visual content, renowned for its extensive archive of licensed editorial and creative imagery, and Stability AI. [78]

The first lawsuit was filed in Delaware and accused Stability AI of misusing over 12 million copyrighted images and their metadata without permission; thereby, violating the site's terms of service, particularly in relation to web scraping. The defendants responded that the Delaware court lacked jurisdiction over its UK entity and requested the case to be transferred to the Northern District of California, where a related lawsuit is already ongoing. Getty noted that, even though it has provided licenses to top tech innovators for AI training that respects intellectual property and personal rights, Stability AI has bypassed these licensing

[78] *Getty Images* (n 18).

opportunities and disregarded any already established legal protections to prioritise its own commercial interests.[79] Additionally, the AI model reportedly created a modified version of Getty's watermark, which may have been used to either disguise the unauthorized use of Getty Images' content or to make it easier to infringe on their copyright without detection

Following the case filed in the US, Getty initiated a lawsuit against Stable Diffusion[80] in the UK, claiming that the latter party '*unlawfully copied and processed millions of copyright-protected images*', to train its software, and subsequently generate synthetic images without obtaining consent or providing compensation.[81] Through this claim, Getty has requested London's High Court of England Wales to issue an injunction banning Stability AI from selling its AI image generation technology in the country.[82] Getty alleged that the action of Stable Diffusion constituted copyright infringement since Stability AI had downloaded and used Getty's copyrighted works on servers or computers in the UK during the development and training of its AI model.

Whilst the legal disputes are unfolding, Getty has launched a new service called Generative AI, developed in collaboration with Nvidia, which allows users to create new images using Getty's extensive library of professionally taken photos.[83] Getty claims this service is distinct from other AI-generated image platforms as it was not trained on publicly available internet images. The new service includes a royalty-sharing model, compensating contributors for their images used in the training set and offering full indemnification for commercial use; thence, addressing the intellectual property concerns that have made some businesses hesitant to adopt generative AI tools.

Nevertheless, the legal battles did not stop there. As the legal landscape surrounding AI and copyright continues to grow, visual artists have achieved a partial victory in court. The case

[79] James Vincent, 'Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its Content' (The Verge, 17 January 2023) <https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit> accessed 8 August 2024.
[80] *Getty Images (US) Inc and others v Stability AI Ltd* [2023] EWHC 000007 (Ch).
[81] Vincent (n 79).
[82] Sam Tobin, 'Getty asks London court to stop UK sales of Stability AI system' (*Reuters*, 1 June 2023) <https://www.reuters.com/technology/getty-asks-london-court-stop-uk-sales-stability-ai-system-2023-06-01/> (accessed 3 August 2024).
[83] 'Getty Images Introduces Updated AI Model with Increased Speed, Quality, and Accuracy' (*Getty Images*, 29 July 2024) <https://newsroom.gettyimages.com/en/getty-images/getty-images-introduces-updated-ai-model-with-increased-speed-quality-and-accuracy> accessed 18 July 2024.

initially brought up by Sarah Anderson, Kelly McKernan, and Karla Ortiz alleged that Stability AI's Stable Diffusion software infringed their copyrights by illegally acquiring billions of copyrighted images, including their own. These unlawfully obtained images were stored as compressed copies and used to create a software library for various AI platforms, such as Stability AI, Midjourney, Runway AI and DeviantArt, the latter two being other AI-powered creative tools.[84] Notably, they alleged that those image generators were trained using Stable Diffusion and the LAION dataset, both of which included the plaintiffs' copyrighted works.[85] As already noted, to facilitate this, LAION provided a utility called 'img2dataset,' which enabled users to generate local copies of these images by extracting them from the URLs embedded in the metadata records.[86] This process ensured that those utilizing the dataset for training machine-learning models could access and store the necessary images independently, by creating their own local repositories for further use, without actually receiving access to the actual images.

Stable Diffusion; however, rebutted the latter allegation, arguing that their training process involves refining millions of parameters that collectively form a learned representation of visual elements, such as lines, colours, shades, and other attributes.[87] Although many of the original claims were dismissed, the artists were permitted to refile their case. In November 2023, the three original plaintiffs were joined by seven others, presenting revised arguments, that Stability AI's Stable Diffusion model unlawfully contained 'compressed copies' of their artwork. At this stage, the court found that the design of Stable Diffusion could indeed lead to copyright infringement by end users, suggesting that it might have been deliberately structured to facilitate such violations. As of yet, on 12th August 2024, U.S.District Judge William Orrick allowed the copyright infringement claims, addressing the unauthorized inclusion of artists' works in image generation systems, to proceed.

[84] Carlie Porterfield, 'Judge dismisses most of artists' copyright lawsuit against AI image generators' (*The Art Newspaper*, 31 October 2023) <https://www.theartnewspaper.com/2023/10/31/california-judge-dismisses-most-of-artists-ai-copyright-lawsuit> accessed 4 August 2024.
[85] *Andersen v Stability AI Ltd* (ND Cal, 3:23-cv-00201, 2023).
[86] ibid [14].
[87] Motion to Dismiss filed by Stability AI Ltd, at 1, *Andersen* (No 23-cv-00201-WHO, ND Cal, 18 April 2023).

As regards all these court charges, indirectly touching upon LAION and the access to data it has provided, LAION claims that each image incorporated in its dataset is independently protected by its own copyright.[88] Rather than directly hosting the images, LAION merely indexes URLs that link to the content. In fact, by doing so, LAION skirts any kind of liability, relying on the notion that the images are publicly accessible and remain under the ownership of their original creators.[89] This creative freedom, allows database curators, model developers, AI art creators, and scrapers, to freely create without giving credit or compensation to the IP holders. The lack of consent given by artists, who in many instances see AI works replicating their style, or their name attributed to art they did not create, is justified on the grounds that AI models do not store images. Yet, the freedom bestowed upon AI companies conflicts with the rights of the artists who have no means to prevent this from occurring or to have their works removed. In some instances, removal requests of the affected parties have remained completely ignored.

Without clear legal guidance, companies such as Stability AI will continue benefiting from harvesting text and data, rationalising their actions on the basis that such activities are carried out for the purpose of academic research. Indirectly, companies involved in the creation of AI art sidestep addressing intellectual property concerns, by approaching it from a scientific research standpoint, rather than a profit-orientated one.[90] With respect to LAION, on their website, the company stipulates that their datasets are intended solely for research purposes. However, this raises questions as to their partnership with Stability AI, which is a purely private company, whose users utilise the service to obtain monetary gains. While this partnership can be reframed by emphasising that the art tool Stable Diffusion was developed by university researchers with funding from the Stability AI; thereby, creating a collaboration between the private entity and academia, this does not diminish the significant influence of the private company over the entire process. The activities of big AI companies, hiding behind the guise of academic research, do not negate the fact that they frivolously engage with authorship material that was not initially intended as training data. Simply put, their engagement with

[88] Christoph Schuhmann and others, 'LAION-5B: An open large-scale dataset for training next generation image-text models' (2022) NeurIPS 2022 Datasets and Benchmarks 1, 1-17.
[89] Burger (n 65) 29.
[90] ibid 31.

protected material extends far beyond simply accessing the content of the image. Couldry and Mejias' discussion of data colonialism and its threat to the 'bare self' highlights how systems that quantify and abstract human experiences increasingly shape not only the economic value of individuals but also the core of their identities. In the context of AI-generated art, this concept becomes particularly relevant, as these systems commodify human creativity, turning it into a resource to be exploited by those with the technological means to do so.[91]

Although the judgements are yet to be delivered, what is known is that their outcomes will have a substantial influence on how the law will navigate the intersection of human-generated art and the expanding capabilities of AI systems, particularly concerning artists' rights. Authors argue that AI-generated images closely mimic their personal artistic styles, often intentionally incorporating their creations into AI systems, without proper authorisation. The key question is whether AI can coexist with human creativity, potentially enhancing artistic expression, rather than replacing it, and whether the law can adequately balance innovation in AI with the protection of artists' intellectual property. Once decided, these cases will serve as *stare deciris.*

[91] Nick Couldry and Ulises Mejias, 'Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject' (2019) 20(4) Television & New Media 336, 343.

CHAPTER 3

WHAT ARE THE POSSIBLE COPYRIGHT ISSUES ARISING FROM HARVESTING AND TDM?

With AI image generators being trained on large datasets, pressing concerns have emerged as to whether they employ authorship work within the training processes and if so, whether the rights of creators are respected. Antagonists argue that these machines often scrape original art, detaching it from its context and original intent in order to create new images.[92] By removing artists from their artwork, algorithms strip away both credit and autonomy, disregarding the years they spent honing their crafts. A major concern within the art community is that AI systems have been developed without considering creators as stakeholders in the art found online which is utilised for training data. Moreover, some AI models are specifically designed to remove watermarks and signatures, completely circumventing the intellectual property rights of authors. This issue left unaddressed has snowballed into a menace that continues to grow with the increasing popularity and use of AI, turning the mere creation of an image into a battleground of copyright.

The intersection of AI-generated art and intellectual property law has become a legally murky territory, enough to scare many creators and legal experts alike. Describing the legal issues related to art-generating machines as a 'washing machine of intellectual property' is one way to frame the situation.[93] As is often the case with rapidly advancing technology that outpaces any existing legal frameworks, there is no definitive, binding stance on the copyright challenges that emerge when these machines chew up human-created art and then spit out new works based on what they have previously absorbed.[94]

3.1 What Was the Legal Pathway Leading to the Adoption of the DSM Directive?

Statistics indicate that Europe falls behind the United States and Asia, particularly China, in the number of publications and patents related to TDM.[95] This stagnated development of TDM in Europe could be based on several factors such as the legal uncertainties in the EU, fragmentations within national legal frameworks, lack of awareness and expertise, and various

[92] Yuhua Luo (ed), 'Cooperative Design, Visualization, and Engineering' (20th International Conference on Cooperative Design, Visualization and Engineering, CDVE 2023, Mallorca, 1-4 October 2023 39 <https://link.springer.com/book/10.1007/978-3-031-43815-8> accessed 7 August 2024.
[93] Luke Plunkett, 'AI Creating 'Art' Is An Ethical And Copyright Nightmare' (*Kotaku*) <https://kotaku.com/ai-art-dall-e-midjourney-stable-diffusion-copyright-1849388060> accessed 8 August 2024.
[94] ibid.
[95] Filippov (n 37) 1-3.

other infrastructural challenges.[96] According to the European Commission's Impact Assessment accompanying the Proposal for the DSM Directive,[97] there was a broad consensus that TDM is still in its early stages, particularly within the non-commercial research sector, such as universities and research institutes.[98]

If one questions how TDM was being regulated in Europe, the UK was the first Member State (more recently also France, Germany, and Estonia which also have laws addressing this) to incorporate a specific copyright exception, permitting text and data analysis for non-commercial research.[99] On 14 September 2016, the European Commission published a Proposal for the DSM Directive which aimed towards fostering research and innovation. TDM was considered a key enabler given its seminal role in driving a generation of new knowledge which elevates the overall quality of research. Among its provisions, the Proposal introduced one mandatory exception in Article 3, allowing research organizations to reproduce and extract content for the purpose of conducting TDM for scientific research on works and other materials to which they have obtained lawful access. Article 3 was implemented in response to the growing reliance on digital technology in the scientific sphere, amid concerns that the European Union's competitive position as a research hub could be jeopardized without clarifying the legal uncertainty surrounding TDM activities.[100] The idea was to thus guarantee EU scientific leadership and competitiveness, particularly with respect to third countries.[101] Unlike other countries such as the UK, whose TDM exception applies more broadly, the Proposal covered both commercial and non-commercial uses,[102] but limited itself to a narrower range of beneficiaries. This restriction stemmed from the wording of Recital 11 which covered only public universities and research organisations. However, it failed to

[96] ibid.
[97] Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market COM (2016) 593 final.
[98] European Commission, Commission Staff Working Document, SWD(2022)366, 29 November 2022 16-17.
[99] Rosati (n 15) 6.
[100] Maria Bottis and others, 'Text and Data Mining in Directive 2019/790/EU Enhancing Web-Harvesting and Web-Archiving in Libraries and Archives' (2019) 9(3) Open Journal of Philosophy 369, 376.
[101] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Towards a modern, more European copyright framework, COM(2015)626, 9 December 2015, 7.
[102] Hargreaves and others (n 34) 3.

delineate how this limitation would be applied in practice.[103] The situation with individual researchers without a specific link to an institution also remained unclear. Such constraints risked creating a provision which could quickly become ineffective, specifically in the context of AI, as well as other vital research activities falling outside the scope of activities of many public institutions. By restricting access solely to research organizations, the Proposed Directive potentially excluded a broad range of essential research initiatives. This also posed major challenges to the right to research.[104] The term 'research organizations' incorporated within Article 3, refered to universities, research institutes, and non-profit or public interest organizations that are heavily involved in research.[105] While the draft of the DSM Directive did not explicitly exclude public-private partnerships from the TDM exception, it failed to clarify that the exception would not apply if a commercial entity exerts decisive influence and control over the research organization.[106] Not only this, but the formulation left out a reference to start-ups and innovations that conduct TDM for commercial purposes; thence, subjecting their TDM activities to the exclusive rights of authors. All this explains why after presenting the TDM exception, many controversial views were expressed as to its wording and narrow scope of application.

Later, in February 2020, the European Commission announced its ambition towards a digital strategy for the EU setting out the objectives that had to be achieved in two ways - one on data[107] and one on Europe's digital future[108] - both compiled in a White Paper on AI.[109] This transformative change targeted many sectors within the EU.[110] The European Commission's

[103] Christophe Geiger, 'The Missing Goal-Scorers in the Artificial Intelligence Team: Of Big Data, the Fundamental Right to Research and the failed Text and Data Mining Limitations in the CSDM Directive' (2021) 66 PIJIP/TLS Research Paper Series 383, 389.
[104] ibid 390.
[105] European Alliance for Research Excellence, 'Open letter to European Commission, Maximizing the benefits of Artificial Intelligence through future-proof rules on Text and Data Mining' (9 April 2018).
[106] Rosati (n 15) 8.
[107] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - A European strategy for data, COM(2020)66, 19 February 2020 19.
[108] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Shaping Europe's digital future, COM(2020)67, 19 February 2020.
[109] White Paper (n 14).
[110] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Shaping Europe's digital future' COM(2020), 19 February 2020 1.

action plan on intellectual property, released in late November 2020, highlighted the urgent need to update and modernize the legal framework for intellectual property in response to the challenges of the digital age.[111] The document thus stressed that the ongoing technological revolution - encompassing data economy, AI, emerging technologies and the Internet of Things (hereinafter: the IoT), along with new business models - presents an unparalleled opportunity to rethink and enhance the way intangible assets are protected. In response, the Commission sought a comprehensive, forward-looking approach by introducing appropriate legislation to meet the evolving digital needs.

Following the legal contentions, on 6 June 2019, the DSM Directive entered into force, marking a step towards the harmonisation of copyright in the EU.[112] With the adoption of the final text, the European Commission also wanted to boost the overall competitiveness of the EU copyright and related rights not only internally to TDM but also vis-à-vis third countries, such as Japan and the US, which have already dealt with TDM issues in their IP regimes and remain the EU's main trading partners.[113] The overall objective of the Directive was set to foster '*innovation, creativity, investment and production of new content, also in the digital environment, in order to avoid fragmentation of the internal market*'.[114]

3.2 How Is TDM Addressed in the DSM Directive?

Article 2(2) DSM Directive defines TDM as '*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*'. As already noted, TDM utilizes automated software tools that allow individuals with the requisite expertise to gather vast amounts of text and data, such as Big Data, for the purpose of relealing new insights and patterns. In this context, digital automated computation analysis shall, besides text, include sounds, images, or data.[115] The definition provided is broad and aptly identifies the potential of a tool capable of

[111] European Commission, 'Making the most of the EU's innovative potential, An intellectual property action plan to support the EU's recovery and resilience' COM(2020)760, 25 November 2020 2.
[112] Federico Ferri, 'The dark side(s) of the EU Directive on copyright and related rights in the Digital Single Market' (2020) 7(1-4) China-EU Law Journal 21, 26.
[113] Impact Assessment (n 35) 27.
[114] DSM Directive, recital 1 and 2.
[115] DSM Directive, recital 8.

autonomously or semi-autonomously analysing datasets. Hence, it could be said that it is forward-looking as it covers and regulates most areas of AI and machine learning, whether currently known or yet to be developed, as long as they rely on data analytics.[116] However, when such a broad definition is applied within a narrowly defined exception, as in the case at hand, it may restrict rather than advance technological and cultural practices, contrary to the initial intent of the drafters.[117]

The idea behind the introduction of Article 4 within the DSM Directive was to address the negative appraisal of the Proposal. This stems from Recital 18 which stipulates 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*.'

Both exceptions within Articles 3 and 4 DSM Directive had to be transposed by Member States within their national laws. The two articles delineate that not every act of TDM requires obtaining prior authorisation, especially if the activities fall within the exceptions and limitations outlined in the law. To start with, Article 3 DSM Directive introduced a new copyright exception for research organizations and cultural heritage institutions, allowing them to reproduce and extract content for scientific research. This Article permitted TDM out on lawfully accessed works and other subject matter, such as databases and press publications.[118] However, it excluded computer programs protected under the Software Directive[119], which may require a license for certain restricted acts.[120] Since nowadays TDM has become a vital tool for research, both public and private organizations were given the right to utilise it.[121] The new Article 4 DSM Directive, which mirrors Article 3 DSM Directive, is

[116] Thomas Margoni and Martin Kretschmer, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2020) 71(8) GRUR International 685, 687.
[117] ibid.
[118] Kristina Christensen, 'A European solution for Text and Data Mining in the development of creative Artificial Intelligence With a specific focus on articles 3 and 4 of the Digital Single Market Directive' (2021) 4(2) Stockholm Intellectual Property Law Review 18, 28.
[119] Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs [2009] OJ L111/16 (Software Directive).
[120] ibid art 4.
[121] DSM Directive, recital 8.

broader in scope and permits anyone with lawful access to works and other subject matter, including databases, press publications, and computer programs, to engage in TDM for any type of purpose, including commercial ones. Although Article 4 DSM Directive is open to all kinds of beneficiaries and uses, rightsholders can still reserve their rights through opt-out or contract-out mechanisms.[122] Notably, both exceptions also allow for TDM to be carried out for commercial purposes.

By allowing the replication, storage, and processing of training data to facilitate the development of new solutions under Article 3 and 4 DSM Directive, this legislative instrument aimed at establishing a uniform framework across the Union, ensuring that system developers can lawfully engage in TDM activities. Recital 5 DSM Directive states that already existing exceptions and limitations under EU law should continue to apply to activities such as text and data mining, education, and preservation, provided they do not restrict the mandatory exceptions or limitations outlined in Articles 3 and 4 DSM Directive. Another positive impact of the DSM Directive is that it allows the unauthorised use of TDM under specific circumstances which introduced important changes to the long-established copyright principles. It also strongly supports public interest by encouraging the creation of new knowledge through TDM, which might otherwise be hindered by prohibitive transaction costs. At the same time, the Directive ensures that rightsholders can share in the economic value of their works by allowing them to claim remuneration for the actual use of their protected materials, as both exceptions require lawful access but do not automatically grant it.

3.3 How Is Article 4 DSM Directive Addressing the Copyright Challenges?

Article 4 DSM Directive is designed to create a standardised level playing field across the EU, allowing broad access to TDM activities. Not only was the introduction of this provision focused on stimulating innovation, but it was also offering greater legal certainty for TDM activities that are not covered by Article 3 DSM Directive. In any event, the reservation within Article 4 DSM Directive, does not affect the substance of Article 3 DSM Directive. This derogation successfully encompassed not only research purposes but any other TDM activity

[122] Margoni and Kretschmer (n 116) 686.

that falls within the definition of TDM stipulated under Article 2(2) DSM Directive. Recital 18 DSM Directive specifically denotes that Article 4 DSM Directive cannot unduly affect the licensing opportunities for rightholders regarding uses of their work falling outside of the ambit of Article 3 DSM Directive and of the existing exceptions and limitations part of the EU legal order.

To start with, Article 4(1) DSM Directive states that Member States shall establish an exception or limitation to the exclusive rights of authors, including those related to computer programs and database work, as well as the rights of database creators, and press publishers, allowing reproductions and extractions of legally accessible works and other materials for TDM. As per the new exception, all reproductions and extractions of works or other subject matter made for TDM purposes are exempt from the rightsholder's monopoly, provided that they are lawfully accessible, and the reproductions and extractions are kept only as long as necessary for the respective TDM.[123] This exception can, in principle, be invoked *a priori* in any machine learning project, covering all reproductions and extractions related to the analytical processing of protected training data. The provision does not impose any *ratione personae* or purpose limitation, meaning it applies broadly to all entities and for any purpose. However, Article 4 DSM Directive should not disproportionately limit rightholders' ability to license their works for uses beyond the scope of Article 3 DSM Directive and the existing exceptions and limitations within the EU's copyright framework. The goal is to ensure that while data use for AI training is essential, it should not overly restrict the authors' right to exploit their works

Further to the above, Article 4(3) DSM Directive introduced the so-called opt-out option, allowing the rightsholder to strike a balance between their interests and those of the potential users. Within paragraph 3 of the same Article, it is noted that the TDM limitation '*shall apply on condition that the use of works and other subject matter ... has not been expressly reserved by their rightsholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online'*. In particular, the opt-out could be exercised through technical measures, such as machine-readable means, metadata, or contractual

[123] DSM Directive, art 4(2).

agreements,[124] such as terms and conditions of a website or service.[125] Rightsholders may also issue unilateral declarations,[126] like disclaimers, to reserve the right to control reproductions and extractions for data analysis purposes.

The Directive sought to implement the exception in compliance with the three-step test, as outlined in Article 7 DSM Directive, which references Article 5(5) InfoSoc Directive[127], where the three-step test is provided. In fact, the three-step test is a long-established principle in copyright law, recognised not only in the Berne Convention[128] but also in the TRIPS Agreement[129] and the WIPO Copyright Treaty (hereinafter: WCT)[130.] This test, as incorporated in Article 5(5) InfoSoc Directive, has solidified its role as a foundational principle governing nearly all exceptions to intellectual property rights at the international and European levels and established three critical, cumulative criteria. The first criterion allows limitations to copyright law only in specific cases. The second criterion necessitates that such limitations do not conflict with the normal exploitation of the work; whereas, the third criterion mandates that these limitations should not cause unreasonable harm to the legitimate interests of the author, right-holder, or user, depending on the interpretation. Therefore, if a limitation fails to meet one of the conditions, the restriction does not pass the test and would have to be abolished.

3.4 Criteria for Establishing Copyright Protection

The three main international treaties in the field of copyright; namely, the Berne Convention, TRIPS Agreement and WCT, necessitate their contracting states, including the EU Member

[124] DSM Directive, art 7(1).
[125] European Commission, Legal and regulatory implications of Artificial Intelligence. The case of autonomous vehicles, m-health and data mining, JRC116235, 2019, 28.
[126] DSM Directive, recital 18.
[127] 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), art 5(5).
[128] Berne Convention for the Protection of Literary and Artistic Works (adopted 9 September 1886, entered into force 5 December 1887) 1161 UNTS 3 (Berne Convention), art 13 and 9(2).
[129] Agreement on Trade-Related Aspects of Intellectual Property Rights (adopted 15 April 1994, entered into force 1 January 1995) 1869 UNTS 299 (TRIPS), art 13.
[130] WIPO Copyright Treaty (adopted 20 December 1996, entered into force 6 March 2002) 2186 UNTS 121 (WCT), art 10.

States to provide international protection of eligible authors of 'works' according to the minimum standards set out therein.[131] The Berne Convention definition of 'literary and artistic works'[132] is well incorporated in TRIPS and WCT[133], as well as indirectly embodied within the EU legal order.

There are certain conditions that have to be satisfied upon granting copyright protection. As per the first requirement; namely, originality, the Berne Convention on multiple instances refers to 'original works'[134] but sets down no clear definition of this criteria. This requirement carries a dual implication. First, the work must constitute 'the author's own', meaning it cannot be a copy of another's creation. Second, and more crucially, it must represent an 'intellectual creation' by the author. In the *Levola* and *Cofemel* cases, the CJEU emphasized this principle, albeit in a somewhat circular fashion, by stating that '*the subject matter concerned must be original in the sense that it is the author's own intellectual creation*'.[135] It is important that the author '*was able to express his creative abilities in the production of the work by making free and creative choices'*.[136] There is no need for the author's personality and creativity to be distinguishable in the final output.[137] As per the jurisprudence of the Court, the availability of sufficient 'creative space' is already a strong indication of originality. The CJEU further elaborated on this in *Painer*[138] and *Funke Medien*[139] that intellectual creation inherently reflects originality, which in turn necessitates the making of personal and creative choices by the author. This interpretation was reaffirmed in the recent *Cofemel* ruling, solidifying the understanding that originality, as a cornerstone of copyright protection, must stem from the author's unique creative decisions.[140] It shall be also considered that some AI-

[131] European Commission, 'Trends and Developments in Artificial Intelligence - Challenges to the Intellectual Property Rights Framework' (September 2020) 67 < https://op.europa.eu/en/publication-detail/-/publication/394345a1-2ecf-11eb-b27b-01aa75ed71a1/language-en> accessed 18 August 2024.
[132] Berne Convention, art 2(1).
[133] TRIPS, art 9(1) and WCT, art 3.
[134] Berne Convention, art 2(3) and art 14bis.
[135] Case C-310/17 *Levola Hengelo v Smilde Foods BV* ECLI:EU:C:2018:899, para 36 and Case C-683/17 *Cofemel - Sociedade de Vestuário SA v G-Star Raw CV* ECLI:EU:C:2019:721, para 29.
[136] Case C-469/17 *Funke Medien NRW GmbH v Bundesrepublik Deutschland* ECLI:EU:C:2019:623, para. 19 and Case C-145/10 *Eva-Maria Painer v Standard Verlags GmbH and Others* ECLI:EU:C:2011:798 [2011] ECR I-12533, paras 87-88.
[137] European Commission (n 130) 73.
[138] *Painer* (n 136) [94].
[139] *Funke Medien* (n 136) [19]
[140] *Cofemel* (n 135) [30].

generated images involve a redaction stage in the creation process. The case *Painer* noted that this final stage may require a lot of creative choices, which depending on the circumstances, may suffice for finding originality in the production process.[141] With regards to databases, the criterion for originality '*is satisfied when, through the selection or arrangement of the data which it contains, its author expresses his creative ability in an original manner by making free and creative choices*'.[142] Nevertheless, economic investments cannot as such satisfy this criterion.

Another essential prerequisite for granting copyright is that the work must be a product of human intellectual effort. Both the Berne Convention and the EU copyright framework are fundamentally rooted in the tradition of author's rights, or *droit d'auteur*, wherein copyright protects original expressions that emanate directly from a human creator.[143] While EU copyright law does not explicitly state that a human creator is essential for copyright protection, the human-centred approach is particularly apparent in the jurisprudence of the CJEU on originality, which is deeply grounded in the idea of human engagement in creative acts that involve making 'creative choices'.[144] For instance, in *Painer*, the Court emphasized that through creative choices, the author of a portrait photograph can imprint the work with his 'personal touch'.[145]

Lastly, in order for copyright protection to be granted, the form of creativity needs to be expressed in the final product. The CJEU has ruled on numerous occasions that expression constitutes a *sine qua non* for granting copyright protection. For instance, in the cases *Infopaq* and *BSA*, the CJEU spelled out the requirement that the author must have '*express[ed] his creativity in an original manner*'.[146] Also, in *Painer*, the Court established that to have an original work, the author must be able to '*express his creative abilities in the production of the work by making free and creative choices*'.[147] Creative choices made by the author must be

[141] *Painer* (n 136) [116].
[142] Case C-604/10 *Football Dataco Ltd and Others v Yahoo! UK Ltd and Others* ECLI:EU:C:2012:33, paras 38- 39.
[143] European Commission (n 131) 69.
[144] ibid.
[145] *Painer* (n 136) [92].
[146] Case C-5/08 *Infopaq International A/A v Danske Dagblades Forening* ECLI:EU:C:2009:465 [2009] ECR I-6569, para 45 and Case C-393/09 *Bezpečnostní softwarová asociace -Svaz softwarové ochrany v Monisterstvo kultury* ECLI:EU:C:2010:816 [2010] ECR I-13971, para 50
[147] *Painer* (n 136) [89].

perceptible in the work itself; mere ideas or concepts that lack tangible form or expression do not qualify as protected works. Therefore, it could be said that a causal connection between the author's creative act and the expression in the form of the work is necessary for the expression criterion to be satisfied.

For this reason, in order to determine if TDM violates the copyright of original creators, the work that is being mined needs to first be evaluated against the established criteria of copyright protection. A comprehensive evaluation of the underlying work's copyright status is imperative for establishing the boundaries of permissible use. Only when one work is deemed to be protected can it be assessed if potential infringements materialise from harvesting practices.

3.5 Conditions for Legal Use of Artworks and Copyright Limitations

There has always been a contention between intellectual property and TDM techniques. As a general rule, TDM should not impinge upon the copyright of creators. If copyright is established for a work, it needs to be established in which situations its utilisation in AI systems is allowed and when such could lead to a copyright infringement.

One could argue that TDM falls outside the scope of intellectual property rights, and imposing any restriction on the activity would undermine the rationale of copyright protection and result in a restriction of the fundamental principles of freedom of expression and information, as enshrined in the European Convention of Human Rights (hereinafter: ECHR)[148] and the Charter of Fundamental Rights (hereinafter: CFR)[149].[150] In the chain of TDM activities, some actions touch upon intellectual property rights, and unless an authorisation is obtained, an infringement may be established.

[148] Convention for the Protection of Human Rights and Fundamental Freedoms as Amended by Protocols No.11 and No.14 (adopted on 4 November 1950, entered into force 3 September 1953) CETS No. 005 (ECHR), art 10.
[149] Charter of Fundamental Rights of the European Union [2012] OJ C 326/391 (CFR), art 11.
[150] DSM Directive, art 4.

Datasets often consist of individual numbers, facts, statistics, or other types of information that can be gathered. Some elements, like text, photographs, music, poems, or artworks, may qualify for copyright protection. However, where certain datasets are generated automatically, they fail to satisfy the requirement of originality. As a result, raw data, numbers, factual elements, as well as information already in the public domain, collected without human involvement, may not be eligible for copyright protection and could remain freely accessible. Hence, performing TDM on such data would not constitute a violation of copyright or related rights.

It needs to be highlighted that not all TDM activities involve copying or extraction. In order to bypass any legal complications, data mining operations should focus on sourcing data from legitimate channels, such as public domain works or data collected with the creator's consent.[151] The need for such activities largely depends on how the material will be used, the technical tools employed, and the scope of the mining process.[152] Additional legitimate sources include open-access datasets or works available under licenses like Creative Commons or open-source software licenses. Nonetheless, it should be acknowledged that relying solely on public domain works may not be always sufficient, as it could result not only in bias but the exclusion of vital information. In attempts to avoid those problems, system generators often scrap data from every available website, which, however, may potentially trigger copyright infringements.

Many scholars also note that if a computer has the right to read data, it is also allowed to analyse or else mine it. This concept, often summarized as '*the right to read is the right to mine*,' conveys the idea that simply accessing and processing information through TDM should not infringe upon copyright, provided that no unauthorized copying or distribution of the original work occurs.[153]

[151] Andres Guadamuz, 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs' (2024) 73(2) GRUP International 111, 116.
[152] Rosati (n 45) 198-200.
[153] Sean Flynn and others, 'Implementing user rights in the field of Artificial Intelligence: A call for international action' (2020) 42(7) European Intellectual Property Review 393, 397.

TDM tools that perform minimal copying of a few words or that crawl through data and process each item separately might operate without encroaching on the author's rights. This could be explained by the fact that copyright law does not protect data itself, but only original expressions within protected subject matter. While the *Infopaq* case recognised that even an 11-word excerpt could be copyright-protected, the European Commission pointed out that minimal extraction techniques, like counting the frequency of a specific word without retaining a copy, may not qualify as copying. In such cases, the activity would not require the rightsholder's permission, eliminating the need for a legal exception.[154] This could be also the case when data analysis software 'crawl' through text or data by processing them individually, analysing one word, pixel, or data point at a time, without copying entire texts.[155]

A question arises in terms of datasets when metadata is extracted rather than the work itself being directly copied. A notable example of where image links and text descriptions are extracted and not copied is the LAION dataset. Although during the extraction process, copies could be made, they are not retained in the dataset. One might argue that including links in a database could constitute the right of communication to the public. However, it is unlikely that these links would be communicated to the public in the strict sense, as the links themselves are not made public. Existing CJEU case law, delineating what the right of communication to the public is, suggests that the link itself will lead to a copy that has been made available by the owner, rather than infringe it.[156] However, there could still be a violation if the reproduction appears to not be authorised. Usually, after a model is trained, the copied data is not needed anymore, unless additional models are developed, based on the same information.

Many AI models, like LAION, are trained to make only a temporary copy which is turned into an abstract adaptation, and later used in the model for the creation of the final output. As already noted, during training, images are encoded into latent space, clustered with words, and processed through a noise-and-denoise cycle typical of diffusion models. The result is a

[154] Triaille, Meeûs d'Argenteuil and de Francquen (n 49) 31.
[155] ibid.
[156] See Case C-466/12 *Nils Svensson and Others v Retriever Sverige AB* ECLI:EU:C:2014:76, para 32 and Case C-160/15 *GS Media BV v Sanoma Media Netherlands BV and Others* ECLI:EU:C:2016:644, paras 47-55.

model with representations of various dataset elements which are not precise copies but lossy approximations. Although these copies are transient, they are essential for extracting information; without them, the training of the model would be impossible. As per the browsers capable of sifting and searching through these links to collect images, a process akin to browsing the web, their activity further stretches the definition of public accessibility.[157]

Recital 9 DSM Directive states that mandatory exception for temporary acts of reproduction, as outlined in the EU *acquis*, might apply to limit TDM. This in the case when the process creates temporary reproductions that are 'transient or incidental' and which form an integral part of a technological process; thus, enabling the lawful use of the work, without offering 'independent economic significance'.[158] These conditions are cumulative and subject to a strict application ensuring that the exception is interpreted narrowly to prevent misuse of copyrighted content. The case law of the CJEU acknowledges that 'transient or incidental' copies may persist only for as long as necessary to complete the technological process, after which they must be deleted.[159] According to the Court, reproduction is considered to have independent economic significance if it generates an additional economic benefit beyond that obtained from the lawful use of the protected material or if it leads to a modification of the material itself.[160] When these cumulative criteria are applied to TDM, it becomes evident that they are not easily fulfilled.[161]

In the case *Infopaq*,[162] the CJEU ruled that an electronic news-cutting service, which created temporary copies during a data capture process, met the criteria for temporary copies under Article 5(1) InfoSoc Directive. Similarly, *Meltwater*,[163] the CJEU determined that transient copies made by users while browsing the internet also fall within the scope of Article 5(1) InfoSoc Directive. While these cases do not directly address the training of models, they

[157] Guadamuz (n 151) 116.
[158] InfoSoc Directive, art 5(1).
[159] *Infopaq* (n 146) [33], [64].
[160] ibid [51], [53].
[161] Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, 'Text and Data Mining in the Proposed Copyright Reform: Making the EU ready for an age of Big Data?' (2019) 49(7) International Review of Intellectual Property and Competition Law 814, 823-824.
[162] *Infopaq* (n 146) [64].
[163] Case C-360/13 *Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd and Others* ECLI:EU:C:2014:1195, para 63.

suggest that the Court may adopt a broad interpretation of what constitutes a temporary copy with respect to model training.

3.6 When Is TDM Considered Unlawful?

The collection of protected works without permission could infringe on copyright, particularly if it involves exercising one of the author's exclusive rights, such as reproduction, distribution, adaptation, or public communication. Article 3, as well as Article 4 DSM Directive provide exceptions for potential infringements related to the right of reproduction, not the right of distribution or communication to the public, nor of the right of adaptation.[164] Where the TDM is allowed under the exceptions and limitations included in the EU actus, additional authorisation would not be required since the harvesting activity is considered to serve some public interest and guarantee some of the core values enshrined within the fundamental rights.[165] In terms of copyright, TDM could violate the right of reproduction of the database's author, particularly during the pre-processing stages, where the original selection or arrangement of the database is copied entirely.

It is strenuous to determine if the data originated from a specific database, or if the outcome is actually an adaptation of different images. This difficulty occurs because the entire process involves analysing, clustering, indexing, and aggregating similar items, followed by passing them through a noise filter. For related rights, the key criterion shifts from ‘originality’ to ‘recognizability,’ meaning that even small fragments of a larger work can attract their own related rights protection.[166] As a result, since TDM often involves reproducing protected material, it is likely to infringe the related rights of the original creator, especially if the process involves copying recognizable portions of the material without significant altering, adding or selective extraction that reduces recognizability.

There may also be situations where the TDM occurs without any acts of reproduction. In situations where reproductions do happen, they might qualify for the mandatory exception

[164] Margoni and Kretschmer (n 116) 695.
[165] DSM Directive, recital 1 and 3 and InfoSoc Directive, recital 3.
[166] Rosati (n 45) 206.

for temporary acts of reproduction, as specified in Article 5(1) InfoSoc Directive.[167] However, generally, such acts of reproduction require authorization from the original authors and failing to obtain this permission may result in a copyright infringement.

It is worth noting that converting materials into machine-readable formats, known as pre-processing, could potentially violate the right of reproduction, as it involves creating copies of protected works. Similarly, when this pre-processed material is uploaded to a platform, it may infringe the latter right, particularly if the TDM method involves software that crawls data directly from its original source. Not only this but depending on the type of mining software involved and the type of extraction, copyright violations may also arise.[168]

Although some AI system generators may claim they are trained on temporary copies, the line between temporary and permanent reproductions is often blurred. In many instances, copies that are created during TDM processes do not remain as merely 'transient or incidental' but are intended to be retained for extended periods. This retention is helpful for allowing these copies to be pre-processed, uploaded into various mediums, and mined for the development and training of creative AI systems. Not only this but, TDM techniques rarely involve an automatic or manual deletion process for the created copies. In effect, the duration of their retention is heavily dependent on those performing TDM and their subsequent users, such as AI developers, whose motives are often unclear and difficult to anticipate.[169]

[167] Bottis (n 100) 384.
[168] DSM Directive, art 4.
[169] Rosati (n 45) 203-204.

CHAPTER 4

ASSESSING THE EFFECTIVENESS OF ARTICLE 4 DSM DIRECTIVE IN RESOLVING THE COPYRIGHT CHALLENGES IN TDM

Article 4 DSM Directive has proven insufficient for the context of AI advancements, as the EU legislators failed to consider the need for data analysis specific to AI when the Directive was adopted. Namely, the broad definition of TDM included in the Directive, as well as the purpose behind Article 4 DSM Directive and the fact that the background of the provision included discussions on AI training, support the idea that making copies of data during AI training could be seen as a type of data mining.[170]

The mandatory exception within Article 4 DSM Directive presents an unresolved legal disparity. Given that the Article granted Member States a certain degree of discretion in their implementation,[171] this latitude posed a peril, sparking additional fragmentation across national legal orders. Consequently, this hindered the harmonisation of the internal market which the Directive was seeking to achieve. Not only this, but it allowed for a certain level of cherry-picking by the countries.[172] It is questionable if the legal scope of Article 4 DSM Directive proves its usefulness in practice, given some remaining inherent ambiguities in its requirements. The way Article 4 DSM Directive is framed today is rather restricted and may thus not unleash the full capacity of AI in Europe.[173]

4.1 Article 4(3) – The Achilles' Heel Within the DSM Directive

Besides all challenges stemming from Article 4 DSM Directive, the exception within 4(3) DSM Directive has potentially solidified its status as a primary obstacle to the advancement of AI systems in the EU.[174] The opt-out mechanism established by Article 4(3) DSM Directive allows the legal exception to be easily overridden by contractual agreements or technical measures; thus, subjecting it solely to private will. In fact, the language of the provision creates a potential Schrödinger's paradox, where acts like observing, testing, studying a function of a computer program, or even simply using a database, could be restricted if the exception within

[170] Margoni and Kretschmer (n 116) 686.
[171] DSM Directive, art 25.
[172] Lucie Guibault, 'Why Cherry-Picking Never Leads to Harmonisation: The Case of the Limitations on Copyright under Directive 2001/29/EC' (2010) 1(2) JIPITEC 55, 55-64.
[173] Ziaja (n 22) 454.
[174] Artha Dermawan, 'Text and data mining exceptions in the development of generative AI models: what the EU member states could learn from the Japanese "nonenjoyment" purposes?' (2023) 27(1) The Journal of World Intellectual Property 44, 53-54.

paragraph 3 is invoked.[175] Notably, Article 4 DSM Directive is the only provision that explicitly addresses the economic rights conferred by the Software Directive[176] by further extending the 'black box approach' set out therein. The fact that the Article allows for both commercial and non-commercial purposes is not entirely compelling, given the opt-out mechanism and retention period restrictions it prescribes. This limitation significantly curtails the practical utility of the exception, raising concerns about the extent to which software can be studied, tested, or observed, as well as the degree to which economic rights can be exercised in commercial settings, given the rightholder's power to impose limitations.[177]

The idea behind the introduction of Article 4(3) DSM Directive was to reinforce the position of creators, believing that such would strengthen the negotiation power of artists. Nonetheless, the reality proved to be different. Given the concentration of power in creative labour markets and the significant influence wielded by dominant corporations, artists are often pressured to accept contractual terms that require them to forfeit their 'training rights' in exchange for lower compensation.[178] The dynamics followed in the Directive could further entrench the power of large companies, ultimately reducing the control and compensation available to artists in the medium to long term.[179] Others also argue that the freedom entrusted upon authors to reserve, *inter alia*, their right of reproduction for harvesting activities would stifle the application of the exceptions set out in Article 4 DSM Directive by impeding its intended purpose.[180]

Yet, still, a lot of chanciness remains within Article 4(3) DSM Directive. The Directive does not give a clear answer as to whether creators can declare an opt-out *ex post*, after the work has been made available.[181] Specifically, it is ill-defined whether an artist can prohibit the use of their works for TDM, following publication, even if such was initially permitted. Not only this,

[175] Christensen (n 117) 31.
[176] Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs [2009] OJ L111/16 (Software Directive).
[177] Christensen (n 117) 31.
[178] Ziaja (n 22) 454.
[179] João Quintais, 'Generative AI, Copyright and the AI Act' (*Kluwer* Copyright Blog, 9 May 2023) <https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/ > (accessed 11 July 2024).
[180] Ziaja (n 22) 455.
[181] Alexander Peukert, 'Copyright in the Artificial Intelligence Act -A Primer' (2024) 73(6) GRUR International 497, 503.

but the Directive does not delineate if, in a situation where TDM opt-out is invoked, such would apply to all versions and formats of the work, or just to a few specific copies of the latter.[182] As the law stands, the EU legislators failed to also clarify in what machine-readable manner should the out-out be expressed so that it is appropriate under the law.[183] Currently, no universally accepted standards or protocols for the machine-readable expression of opting out exist. While several emerging approaches to this issue are being developed by various market participants, it remains largely unclear which, if any, will be adopted by leading AI model providers.[184]

Although Article 4(3) of the DSM Directive addresses certain aspects of copyright in the context of AI, there is no comprehensive EU Copyright Regulation specifically governing AI. The EU regulates AI-related copyright issues primarily through Directives, including the DSM Directive. While national copyright laws have been largely harmonised, the EU lacks a dedicated copyright Regulation for AI. The only exception is the AI Act[185], which, through Article 53 and Recitals 104-109, imposes an obligation on general-purpose AI providers to comply with existing copyright legislation. Even then, the primary source of copyright regulation on AI applications comes from pre-existing secondary EU copyright law.

The CJEU case law stipulates that even a clear, precise, and unconditional provision of a Directive that confers rights or imposes obligations on individuals cannot apply in legal disputes between private parties.[186] This implies that AI developers have to swiftly navigate a complex patchwork of 27 national copyright legislations which contain different rights, with varying limitations and exceptions. For instance, in certain jurisdictions, it is compulsory for the TDM opt-out to be provided in a machine-readable format, while in others it either

[182] Paul Keller and Zuzanna Warso, 'Defining best practices for opting out of ML training' (*Open Future*, 28 September 2023) 14 <https://openfuture.eu/publication/defining-best-practices-for-opting-out-of-ml-training/> accessed 14 July 2024.
[183] DSMD, recital 18.
[184] Paul Keller 'Generative AI and Copyright: Convergence of Opt-outs?' (*Kluwer Copyright Blog*, 23 November 2023) <https://copyrightblog.kluweriplaw.com/2023/11/23/generative-ai-and-copyright-convergence-of-opt-outs/> accessed 15 July 2024.
[185] 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) [2024] OJ L2024/1689.
[186] Case C-351/12 *OSA v Léčebné lázně Mariánské Lázně a.s*. ECLI:EU:C:2014:110, para 42.

remains optional or is simply not required.[187] Further to that, Article 4 DSM Directive does not apply to model training acts which were executed before 7 June 2021. This could be crucial for certain AI models, as they may not be able to take advantage of the TDM exceptions outlined in Article 4 DSM Directive.[188]

4.2 Territorial Limitations and Disparities in Article 4 DSM Directive

Another issue arising out of Article 4(3) DSM Directive is its territorial scope. If the model training does not fall within the scope of Article 4 DSM Directive, there is no obligation for AI developers to recognise and comply with the reservations included in Article 4(3) DSM Directive. As per the principle of territoriality, the Directive does not apply outside of the Union. In the same fashion, EU copyright law will only apply to acts that occur in the territory of the EU.[189] This means that rightsholders are allowed to control access to their art only if training takes place in the EU, or if their artwork is hosted and restricted within the Union, even if the model training is executed outside the EU.[190]

Speaking of other jurisdictions, the governance of TDM in Europe seems to be much more restrictive, in comparison to other jurisdictions, which rely on fair use clauses or opening models permitting a larger number of TDM activities. In this manner, AI developers, as well as other innovators may find themselves in a disadvantageous situation, in contrast to their counterparts.[191]

In any event, the opt-out mechanism within Article 4(3) DSM Directive crippled the assumption that text and data harvesting should be more widely accessible. From a market standpoint, this could hinder innovation opportunities for a wide range of market participants,

---

[187] Countries requiring opt-out to be provided in a machine-readable format include Belgium, Germany, Irelands, Hungary, Bulgaria, Czechia, Latvia, Lithuania, Portugal and Slovenia.
[188] DSM Directive, art 26(2).
[189] Case C-523/10 *Wintersteiger AG v Products 4U Sondermaschinenbau GmbH* ECLI:EU:C:2012:220, para 25.
[190] Peukert (n 181) 506.
[191] Tatiana Synodinou, 'Comment of the European Copyright Society addressing selected aspects of the implementation of articles 3 to 7 of Directive (EU) 2019/790 on Copyright in the Digital Single Market' (*Kluwer Copyright Blog*, 27 July 2022) <https://copyrightblog.kluweriplaw.com/2022/07/27/comment-of-the-european-copyright-society-addressing-selected-aspects-of-the-implementation-of-articles-3-to-7-of-directive-eu-2019-790-on-copyright-in-the-digital-single-market/> accessed 9 August 2024.

varying from large corporations to startups. In light of this, the DSM Directive should have considered the possibility of extending the non-overridable limitations to certain categories of beneficiaries.

4.3 Retention Limits and Challenges of Temporary Reproductions in TDM

Nevertheless, another issue emerges from Article 4(2) DSM Directive which partially recalibrates the period for retaining reproductions and extraction made during TDM. According to Article 4(2) DSM Directive, such reproductions can be retained 'for as long as necessary'. In practice, this wording implies that only temporary reproductions are permitted, a concept closely aligned with the mandatory exception found in Article 5(1) of the InfoSoc Directive.[192] However, as previously discussed, the temporary reproduction exception is insufficient to accommodate unlicensed TDM for AI-driven creative purposes. This limitation creates a scenario where neither public, nor private AI developers would be willing to make the substantial investments required for data mining if the resulting copies could only be retained temporarily.[193]

4.4 Licensing Constraints and Their Unintended Economic Implications

In order to comply with the three-step test, as expressly provided for in Article 7(2) DSM Directive, the entirety of a rightsholder's work cannot be used, even in situations where they have not opted out.[194] Nevertheless, where authors do exercise their right to opt out, AI system developers will be required to obtain licenses, particularly when the protected data in question is essential for training the respective AI systems. Imposing such a requirement for authorisation in the context of training AI, introduces an additional financial burden for system developers, in the form of licensing fees and transaction costs.[195] In the long term, this could

[192] League of European Research Universities, 'Europe needs a broad & mandatory TDM exception' (*LERU*, 14 November 2018) <https://www.leru.org/news/europe-needs-a-broad-mandatory-tdm-exception> accessed 14 July 2024.
[193] Christensen (n 118) 31.
[194] Eleonora Rosati, 'No step-free copyright exceptions: the role of the three-step in defining permitted uses of protected content (including TDM for AI-training purposes)' (2023) 46(5) Stockholm Faculty of Law Research Paper Series 1, 20.
[195] Ziaja (n 22) 455.

diminish the economic viability of licensing copyright-protected content for AI training purposes; thus, weaken the EU's competitive position in relation to other jurisdictions with more robust legal frameworks governing AI protection. This is likely to have a great impact on the investment and development of AI, as opposed to the objectives of the Directive. Paradoxically, it may also facilitate the development of biased AI systems. This unintended consequence may arise from the creation of perverse incentives, influenced by the accessibility and cost of training data. In an effort to avoid licensing fees, it may become economically advantageous for developers to train AI systems on outdated and biased datasets, or to rely on pre-trained algorithms developed using unverifiable data.[196] Such practices could ultimately compromise the quality, fairness, and reliability of AI systems within the EU; thereby, undermine the Union's aspirations for technological advancement and ethical innovation.

4.5 Lawful Access, but Is It for All?

Article 4 DSM Directive does not define the requirement for lawful access, but instead refers to the Guidance found in Recital 14 DSM Directive. This approach subjects the TDM exception to private ordering, effectively making the exercise of the TDM limitation dependent on the decisions of creators. Hence, start-ups and small and medium-sized enterprises may face significant challenges in negotiating reasonable TDM licenses with owners of large datasets. Essentially, this framework will increase the danger of their exclusion from TDM activities and AI development, thus severely limiting any potential for innovation.[197]

4.6 Possible Solutions to Remedy the Challenges Posed by Article 4 DSM Directive?

The overall objective of the Directive - to balance the interests of rightsholders while avoiding negative impacts on AI innovation in Europe - remains hard to strike. It is apparent that Article 4 DSM Directive rather than acting as an enabler of innovation, has emerged as an obstacle.[198]

[196] Margoni and Kretschmer (n 116) 700.
[197] Pamela Samuelson, The EU's Controversial Digital Single Market Directive (2018) 61(11) Communications of the ACM 20, 23.
[198] Rosati (n 45) 217.

Although the DSM Directive adopts a more digitally attuned approach than its legislative predecessors, it still falls short in meeting the demands of the Fourth Industrial Revolution, particularly interlinked with AI. This conclusion is supported by the apparent prioritization of the copyright and related rights of rightsholders who control content. Such prioritization is *prima facie* evident in the requirement for lawful access and the opt-out mechanism outlined in Article 4 DSM Directive. Although technological revolutions may come with a significant amount of legal strife, the gaps found within Chapter 4 DSM Directive could be alleviated in several ways. Embarking on initiatives to address the lacunae in the law will help create a more adaptive legal framework capable of responding to the complexities introduced by the ever-evolving AI systems.

A more effective solution, as opposed to Article 4(3) DSM Directive, could have been the introduction of remuneration rights for TDM that is conducted for commercial purposes. This would not only place financial responsibility on users without resorting to a blanket opt-out option but will better align the Directive with the realities of AI development. The mandatory exception within Article 4(3) DSM Directive places small and medium-sized enterprises at a disadvantageous position; unlike, major corporations such as OpenAI, Stability AI and Midjourney. Namely, these corporations, which have already secured access to vast datasets, enjoy a competitive advantage, as they are not dependent on authors' authorisations. This allows them to not only continue utilising the data they have trained their AI systems with but also leverage their extensive datasets to develop more advanced systems. By doing so, they can sidestep the costs associated with creating proprietary datasets from scratch - a prohibitive expense for new market entrants attempting to compete on an equal footing with established industry giants.

Another alternative remedy for the challenges posed by Article 4 DSM Directive would be the establishment of data pools which are centralised repositories where diverse datasets and information can be stored, maintained, and exchanged among various market participants. The creation of a single set of works and other subject matter will significantly simplify the process for AI developers to create images, obliterating the need to access individual works. Not only this but such a solution could streamline access to relevant data for small businesses that lack the resources of larger corporations. Inevitably, this may also result in the

development of more diverse and ethically robust AI systems. If authors were allowed to license their works within these data pools, they would have the incentive to create high-quality datasets; thereby, generating new revenue streams while simultaneously supporting the development of AI systems.

Additionally, the development of data pools could address the legal and logistical complexities tied to data ownership, licensing, and copyright. By centralizing datasets and implementing standardized licensing practices, data pools would reduce the menace of fragmentation and legal disputes, ensuring that developers and rightsholders alike have more cohesive guidelines in the use and distribution of data. That being said, positioning data pools as a solution that balances the protection of intellectual property rights with the facilitation of technological innovation, would be a way to overcome many of the barriers currently hindering the AI development in Europe

Another possible solution targeting the fragmentation across Member States in implementing Article 4 DSM Directive, would be the implementation of additional guidelines to limit the degree of discretion afforded to States, particularly as regards the opt-out mechanism. Such guidelines could also provide clearer rules regarding retention periods and data extraction practices during TDM activities and play an important role in promoting ethical data usage in AI training, by reducing the risks associated with outdated or biased datasets.

After exhausting all alter-teleological limits of the existing rules, a more thorny and arduous task that the EU could potentially involve itself in would be the adoption of a new Regulation aimed at providing a resilient solution to the fragmentation caused by the current implementation of Article 4 DSM Directive. Unlike directives, which give Member States discretion for the transposition of provisions into national law, regulations establish a uniform and directly applicable legal framework across the EU. This uniformity would ensure that the TDM exceptions and opt-out mechanisms are applied consistently in all Member States, leaving no ambivalence and disparity in both interpretation and enforcement. By adopting a new Regulation, national laws would not only be harmonised, but a greater degree of legal certainty would be provided for AI developers, preventing rightsholders from exploiting national discrepancies to impose restrictive conditions. In doing so the initial objectives

sought under the DSM Directive will be more effectively realised. This approach will transform TDM from a mere source of friction into a collaborative opportunity, aligning the interests of rightsholders with the broader goal of fostering technological progress.

4.7 Concluding remarks

TDM stands as a crucial pillar for the development of AI, particularly in today's data-driven socio-economic context. With the proliferation of technology, stakeholders recognised the transformative potential of TDM across different heterogeneous fields.[199] The creative industries, notably those involved in the generation of AI art, have not been immune to this revolution. In particular, companies engaged in the development of AI art have harnessed extensive datasets to train their models. However, concerns have transpired around the use of copyrighted works as inputs for these systems.

While the DSM Directive has successfully introduced a more modern approach to copyright in the digital age, it fell short in accommodating the full demands and possibilities of AI, particularly with regard to TDM. Article 4 DSM Directive alongside the opt-put mechanism set out thein, which was intended to bridge this regulatory gap, has instead exacerbated the levels of fragmentation. Namely, this divergence has added an additional layer of uncertainty regarding the rights of authors and the obligations of AI developers.

Unless resolved, the ambiguities about TDM threaten to undermine the EU's position as a hub for research, innovation and creativity on the global stage. While the European Union has already laid a good foundation, much remains to be done to ensure legal certainty and robust support for the future of AI-driven innovation. Having the legal loopholes now identified, only time will tell what steps the European legislators will take to pave the way for a comprehensive TDM framework that balances innovation with the protection of copyright. The right solution may not be yet visible but with decisive actions, the EU can illuminate its path forward, driving AI development with both clarity and prudence.

[199] Rosati (45) 23.

BIBLIOGRAPHY

Primary Sources:

Cases Court of Justice of the European Union

Case C-393/09 Bezpečnostní softwarová asociace -Svaz softwarové ochrany v Monisterstvo kultury ECLI:EU:C:2010:816 [2010] ECR I-13971

Case C-683/17 Cofemel -Sociedade de Vestuário SA v G-Star Raw CV ECLI:EU:C:2019:72

Case C-145/10 Eva-Maria Painer v Standard Verlags GmbH and Others ECLI:EU:C:2011:798 [2011] ECR I-12533

Case C-604/10 Football Dataco Ltd and Others v Yahoo! UK Ltd and Others ECLI:EU:C:2012:33

Case C-469/17 Funke Medien NRW GmbH v Bundesrepublik Deutschland ECLI:EU:C:2019:623

Case C-5/08 Infopaq International A/A v Danske Dagblades Forening ECLI:EU:C:2009:465 [2009] ECR I-6569

Case C-310/17 Levola Hengelo v Smilde Foods BV ECLI:EU:C:2018:899

Case C-360/13 Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd and Others ECLI:EU:C:2014:1195

Case C-523/10 Wintersteiger AG v Products 4U Sondermaschinenbau GmbH ECLI:EU:C:2012:220

Cases United States of America

Andersen v Stability AI Ltd (ND Cal, 3:23-cv-00201, 2023)

Getty Images (US), Inc. v. Stability AI, Inc. 1:23-cv-00135 (D. Del. 2023)
Motion to Dismiss filed by Stability AI Ltd, at 1, Andersen (No 23-cv-00201-WHO, ND Cal, 18 April 2023)

Cases United Kingdom

Getty Images (US) Inc and others v Stability AI Ltd [2023] EWHC 000007 (Ch)

EU legislation

Charter of Fundamental Rights of the European Union [2012] OJ C 326/391 (CFR)

Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market [2019] OJ L 130/92 (DSM Directive)

Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs [2009] OJ L111/16 (Software Directive)

Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market, COM(2016)593, 14 September 2016

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) [2024] OJ L2024/1689

Council of Europe legislation

Convention for the Protection of Human Rights and Fundamental Freedoms as Amended by Protocols No.11 and No.14 (adopted on 4 November 1950, entered into force 3 September 1953) CETS No. 005 (ECHR)

International Treaties

Agreement on Trade-Related Aspects of Intellectual Property Rights (adopted 15 April 1994, entered into force 1 January 1995) 1869 UNTS 299 (TRIPS)

Berne Convention for the Protection of Literary and Artistic Works (adopted 9 September 1886, entered into force 5 December 1887) 1161 UNTS 3 (Berne Convention)

WIPO Copyright Treaty (adopted 20 December 1996, entered into force 6 March 2002) 2186 UNTS 121 (WCT)

Secondary Sources:

Books

Elmasri R and Navathe S, *Fundamentals of Database Systems* (7th edition, Pearson 2016)

Hutchinson T, *Researching and Writing in Law* (3rd Edition, Thomson Reuters 2010)

McConville M and Hing Chui W (eds), *Research Methods for Law* (2nd Edition, Edinburgh University Press 2007)

M Patel J, *Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale* (Apress 2020)

Sheikh H, Prins C and Schrijvers E, *Mission AI* (Springer 2023)

Articles:

Ahmad Ali Elfa M and Eshaq Tawfilis Dawood M, 'Using Artificial Intelligence for enhacing Human Creativity' (2023) 2(2) Journal of Art, Design & Music 106

Alwidian S, Bani-Salameh H and Alslaity A, 'Text data mining: a proposed framework and future perspectives' (2015) 18(2) International Journal of Business Information Systems 127

Ang D, 'The Web Scraper's World of Copyright Exceptions and Contractual Overrides' (2022) 13 Singapore Law Review 1

Ang T and others, 'The rise of artificial intelligence: addressing the impact of large language models such as ChatGPT on scientific publications' (2023) 64(4) Singapore Medical Journal 219

A. Hearst M, 'Untangling Text and Data Mining' (1999) Proceedings of the 37th Annual meeting of the Association for Computational Linguistics 3

Asim khan M and Ahmad S, 'An Emergence of AI in Data Mining and KDD: ANN its Strength & Weakness' (2020) 9(1) International Journal of Recent Technology and Engineering 380

Bell C, 'A Note on Participant Observation' (1969) 3(2) Sociology 417

Bottis M and others, 'Text and Data Mining in Directive 2019/790/EU Enhancing Web-Harvesting and Web-Archiving in Libraries and Archives' (2019) 9(3) Open Journal of Philosophy 369

Casini M and others 'Current trends on the use of deep learning methods for image analysis in energy applications' (2024) 15 Energy and AI 1

Chatterjee A, 'Conceptual Analysis' (2022) 13 Frontiers 1

Christensen K, 'A European solution for Text and Data Mining in the development of creative Artificial Intelligence With a specific focus on articles 3 and 4 of the Digital Single Market Directive' (2021) 4(2) Stockholm Intellectual Property Law Review 18

Couldry N and Mejias U, ‘Data Colonialism: Rethinking Big Data’s Relation to the Contemporary Subject’ (2019) 20(4) Television & New Media 336

Dermawan A, ‘Text and data mining exceptions in the development of generative AI models: what the EU member states could learn from the Japanese “nonenjoyment” purposes?’ (2023) 27(1) The Journal of World Intellectual Property 44

E Hinton G, Osindero S and Teh Y, ‘A fast learning algorithm deep belief nets’ (2006) 18(7) Neural Computation 1527

Ferri F, ‘The dark side(s) of the EU Directive on copyright and related rights in the Digital Single Market’ (2020) 7(1-4) China-EU Law Journal 21

Filippov S, ‘Mapping Text and Data Mining in Academic and Research Communities in Europe’ (2014) 16 The Lisbon Council 1

Geiger C, Frosio G and Bulayenko O, ‘Text and Data Mining in the Proposed Copyright Reform: Making the EU ready for an age of Big Data?’ (2019) 49(7) International Review of Intellectual Property and Competition Law 814

Geiger C, ‘The Missing Goal-Scorers in the Artificial Intelligence Team: Of Big Data, the Fundamental Right to Research and the failed Text and Data Mining Limitations in the CSDM Directive’ (2021) 66 PIJIP/TLS Research Paper Series 383

Guibault L, ‘Why Cherry-Picking Never Leads to Harmonisation: The Case of the Limitations on Copyright under Directive 2001/29/EC’ (2010) 1(2) JIPITEC 55

Knoedler L, ‘Application possibilities of artificial intelligence in facial vascularized composite allotransplantation-a narrative review’ (2023) 10 Frontiers 1

Kugel P ‘Artificial Intelligence and Visual Art’ (1981) 14(2) Leonardo 137

Magdy Hanna D, ‘The Use of Artificial Intelligence Art Generator "Midjourney" in Artistic and Advertising Creativity’ (2023) 4(2) Journal of Design Science and Applied Arts 42

Maria Ziaja G, ‘The text and data mining opt-out in Article 4(3) CDSMD: Adequate veto right for rightholders or a suffocating blanket for European artificial intelligence innovations?’ (2024) 19(5) Journal of Intellectual Property Law 453

Mariscal G, Marbán O and Fernández C, ‘A Survey of Data Mining and Knowledge Discovery process Models and Methodologies’ (2010) 25(2) KER 137

Margoni T and Kretschmer M, ‘A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology’ (2020) 71(8) GRUR International 685

M. Townsend D and A.Hunt R, ‘Entrepreneurial action, creativity, & judgment in the age of artificial intelligence’ (2019) 11 Journal of Business Venturing Insights 1

Patil N, Patel S and Lawsnd S, ‘Research on Artificial Intelligence and its Applications’ (2023) 44(8) Journal of Advanced Zoology 229

Peres R and others, ‘On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice’ (2023) 40(2) International Journal of Research in Marketing 269

Peukert A, ‘Copyright in the Artificial Intelligence Act -A Primer’ (2024) 73(6) GRUR International 497

Pinhero A and others, ‘Malware detection employed by visualization and deep neural network’ (2021) 105 Computers & Security 1

Prior F and others, ‘Open access image repositories: high-quality data to enable machine learning research’ (2020) 75(1) Clinical Radiology 7

Quang J, ‘Does Training AI Violate Copyright Law’ (2021) 36(4) Berkeley Technology Law Journal 1407

R. Sarvamangala D and V. Kulkarni R, ‘Convolutional neural networks in medical image understanding: a survey’ (2022) 15(1) Evol Intell=

Rosati E, ‘Copyright as an obstacle or an enabler? A European perspective on text and data mining and its role in the development of AI creativity’ (2019) 27(2) Asia Pacific Law Review 198

Rosati E, ‘No step-free copyright exceptions: the role of the three-step in defining permitted uses of protected content (including TDM for AI-training purposes)’ (2023) 46(5) Stockholm Faculty of Law Research Paper Series 1

Samuelson P, The EU’s Controversial Digital Single Market Directive (2018) 61(11) Communications of the ACM 20

Schuhmann C and others, ‘LAION-5B: An open large-scale dataset for training next generation image-text models’ (2022) NeurIPS 2022 Datasets and Benchmarks 1

Suganthi R and others, ‘Utilizing Web Scraping for Big Data: An Exploratory Analysis’ (2023) 44(6) Journal of Propulsion Technology 6150

Tromble R, ‘Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age’ (2021) 7(1) Sage Journals 1

Tyagi K, ‘Copyright, text & data mining and the innovation dimension of generative AI’ (2024) 19(7) Journal of Intellectual Law & Practice 557

Velez-Estevez A and others, ‘New trends in bibliometric APIs: A comparative analysis’ (2023) 60(40) Information Processing and Management 1

W. Opderbeck D, ‘Copyright in AI Training Data: A Human-Centered Approach’ (2024) 76(4) Oklahoma Law Review 1

Xu Y, ‘Artificial intelligence: A powerful paradigm for scientific research’ (2021) 2(4) Cell Press 1

Yamashita R and others, ‘Convolutional neural networks: an overview and application in radiology’ (2018) 9 Insights Imaging 611

Zhao X, ‘A review of convolutional neural networks in computer vision’ (2024) 57 Artificial Intelligence Review 1

Zwolenski M and Weatherill L, ‘The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things’ (2020) 2(3) Australian Journal of Telecommunications and the Digital Economy 1

Websites:

Antonio Lanz J, ‘Greg Rutkowski Was Removed From Stable Diffusion, But AI Artists Brought Him Back’ (*Decrypt*, 30 July 2023) <https://decrypt.co/150575/greg-rutkowski-removed-from-stable-diffusion-but-brought-back-by-ai-artists> accessed 17 August 2024

Baio A, ‘Exploring 12 Million of the 2.3 Billion Images.’ (*WAXY*, 30 August 2022) <https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/> accessed 7 July 2024

Blank S, 'Artificial Intelligence and Machine Learning Explained' (*Gordian Knot Center for National Security Innovation*, 17 May 2022) 2 <https://steveblank.com/2022/05/17/artificial-intelligence-and-machine-learning-explained/> accessed 17 June 2024

Foley J, 'DALL-E 2 access is now open to everyone, and that's kind of terrifying' (*Creative Bloq*, 1 October 2022) <https://www.creativebloq.com/news/dalle2-access-open-to-all> accessed 18 July 2024

'Getty Images Introduces Updated AI Model with Increased Speed, Quality, and Accuracy' (*Getty Images*, 29 July 2024) <https://newsroom.gettyimages.com/en/getty-images/getty-images-introduces-updated-ai-model-with-increased-speed-quality-and-accuracy> accessed 18 July 2024

Heikkilä M, 'This artist is dominating AI-generated art. And he's not happy about it. Greg Rutkowski is a more popular prompt than Picasso' (*MIT Technology Review,* 16 September 2022) <https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/> accessed 5 July 2024.

Keller P and Warso Z, 'Defining best practices for opting out of ML training' (*Open Future*, 28 September 2023) 14 <https://openfuture.eu/publication/defining-best-practices-for-opting-out-of-ml-training/> accessed 14 July 2024

League of European Research Universities, 'Text and Data Mining: The need for change in Europe' (*Liber Europe*) <https://libereurope.eu/topic/text-data-mining/> accessed 1 June 2014

League of European Research Universities, 'Europe needs a broad & mandatory TDM exception' (*LERU*, 14 November 2018) <https://www.leru.org/news/europe-needs-a-broad-mandatory-tdm-exception> accessed 14 July 2024

Ma Y, 'The Ultimate Stable Diffusion LoRA Guide (Downloading, Usage, Training)' (*Aituts*, 24 April 2024) <https://aituts.com/stable-diffusion-lora/> accessed 1 August 2024

Plunkett L, 'AI Creating 'Art' Is An Ethical And Copyright Nightmare' (*Kotaku*) <https://kotaku.com/ai-art-dall-e-midjourney-stable-diffusion-copyright-1849388060> accessed 8 August 2024

Porterfield C, 'Judge dismisses most of artists' copyright lawsuit against AI image generators' (*The Art Newspaper*, 31 October 2023) <https://www.theartnewspaper.com/2023/10/31/california-judge-dismisses-most-of-artists-ai-copyright-lawsuit> accessed 4 August 2024

Shashkevich A, 'Stanford researcher examines earliest concepts of artificial intelligence, robots in ancient myths' (*Stanford Report*, 28 February 2019) <https://news.stanford.edu/stories/2019/02/ancient-myths-reveal-early-fantasies-artificial-life> accessed 1 August 2024

Tobin S, 'Getty asks London court to stop UK sales of Stability AI system' (*Reuters*, 1 June 2023) <https:// www.reuters.com/technology/getty-asks-london-court-stop-uk-sales-stability-ai-system-2023-06-01/> (accessed 3 August 2024)

Vincent J, 'Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its Content' (*The Verge*, 17 January 2023) <https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit> accessed 8 August 2024

Blogs:

Keller P 'Generative AI and Copyright: Convergence of Opt-outs?' (*Kluwer Copyright Blog*, 23 November 2023) <https://copyrightblog.kluweriplaw.com/2023/11/23/generative-ai-and-copyright-convergence-of-opt-outs/> accessed 15 July 2024

Quintais J, 'Generative AI, Copyright and the AI Act' (*Kluwer* Copyright Blog, 9 May 2023) <https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/ > (accessed 11 July 2024)

Synodinou T, 'Comment of the European Copyright Society addressing selected aspects of the implementation of articles 3 to 7 of Directive (EU) 2019/790 on Copyright in the Digital Single Market' (*Kluwer Copyright Blog*, 27 July 2022) <https://copyrightblog.kluweriplaw.com/2022/07/27/comment-of-the-european-copyright-society-addressing-selected-aspects-of-the-implementation-of-articles-3-to-7-of-directive-eu-2019-790-on-copyright-in-the-digital-single-market/> accessed 9 August 2024

Reports:

Hargreaves I and others, 'Standardisation in the Area of Innovation and Technological Development, notably in the Field of Text and Data Mining' (*European Commission*, April 2014) 63 <https://www.researchgate.net/publication/265078277_Final_Report_from_the_Expert_Group_on_Standardisation_in_the_area_of_Innovation_and_technological_development_Notably_in_the_field_of_Text_and_Data_Mining> accessed 1 June 2024

Official documents:

European Alliance for Research Excellence, ‘Open letter to European Commission, Maximizing the benefits of Artificial Intelligence through future-proof rules on Text and Data Mining’ (2018)

European Commission, Commission Staff Working Document - Impact Assessment on the modernisation of EU copyright rules, SWD(2016)301 final, 14 September 2016

European Commission, Commission Staff Working Document, SWD(2022)366, 29 November

European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - A European strategy for data, COM(2020)66, 19 February 2020

European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Shaping Europe’s digital future, COM(2020)67, 19 February 2020

European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Towards a modern, more European copyright framework, COM(2015)626, 9 December 2015

European Commission, Communication on artificial intelligence for Europe, COM(2018)237, 25 April 2018

European Commission, Legal and regulatory implications of Artificial Intelligence. The case of autonomous vehicles, m-health and data mining, JRC116235, 2019

European Commission, ‘Making the most of the EU’s innovative potential, An intellectual property action plan to support the EU’s recovery and resilience’ COM(2020)760, 25 November 2020

European Commission, Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining: Report from the Expert Group, April 2014

European Commission, ‘White Paper on Artificial Intelligence: a European Approach to Excellence and Trust’, Brussels (19 February 2020) COM(2020) 65 final

Rosati E, ‘The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspects PE, 604.942

Conferences:
Luo Y(ed), 'Cooperative Design, Visualization, and Engineering' (20th International Conference on Cooperative Design, Visualization and Engineering, CDVE 2023, Mallorca, 1-4 October 2023 39 <https://link.springer.com/book/10.1007/978-3-031-43815-8> accessed 7 August 2024

Study:
Boucher P, 'Artificial intelligence:How does it work, why does it matter, and what can we do about it?' PE 641.547 - June 2020

Triaille J, de Meeûs d'Argenteuil J and de Francquen A, 'Study on the Legal Framework of Text and Data Mining (TDM)' European Union, March 2014

Thesis:
Burger L, 'LAION: Image Data, AI, and Dispossession' (Master's Thesis, Utrecht University 2023) 21