



**Stanford – Vienna
Transatlantic Technology Law Forum**

A joint initiative of
Stanford Law School and the University of Vienna School of Law



European Union Law Working Papers

No. 150

**Training AI Models on Copyrighted Works
under EU Law**

Violietta Hondza

2026

European Union Law Working Papers

Editors: Siegfried Fina and Roland Vogl

About the European Union Law Working Papers

The European Union Law Working Paper Series presents research on the law and policy of the European Union. The objective of the European Union Law Working Paper Series is to share “works in progress”. The authors of the papers are solely responsible for the content of their contributions and may use the citation standards of their home country. The working papers can be found at <http://ttlf.stanford.edu>.

The European Union Law Working Paper Series is a joint initiative of Stanford Law School and the University of Vienna School of Law’s LLM Program in European and International Business Law.

If you should have any questions regarding the European Union Law Working Paper Series, please contact Professor Dr. Siegfried Fina, Jean Monnet Professor of European Union Law, or Dr. Roland Vogl, Executive Director of the Stanford Program in Law, Science and Technology, at:

Stanford-Vienna Transatlantic Technology Law Forum
<http://ttlf.stanford.edu>

Stanford Law School
Crown Quadrangle
559 Nathan Abbott Way
Stanford, CA 94305-8610

University of Vienna School of Law
Department of Business Law
Schottenbastei 10-16
1010 Vienna, Austria

About the Author

Viолетта Hondza is a law student in the Bachelor's Program International Legal Studies at the University of Vienna in Austria. Her academic work focuses on intellectual property law, in particular EU copyright law and the legal issues raised by generative artificial intelligence. She has also studied finance and management at Kyiv National Economic University, Ukraine, which provide her with an interdisciplinary view of the commercial and regulatory aspects of technology. This paper is based on her bachelor thesis, which she wrote under the supervision of Professor Dr. Siegfried Fina. She intends to pursue a career in intellectual property law.

General Note about the Content

The opinions expressed in this student paper are those of the author and not necessarily those of the Transatlantic Technology Law Forum, or any of TTLF's partner institutions, or the other sponsors of this research project.

Suggested Citation

This European Union Law Working Paper should be cited as:
Viолетта Hondza, Training AI Models on Copyrighted Works under EU Law, Stanford-Vienna European Union Law Working Paper No. 150, <http://tlf.stanford.edu>.

Copyright

© 2026 Viолетта Hondza

Abstract

Training large language models and other generative AI systems requires datasets comprising hundreds of billions of documents drawn from the internet. A substantial proportion of these documents are copyrighted subject matter, raising the question of how this practice aligns with the rights of authors and other rightholders under European Union law. This thesis addresses this issue by first asking whether AI training constitutes copyright-relevant reproduction under Article 2 of the EU InfoSoc Directive, second examining to what extent the text and data mining exceptions of the EU DSM Directive allow for such use, and third evaluating how effectively the rightholder opt-out mechanism under Article 4(3) of that Directive balances the interests of AI developers and rightholders. Through doctrinal legal analysis based on the InfoSoc and DSM Directives and CJEU case law, the thesis concludes that AI training involves the right of reproduction at various stages of the training process. Without the temporary reproduction exception, a license or statutory exception is necessary. Article 3 of the DSM Directive excludes commercial AI developers, while Article 4 may only cover commercial training where lawful access exists and no valid opt-out has been exercised. This leaves data from unauthorized sources outside both provisions. Additionally, the opt-out mechanism is structurally inadequate due to the absence of technical standardization, asymmetric allocation of burdens, inability to address historical training data, and lack of dedicated enforcement. Based on these findings, the thesis concludes that EU copyright law provides a partial but inadequate framework for AI training. Drawing on recent litigation, licensing practices, and regulatory developments, the thesis proposes a set of targeted reforms, including a binding technical standard for machine-readable opt-outs, mandatory compensation for commercial text and data mining, a collective licensing scheme based on Article 12 of the DSM Directive, a harmonization clause linking the DSM Directive to the transparency obligations of the AI Act, and a regular review mechanism. The thesis argues that copyright law and AI development are not inherently irreconcilable and that the European Union has the institutional capacity to implement such a framework; the remaining question is one of political will.

Table of Contents

1. Introduction.....	7
1.1. Problem Statement and Relevance.....	7
1.2. Research Questions.....	9
1.3. Scope and Delimitations	10
1.4. Methodology	11
1.5. Structure of the Thesis	12
2. Technological Background of AI Training and Its Legal Relevance	14
2.1. Machine Learning — Basic Concepts	14
2.2. Large Language Models — Architecture and Training Process.....	16
2.3. The Memorisation Phenomenon	20
2.4. AI Training Distinguished from Other Data Processing.....	22
2.5. Legal Relevance of Technical Architecture.....	24
3. The Reproduction Right under EU Law and Its Application to AI Training	26
3.1. The Reproduction Right in the InfoSoc Directive	26
3.1.1. Statutory Analysis of Article 2	26
3.1.2. Relationship to Other Rights.....	29
3.2. The EU-Autonomous Concept of Reproduction and the Originality Threshold	30
3.2.1. EU-Autonomous Interpretation	30
3.2.2. Infopaq International — The Foundational Case.....	31
3.2.3. Infopaq II	34
3.2.4. Development of the Originality Standard in Subsequent Case Law.....	35
3.2.5. The Operative Standard for AI Training.....	38
3.3. Application of the Findings to the AI Training Pipeline	40
3.3.1. Web Scraping and Dataset Compilation.....	41
3.3.2. Dataset Storage and Preprocessing	42
3.3.3. The Training Process	45

3.3.4.	The Memorization Problem of Model Weights	48
3.4.	The Temporary Copying Exception — Article 5(1) InfoSoc Directive	51
3.4.1.	Structure and Purpose of Article 5(1)	51
3.4.2.	The Five Conditions — Cumulative Requirement	53
3.4.3.	Conclusion on Article 5(1) of the InfoSoc Directive	58
3.5.	Answer to Research Question 1	60
4.	Text and Data Mining Exceptions under the DSM Directive	62
4.1.	Introduction to TDM Exceptions	62
4.1.1.	Legislative History	62
4.1.2.	Policy Rationale	64
4.1.3.	Relationship to Existing InfoSoc Exceptions	66
4.1.4.	Definition of TDM — Article 2(2) DSM Directive	67
4.2.	Article 3 DSM — Exception for Research Organisations	69
4.2.1.	Personal Scope — Research Organisations	70
4.2.2.	Material Scope — Scientific Research Purpose	74
4.2.3.	Lawful Access Requirement	76
4.2.4.	Mandatory Nature and Contractual Override Prohibition	78
4.3.	Article 4 DSM — General TDM Exception	81
4.3.1.	Personal Scope	81
4.3.2.	Material Scope	82
4.3.3.	Lawful Access Requirement	85
4.3.4.	The Opt-Out Mechanism	87
4.3.5.	Temporal Problem — Historical Training Datasets	92
4.4.	The Pirated Database Problem	97
4.4.1.	Lawful Access as Categorical Exclusion	97
4.4.2.	Known Pirated Datasets in AI Training	99
4.4.3.	Legal Consequences	102
4.5.	Common Provisions of Articles 3 and 4	104

4.5.1.	Retention and Storage of Copies.....	105
4.5.2.	Interaction with the Database Right.....	108
4.5.3.	Article 7 — Contractual Override Prohibition.....	110
4.6.	Are TDM Exceptions Sufficient for AI?	112
4.7.	Answer to Research Question 2	121
5.	The Opt-Out Mechanism and Licensing Practices	124
5.1.	The Opt-Out as Policy Instrument	124
5.2.	Technical Implementation of the Opt-Out	125
5.2.1.	Robots.txt — From Niche Signal to Default Battleground.....	126
5.2.2.	Metadata and Embedded Rights Information — A Two-Tier System	128
5.2.3.	Emerging Standards and Initiatives — From Technical Promise to Regulatory Process	129
5.3.	Structural Problems with the Opt-Out Mechanism.....	131
5.3.1.	Burden Allocation Asymmetry	131
5.3.2.	Historical Data Problem.....	133
5.3.3.	Enforcement Gap	134
5.3.4.	Global Mismatch.....	135
5.4.	Licensing Practices and Market Developments	138
5.4.1.	Emerging Direct Licensing Markets	138
5.4.2.	Collective Licensing Models	142
5.4.3.	Critical Assessment.....	144
5.5.	Balancing Innovation and Rightholder Protection.....	145
5.6.	Answer to Research Question 3	151
6.	Liability Issues and Future Regulatory Perspectives	152
6.1.	Copyright Liability for AI Training Under Current EU Law	152
6.1.1.	Direct Infringement.....	152
6.1.2.	Liability for Use of Pirated Datasets.....	155
6.1.3.	Remedies Available to Rightholders	156

6.2.	EU AI Act — Relevance for Copyright Law.....	159
6.3.	Current Litigation Landscape	161
6.3.1.	US Litigation — From Factual Record to Legal Signal	162
6.3.2.	EU Developments — What the Case Law Suggest About Sequencing	163
6.4.	Future Regulatory Perspectives	165
6.4.1.	The Existing Reform Landscape.....	165
6.4.2.	Legislative Reform Options.....	167
6.4.3.	A Boundary Case — The Japanese Model	170
6.4.4.	Preferred Reform Proposal: A Combination of Options A and B	171
7.	Conclusion	174
7.1.	Summary of Findings by Research Questions	174
7.2.	Overall Assessment.....	176
7.3.	Reform Recommendations.....	177
7.4.	Broader Significance.....	179
8.	Bibliography	181

List of Abbreviations

Abbreviation	Definition
Art	Article
AI	Artificial Intelligence
CFREU / the Charter	Charter of Fundamental Rights of the European Union [2012] OJ C326/391
CJEU / the Court	Court of Justice of the European Union
Database Directive	Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L77/20
DSM Directive	Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market [2019] OJ L130/92
Enforcement Directive	Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights [2004] OJ L157/45
ECL	Extended Collective Licensing
EU	European Union
EU AI Act	Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence [2024] OJ L1689/1
EUIPO	European Union Intellectual Property Office
General Data Protection Regulation	Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) [2016] OJ L119/1
GPAI	General-Purpose Artificial Intelligence
InfoSoc Directive	Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L167/10
IP	Intellectual Property
IP address	Internet Protocol Address

IPRED	Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights [2004] OJ L157/45
LLM(s)	Large Language Model(s)
RAM	Random-Access Memory
TDM	Text and Data Mining
TFEU	Treaty on the Functioning of the European Union
The Commission	The European Commission
US	United States

1. Introduction

1.1. Problem Statement and Relevance

The development of large language models and other generative artificial intelligence systems fundamentally relies on access to extensive collections of text, images, and other data sourced from the internet. Training corpora such as Common Crawl, which support many of today's most prominent models, comprise hundreds of billions of individual documents gathered through automated web scraping. Within these corpora, copyrighted works, including news articles, novels, academic papers, photographs, source code, and numerous other forms of original expression, are not incidental but serve as a structural necessity. The statistical training processes that enable LLMs to function depend on the richness, diversity, and expressive qualities of the training material, the majority of which is protected by copyright in its respective jurisdictions. Consequently, the use of copyrighted works for AI training is a central issue in the development of modern AI systems.

The commercial and economic significance of this practice is considerable. Investment in foundational models and the infrastructure required to train them has reached hundreds of billions of dollars globally, with leading AI developers securing successive funding rounds directly linked to training increasingly large models on expanding datasets. Simultaneously, rights holders whose works are included in these datasets, such as authors, journalists, publishers, photographers, and other creative professionals, often did not consent to such use, received no compensation, and were frequently unaware that their works had been incorporated into training data. The resulting conflict between the technology industry and the creative and publishing sectors has emerged as one of the defining intellectual property disputes of the current decade.

The legal status of this practice under European Union law remains largely unresolved. The InfoSoc Directive, which established a harmonised right of reproduction across the European

Union, was adopted in 2001, well before the emergence of current AI training methods. Its provisions must now be applied to a technological context that its drafters could not have anticipated.

The DSM Directive, adopted in 2019, introduced exceptions for text and data mining that are widely regarded as the most significant legislative provisions for AI training under current EU law. However, these exceptions were primarily designed for scientific research and traditional data analysis. Their application to the training of generative AI models raises interpretive questions that neither the legislator nor the Court of Justice of the European Union has yet resolved. Consequently, a significant gap exists between the technological reality of modern AI system development and the legal framework intended to regulate the use of underlying materials.

This gap has become particularly acute for several interrelated reasons. First, legal proceedings concerning the use of copyright-protected works in AI training have increased rapidly, especially in the United States. High-profile cases brought against major AI developers by authors, news publishers, and artists have resulted in settlements and, in some instances, court rulings that may influence the global discourse, despite significant differences between US and EU legislation.¹ Second, national courts across the European Union have begun to address these issues directly, as illustrated by the Hamburg Regional Court's decision concerning the use of images for AI training under the German implementation of the DSM Directive's TDM exceptions, and by a pending reference to the CJEU that may, for the first time, require the Court to engage with the application of the TDM exceptions to AI training. Third, the EU's regulatory framework continues to evolve.² The AI Act introduces transparency obligations regarding training data for general-purpose AI models, and the European Commission and Parliament, through consultations and resolutions adopted in late 2025 and early 2026, have indicated that

¹ See Chapter 4, sub-section 4.4.2 of this thesis; see Chapter 6, sub-section 6.3.1 of this thesis.

² See Chapter 5, section 5.5 of this thesis; see Chapter 6, sub-section 6.3.2 of this thesis.

further legislative clarification of the interaction between copyright and AI may be necessary.³

In light of ongoing doctrinal uncertainty, litigation, and regulatory developments, a systematic legal analysis of how the existing EU copyright system applies to AI training, and where its gaps and contradictions lie, is both timely and necessary.

1.2. Research Questions

This thesis is structured around three research questions, each addressing a distinct stage of the legal analysis necessary to determine the status of AI training under EU copyright law.

RQ1: Does the training of AI models constitute copyright-relevant reproduction under EU law?

RQ2: To what extent do the Text and Data Mining exceptions of the DSM Directive permit the use of protected works for AI training purposes?

RQ3: How effective is the opt-out mechanism under Article 4 of the DSM Directive in balancing innovation interests and rightholders protection?

These research questions are designed to be cumulative. RQ1 determines whether the exclusive reproduction right under Article 2 of the InfoSoc Directive applies, and at which stages of the AI training process. Without an affirmative answer to this threshold issue, further analysis of exceptions would be unnecessary. RQ2 assumes that reproduction occurs and evaluates whether the TDM exceptions introduced by Articles 3 and 4 of the DSM Directive, as well as their interaction with the temporary reproduction exception under Article 5(1) of the InfoSoc Directive, provide a legitimate basis for the activities identified in RQ1. RQ3 focuses on the rightholder's waiver under Article 4(3) of the DSM Directive, which has received significant practical and academic attention, and assesses whether this mechanism, both as designed and

³ See Chapter 6, section 6.2 of this thesis.

as currently implemented, achieves the intended balance between the interests of AI developers and rightholders.

1.3. Scope and Delimitations

This thesis confines its analysis to European Union copyright law, with particular emphasis on the InfoSoc Directive and the DSM Directive as the principal sources of harmonised rules governing the reproduction right and relevant TDM exceptions. The EU AI Act is addressed only insofar as it intersects with the copyright framework, specifically regarding transparency obligations for training data under Article 53 and their relationship to the TDM exemptions regime. The broader regulatory structure of the AI Act, including its risk-based classification system and governance provisions, is excluded from the scope of this thesis unless directly relevant to the copyright analysis.

References to United States legislation, particularly the fair use doctrine under Section 107 of Title 17 of the US Code and ongoing litigation involving major AI developers, are included exclusively for comparative purposes. These references serve to illustrate how a differently structured legal system addresses functionally similar issues and to contextualise practical factors influencing the debate within the EU. However, this thesis does not undertake an independent doctrinal analysis of US copyright law, nor does it treat US case law as binding or persuasive authority for interpreting EU law.

The technical aspects of AI development are addressed only insofar as they are necessary to support the legal analysis. Chapter 2 describes the architecture and training process of large language models, yet this discussion is intentionally limited to technical features that subsequently inform the application of EU copyright concepts, such as reproduction, permanence, and originality in Chapter 3. The thesis does not aim to be technically comprehensive and does

not engage with internal debates within computer science that lack clear relevance to the legal issues examined.

Finally, the scope of this thesis is restricted to the input data of AI systems, specifically the use of copyright-protected works during the training process. It does not address the distinct set of legal issues associated with the outputs generated by AI systems. These include the copyright status of AI-generated content, potential liability for copyright infringement due to reproduction or similarity of outputs to training data, and questions of authorship and ownership of works created using AI tools. While these output-related issues are connected to the topics discussed and may arise as a consequence, they present distinct legal questions that require separate analysis and are therefore excluded from the scope of this thesis.

1.4. Methodology

The thesis employs doctrinal legal analysis as its primary method. This approach relies on relevant provisions of primary and secondary EU legislation, particularly the InfoSoc Directive and the DSM Directive. The objective is to determine the meaning and scope of these provisions using established methods of legal interpretation. Where the wording of a provision is sufficiently clear to resolve the issue, the analysis begins with a literal interpretation. If the wording is ambiguous or its application to AI training was not clearly intended by the legislator, the analysis proceeds to systematic interpretation, examining the provision within the broader structure of the directive and its relationship with other provisions. Further, teleological interpretation is employed to examine the purpose the provision is intended to serve, and historical interpretation draws on legislative history, including Commission proposals, impact assessments, and Parliamentary reports, where these documents clarify the legislator's intentions.

Given the central role of the CJEU in interpreting harmonised concepts such as reproduction and originality, particular attention is devoted to the analysis of the Court's case law. The analysis begins with the landmark judgment in *Infopaq International* and continues with subsequent judgments, including *Painer*, *Pelham*, *Funke Medien*, and *Cofemel*.⁴ As the Court has not yet issued a direct ruling on the application of these concepts to AI training, the principles established in the aforementioned case law are applied to the context of AI training by analogy. The analysis clearly distinguishes where such analogous reasoning introduces genuine uncertainty rather than established doctrine.

Finally, this study draws extensively on secondary literature, including academic commentaries on the InfoSoc Directive and the DSM Directive, working papers, and policy analyses concerning TDM exceptions, particularly in relation to AI. More recent commentaries, such as analyses by practitioners, consultations with regulatory bodies, and reports on ongoing litigation and licensing-related developments, are also considered. These sources reflect the rapidly changing practical environment in which this legal analysis is conducted. The secondary literature is synthesised throughout the study to support the doctrinal analysis and to identify points of genuine disagreement within the academic and practical communities.

1.5. Structure of the Thesis

Following the present introductory chapter, Chapter 2 establishes the technical foundations required for the subsequent legal analysis by describing the architecture of large language models and the stages of the AI training process, from data collection through preprocessing to the training procedure itself. Chapter 3 builds on this foundation to address Research Question 1 by examining whether, and at which stages of the AI training pipeline, the reproduction right

⁴ See Chapter 3 of this thesis.

under Article 2 of the InfoSoc Directive applies. It also addresses the threshold issue of originality, which determines which materials in the training corpus are primarily subject to copyright protection. Chapter 4 addresses Research Question 2 by providing a detailed analysis of the exceptions for TDM under Articles 3 and 4 of the DSM Directive. This includes their respective scopes of application, conditions for application, the rightholder's opt-out mechanism under Article 4(3) of the DSM Directive, and the relationship between these exceptions and the exception for temporary reproduction under Article 5(1) of the InfoSoc Directive. Chapter 5 addresses Research Question 3 by assessing the practical functioning of the opt-out regime under Article 4 of the DSM Directive in light of recent court proceedings, licensing practices, and legislative changes, and evaluates the extent to which it balances the interests of AI developers and rightholders. Chapter 6 summarises the findings of the preceding chapters, identifies key gaps and contradictions in the existing legal framework, and explores possible avenues for legislative or judicial clarification. The thesis closes with a final conclusion summarising of its overall contribution.

2. Technological Background of AI Training and Its Legal Relevance

2.1. Machine Learning — Basic Concepts

Machine learning comprises a set of methods that enable a computer system to improve its task performance through data processing, rather than by executing explicitly programmed rules.⁵

Three principal paradigms are typically distinguished, based on the structure of the data and the objectives of the system.

In supervised learning, the system is trained on examples where each input is paired with a known correct output, such as an image accompanied by a descriptive label. The system learns to predict outputs from inputs, and its performance is evaluated against these known answers.

In contrast, unsupervised learning involves data without labels, requiring the system to identify patterns, regularities, or relationships within the data itself. The third paradigm, reinforcement learning, involves training the system through trial and error, reinforcing actions that yield improved outcomes over time.⁶

LLMs, which are the focus of this thesis, are primarily trained using a form of unsupervised learning often referred to as self-learning. In this approach, the training data consists of unlabelled text, and the system's task is to predict segments of text based on surrounding context, such as predicting the next word in a sequence from preceding words.⁷ Reinforcement learning is introduced in subsequent stages, where a pre-trained model is further refined using feedback on its performance. This thesis concentrates on the initial, large-scale training phase, during

⁵ Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning* (MIT Press 2016) ch 1, 5.

⁶ Goodfellow, Bengio and Courville (n 5) ch 1, 5.

⁷ Nicolas Carlini and others, 'Extracting Training Data from Large Language Models' (2021) USENIX Security Symposium 2633, 2634.

which the majority of the model's training corpus is processed. The later stage of fine-tuning, which involves distinct questions regarding the data, is not addressed here.

Distinguishing between these paradigms is essential for the subsequent analysis, as it clarifies the relationship between individual training data and the resulting system. In supervised learning, data is deliberately selected and annotated, with explicit decisions made regarding the representation of each example. By contrast, in the self-training of large language models, no such individual selection occurs. Each text fragment is one among billions, and the model extracts statistical patterns without any fragment being labelled, selected, or processed differently from others.⁸ This aggregated, statistical approach to training, as opposed to individual selection, forms the basis for this thesis's examination of what the model "learns" from its training data. This issue will be explored further in the following Section and revisited from a legal perspective in Chapter 3.

A neural network, which underpins LLMs, can be conceptualised as a highly complex mathematical function. This function receives input data, such as a sequence of text, and generates output data, such as a prediction of subsequent elements. The network's capacity to learn arises from a vast number of adjustable numerical values, commonly referred to as parameters or weights. Training involves iteratively adjusting these weights based on the discrepancy between the function's output and the expected output for each training example, until the predictions become consistently accurate across the entire training dataset.⁹

For the purposes of this thesis, the primary concern is the state of the model after training is complete. The trained model comprises its weights, a large set of numerical values, and a function structure parameterised by these values.¹⁰ This structure does not function as a storage system for training examples; the model does not, in a strict sense, contain copies of the texts

⁸ Carlini and others (n 7) 2633, 2634.

⁹ Goodfellow, Bengio and Courville (n 5) ch 5.

¹⁰ Goodfellow, Bengio and Courville (n 5) ch 6.

used during training, unlike a hard drive that stores files. The nature of what the weights contain, and the extent to which they may be related to or embody the specific material used in training, is not addressed in this chapter. This issue is examined in detail in Chapter 3, where it emerges as one of the most complex topics considered. The present discussion is limited to defining what a weight is, to clarify the subsequent analysis in Chapter 3 regarding the connection, if any, between weights and the training data from which they are derived.

2.2. Large Language Models — Architecture and Training Process

Building of a LLM involves a series of distinct stages, beginning with the construction of a training corpus and culminating in a deployable model. The present Section outlines each stage in sequence, focusing exclusively on technical aspects. The legal implications of each stage, particularly regarding the right of reproduction under EU copyright law, are addressed in Chapter 3, which follows the same structure.

Data collection

The initial stage involves assembling a training corpus, which is a collection of text used to train the model. This corpus is typically compiled from multiple sources. Web scraping employs automated software to extract content from publicly accessible web pages on a large scale, resulting in copies that are stored on the developer's infrastructure. Curated datasets, which are pre-compiled collections of text assembled either by third parties or by the developer, may also be incorporated in whole or in part. Additionally, a smaller portion of the training data may be sourced through specific agreements with rights holders, as discussed later in relation to licensing.¹¹

¹¹ See Chapter 5 of this thesis (licensing practices).

The key outcome of this stage is the preservation of the collected content on the developer's systems in a form that is nearly identical to the original sources. For example, a web page acquired via scraping and saved as a text or web page file constitutes a copy equivalent to one produced by any other duplication method. This stage is conceptually straightforward, as it requires no specialised technical knowledge: copies of the source material are simply created and stored. The following stages address the subsequent processing of these copies.

Tokenization and preprocessing

Prior to training, the collected text must be transformed into a format suitable for neural network processing, as neural networks operate on numerical data rather than raw text. This transformation, known as tokenisation, involves segmenting the text into units called tokens, each represented by a numerical value. Depending on the tokenisation scheme, a token may correspond to a word, a subword, or a single character. To illustrate, the sentence “the cat sat down” can be tokenised into the units “the”, “cat”, “sat”, and “down”, each mapped to a unique number for the network's use. In practice, tokenisation schemes are often more complex, frequently dividing uncommon words into smaller components. Nevertheless, the fundamental principle remains: a text fragment is converted into a sequence of numbers that represents it.¹²

Before or concurrently with tokenisation, the collected corpus typically undergoes pre-processing. This may involve identifying and removing duplicate content, filtering out specific categories of material, and normalizing formatting across sources with differing structures. As a result of tokenisation and pre-processing, the corpus is transformed from its original form into a large array of numerical sequences immediately prior to training.

¹² Goodfellow, Bengio and Courville (n 5) ch 5; Sebastian Raschka, *Build a Large Language Model (From Scratch)* (Manning 2024) ch 2.

The training process

Training is conducted through the repeated execution of a loop over the entire tokenised corpus. At each iteration, the model in its current state receives a sequence of tokens and generates a prediction, typically regarding the most probable next token in the sequence. This prediction is then compared to the actual subsequent token in the training data. The difference between the prediction and the actual token informs an adjustment to the model's weights, which, if applied earlier, would have improved the model's prediction accuracy. The process of calculating this adjustment is referred to as gradient calculation, while the iterative updating of weights is known as gradient descent.¹³ Such an adjustment consists of a set of numerical values indicating the direction and magnitude of change for each weight.

This cycle of prediction, comparison, and adjustment is repeated for every position within every sequence in the training corpus. The entire corpus may be processed multiple times, with each complete pass referred to as an epoch.¹⁴ Through numerous adjustments, the model's weights transition from an initial random configuration to one that enables accurate predictions. During this process, no copy of the training example is retained in memory as a result of each step; instead, only the model's weights are updated, rather than creating new stored artifacts resembling the input data.

What the model learns

The cumulative effect of this process, which encompasses a large and diverse corpus, is a set of weights that, when considered collectively, encode statistical regularities. These include patterns of language use, structures, and associations that recur across numerous training examples and are generalised to text on which the model has not been trained. This characterisation of what a trained model 'contains' is standard, and for the vast majority of any given training

¹³ Goodfellow, Bengio and Courville (n 5) ch 4, 8.

¹⁴ Goodfellow, Bengio and Courville (n 5) ch 8.

corpus, it remains accurate: no single example can be identified within the weights, as these weights reflect the aggregate signal accumulated over billions of training steps, with each example contributing only a minimal adjustment among countless others.¹⁵

However, this characteristic does not apply universally. Empirical studies of trained language models have demonstrated that, for certain training examples, the model can be tuned to produce outputs that correspond exactly or nearly exactly to a specific example. These outputs do not merely reflect general patterns; instead, they reproduce the precise sequence of words (or, for image models, specific visual content) present in the training data. This phenomenon, known as memorisation, occurs more frequently for content that appears repeatedly in the training corpus and for content that is exceptionally distinctive, such as sequences unlikely to arise from general patterns learned by the model.¹⁶ Memorisation is discussed in detail in the following Section, being introduced here to clarify the general characterisation provided above and since the discussion in Chapter 3 regarding what the weights of a trained model represent depends on both the general case, where no single instance can be identified, and the refinement, where identification is possible in some cases.

The pipeline as a whole

The stages described in the present Section form a sequence. Content is typically collected through parsing and stored as a copy of the source material. The stored corpus is then tokenised and pre-processed into numerical sequences. These sequences are used in iterative cycles of prediction, comparison, and adjustment to train the model's weights. The resulting weights, which primarily encode generalised patterns yet in some cases reflect specific memorised content, constitute the trained model, which is subsequently deployed and made available for use by others, typically through an application or interface that allows users to submit queries and

¹⁵ Patrick Henderson and others, 'Foundation Models and Fair Use' (2023) 24 *Journal of Machine Learning Research* 1, 7.

¹⁶ Carlini and others (n 7) 2633, 2635–2636.

receive generated responses. Each of these stages (collection and storage, tokenisation, training, resulting weights, and deployment) raises distinct questions regarding the right of reproduction under EU copyright law. These issues are examined in this thesis in the order indicated, beginning with the Chapter 3.

2.3. The Memorisation Phenomenon

The preceding Section introduced memorisation as a clarification of the general characteristic of trained model weights, which encode statistical patterns rather than specific examples. The present Section directly examines this phenomenon, summarising its definition, empirical findings regarding its occurrence, and approaches to organising these insights for subsequent legal analysis.

Memorisation is the capacity of a trained model to generate, in response to a relevant query, an output that exactly or closely matches a specific element from its training data, rather than merely reflecting patterns common to many examples. The distinction is one of traceability: a generalised output reflects the corpus as a whole, with no single training example identifiable as its source, whereas a memorised output can, at least in principle, be traced to a specific example.

The existence of memorisation in trained models is well established in the literature; ongoing research focuses on clarifying its scale. Carlini et al. demonstrated that language models can reproduce significant verbatim fragments from their training data, a finding confirmed by subsequent larger-scale studies and analyses across various model sizes.¹⁷ Three factors are frequently identified as correlating with memorisation: the frequency of a fragment in the training corpus, the uniqueness of the fragment (i.e., its improbability of arising from general patterns),

¹⁷ Carlini and others (n 7) 2633, 2635; Milad Nasr and others, ‘Scalable Extraction of Training Data from (Production) Language Models’ (2023) arXiv:2311.17035.

and the size of the model, with larger models generally exhibiting greater memorisation capacity.¹⁸ Reported measures of memorisation vary considerably depending on the model, methodology, and measurement approach. Accordingly, this analysis does not treat any specific figure as definitive, but relies on the robust conclusion that memorisation exists, varies according to the factors described, and is, in principle, measurable.

For analytical purposes, memorisation can be conceptualised as a spectrum, ranging from the most directly observable forms of recall to the least. At one end, the model may reproduce a passage exactly as it appeared during training (verbatim recall). Slightly further along, the model may generate a passage with only minor modifications (near-verbatim recall). Progressing along the spectrum, outputs may reflect distinctive structural or expressive elements of a specific work, such as a particular narrative structure, characteristic phrasing, or a recognisable sequence of code, without reproducing the text verbatim. At the opposite end, outputs may reflect only stylistic or genre conventions common to many works, making it impossible to identify any single source.

This spectrum serves solely to organise the empirical findings described in the literature and does not establish legal boundaries. Nonetheless, positions at either end of the spectrum differ significantly in relation to issues commonly addressed by copyright law. Verbatim or near-verbatim reproduction of a work is the type of act traditionally addressed by copyright. In contrast, reflecting common stylistic or genre conventions has not been regarded as infringing the rights of individual authors; it is not generally asserted that a writer infringes the rights of every author whose style has influenced their own. The question of where a legally significant boundary may lie between these extremes, and its implications for characterising the weights of a

¹⁸ Stella Biderman and others, ‘Emergent and Predictable Memorization in Large Language Models’ (2023) *Advances in Neural Information Processing Systems* 37; Carlini and others (n 7) 2636.

trained model, is addressed in the following Chapter, which considers multiple perspectives without reaching a definitive conclusion.

This Section, therefore, establishes that memorisation is a real and well-documented phenomenon, not merely a theoretical possibility; that it varies in degree along a spectrum from verbatim reproduction to general stylistic influence; and that its occurrence correlates with measurable factors such as duplication, distinctiveness, and model scale. However, it does not define the implications of any point on this spectrum for the right to reproduce under EU copyright law or for the characterisation of trained model weights. Chapter 3 addresses this issue, returning to the spectrum to consider the unresolved question of the relationship, if any, between model weights and the data used to train them.

2.4. AI Training Distinguished from Other Data Processing

Previous Sections have characterised AI training as a process that produces a stable artefact, namely a trained model, from a corpus of collected text. The model's weights encode generalised patterns and, in some instances, memorised content that can be traced. Prior to addressing the legal analysis, it is necessary to clarify how this process relates technically to other forms of large-scale data processing encountered in copyright law. Three comparisons are particularly instructive: search engine indexing, database queries, and the text and data methods that predate large-scale AI training.

Search engines scan the web and store copies of discovered content for indexing purposes, a practice that has prompted numerous copyright questions, often focusing on the temporary or intermediate nature of these copies or their function in redirecting users to the original source.¹⁹

While AI training shares the initial step of scanning and storing copies, it diverges in two significant ways. First, a copy included in a training corpus is not temporary as a search index

¹⁹ Case C-360/13 *Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd* [2014] ECLI:EU:C:2014:1195.

entry might be; it is retained within the corpus and may be used repeatedly over an extended period. Second, the output of AI training does not serve as a pointer to its sources: a trained model's response to a query does not direct users to the underlying material, unlike a search result. Whether these differences influence the application of existing legal regulations on search caching to AI training will be addressed in the next chapter. At this stage, it is sufficient to note that, from a technical perspective, the two processes are not equivalent.

A traditional database is a structured repository that enables the retrieval of specific records in response to specific queries, establishing a direct and traceable link between query and result.²⁰ AI training operates differently. As previously discussed, the output of a trained model is generated based on the properties of the entire corpus and, except in cases involving memorisation, cannot typically be traced to any particular element in the manner of a database search. Therefore, in a technical sense, a trained model is not merely an extensive database. The question of how the sui generis database right interacts with TDM exceptions will be addressed later in sub-section 4.5.2 of this thesis. The present point is narrower: the relationship between AI training and its input data is fundamentally distinct from the relationship between a database and its records, depending on the underlying process.²¹

The most significant comparison is with traditional text and data analysis practices that predate large-scale foundational models. Conventional training based on TDM typically involves assembling a specific corpus for a defined analytical purpose and applying a particular analytical process, resulting in outputs such as statistics, trends, or summaries.²² This approach does not transform the corpus into a new artefact that persists independently of the analysis. In contrast, training based on foundational models produces an output that is itself a model — an artefact capable of generating results independently of any specific query about the training data and

²⁰ Database Directive, art 7.

²¹ See Chapter 4, sub-section 4.5.2 of this thesis.

²² Patrick Henderson and others, 'Foundation Models and Fair Use' (2023) 24 *Journal of Machine Learning Research* 1, 9–11.

continuing to function after training concludes. This distinction concerns the nature of the output produced, rather than its scale. This thesis will revisit this distinction when evaluating the suitability of existing legal frameworks for foundational model training.

These three comparisons highlight a general feature that distinguishes AI training from the other processes discussed. Search caching, database queries, and traditional TDM analytics maintain a persistent connection between the base corpus or derived results and their sources: caches reference specific pages, database records remain identifiable, and analytical conclusions pertain directly to the analysed corpus. In contrast, AI training transforms the corpus into a model whose value is independent of any ongoing connection to its sources. This thesis refers to this phenomenon as permanent value extraction, wherein the value contributed by the training corpus is embedded within the model, severing ties to its origin. If a model's output can substitute for the content on which it was trained, this extraction may also introduce a competitive dimension. However, this possibility is only noted here, as questions regarding the model's output fall outside the scope defined earlier in this Chapter.

2.5. Legal Relevance of Technical Architecture

The preceding Sections outline the following sequence: content is collected and stored as a copy of its original form; this copy is tokenised and pre-processed into numerical sequences; these sequences, through repeated adjustment, determine the weights of the trained model; these weights generally encode generalised patterns, but in some instances reflect specific memorised content; and the resulting model is subsequently used. Each stage of this sequence raises a distinct question regarding the reproduction right under Article 2 of the InfoSoc Directive.²³ These questions are examined in the following Chapter in the order presented.²⁴

²³ InfoSoc Directive, art 2.

²⁴ See Chapter 3 of this thesis.

At the collection and storage stage, a copy of the source material exists; the relevant question is whether it satisfies the conditions for reproduction established in Article 2 of the InfoSoc Directive. During tokenisation, this copy is converted into a numerical representation, raising the issue of whether such conversion excludes it from the scope of Article 2 or whether it remains the same work in a different form. At the training stage, the model's weights are adjusted using a signal derived from each training example, prompting the question of whether this adjustment constitutes reproduction, and, if not, what implications arise from the signal's specificity to the example. Regarding the trained weights, the unresolved question (particularly in light of the data on memorisation discussed above) is whether the weights embody the training data in any manner relevant to Article 2. Upon deployment, the model becomes available for use; the question of the legal status of its output data is not addressed here, although the continued existence of the deployed model may be relevant to potential remedies for infringements occurring at earlier stages.

Chapter 3 examines this sequence in the order outlined above, beginning with collection and storage, where the technical description most closely aligns with traditional notions of copying. The analysis then proceeds to tokenisation, training, and trained weights, where the technical description corresponds less directly with categories previously addressed in copyright law.

3. The Reproduction Right under EU Law and Its Application to AI Training

3.1. The Reproduction Right in the InfoSoc Directive

Prior to assessing whether the training of AI falls within statutory exceptions permitting the use of copyrighted works, it is necessary to determine whether such training implicates copyright protection. It would be illogical to analyse existing exceptions without confirming that the relevant right is engaged. Accordingly, the central issue is whether the process of training AI models constitutes reproduction within the meaning of the InfoSoc Directive²⁵. This is precisely the question this Chapter addresses.

3.1.1. Statutory Analysis of Article 2

Article 2 of the InfoSoc Directive grants authors “the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part” of their works²⁶. Notably, this provision is broadly formulated, and each of its core elements holds distinct legal significance for analysing the AI training process.

The phrase “direct or indirect” reproduction ensures that the exclusive right is not limited to acts of direct copying. Instead, it also covers reproduction acts conducted through intermediate processes or automated technical systems, including those in which a human operator does not personally perform each individual act of copying.²⁷ In the context of AI training, works are reproduced through automated data collection tools and machine learning pipelines without direct human intervention at the level of individual copies, which makes this element particularly significant, since it assures that the automated and indirect nature of the AI reproduction

²⁵ InfoSoc Directive.

²⁶ InfoSoc Directive, art 2.

²⁷ Thomas Dreier and Bernt Hugenholtz (eds), *Concise European Copyright Law* (2nd edn, Kluwer Law International 2016) 341.

process does not, by itself, place this activity outside the scope of Article 2 of the InfoSoc Directive.

The explicit reference to both “temporary” and “permanent” reproduction confirms that the duration of a copy’s existence is irrelevant to its classification as a reproduction under Article 2 of the InfoSoc Directive. In fact, there is no minimum duration threshold. A copy need not exist either indefinitely or for a specific period; even a reproduction that occurs only for a brief technical moment falls within the scope of this provision.²⁸ This element also directly affects the assessment of AI training: the fact that some copies created during the training process (such as data loaded into memory during individual processing cycles) may be technically transient does not exempt them from the reproduction right. However, the temporary nature of such copies is relevant to the separate question of whether the exception provided in Article 5(1) of the InfoSoc Directive applies, which is discussed in Section 3.4 below.

The expression “by any means and in any form” underlines the technological neutrality of Article 2 of the InfoSoc Directive. The right of reproduction is not limited to specific methods or formats of copying (known at the time of the Directive’s adoption). Instead, it covers any technical means by which a copy may be made, including digital reproduction, format conversion, and automated data processing.²⁹ This technological neutrality is particularly significant in the context of AI training, which involves such reproduction methods as large-scale web scraping, data set compilation, and iterative computational processing. Although these methods were clearly not specifically contemplated by the EU legislature in 2001, they fall fully within the scope of this provision.

Finally, the phrase “in whole or in part” establishes that reproducing a mere fragment or portion of a work is sufficient to engage an exclusive right. Article 2 of the InfoSoc Directive does not

²⁸ Jörg Reinbothe and Silke von Lewinski, *The WIPO Treaties on Copyright: A Commentary* (2nd edn, Oxford University Press 2015) 112.

²⁹ Martin Senftleben, ‘Breathing Space for Cloud-Based Business Models’ (2013) 4 JIPITEC 87, 89.

set a quantitative threshold, such as a minimum percentage or a minimum number of words or elements.³⁰ The legal significance of partial reproduction is determined not by the quantity of material copied but, as the Court of Justice of the European Union has confirmed in its case law, by the qualitative originality of the reproduced fragment. This issue is examined in detail in Section 3.2.

Despite its broad wording, Article 2 of the InfoSoc Directive does not define the term “reproduction”. The provision outlines the scope of the exclusive right in formal terms but does not specify the conditions under which a particular act of copying constitutes a legally relevant reproduction for copyright law purposes. This legislative omission is intentional; the InfoSoc Directive was intended to establish flexible yet universally applicable principles that could adapt to future technological developments, as indicated in Recital 21, which states that the scope of acts covered by the right of reproduction should be defined broadly.³¹ As a result, the substantive content of the reproduction right, particularly the threshold for when a partial or fragmentary act of copying falls within the scope of copyright protection, has been shaped exclusively through the interpretive jurisprudence of the CJEU. Consequently, the case law of the CJEU serves as the primary source for determining whether AI training activities constitute reproduction within the meaning of Article 2. The relevant jurisprudence is further examined in Section 3.2.

The right of reproduction, as set out in Article 2 of the InfoSoc Directive, must be perceived as an autonomous legal concept within the EU, the content of which is determined by EU law regardless of the legal traditions of individual Member States. The InfoSoc Directive seeks to harmonise specific aspects of copyright law within the internal market, and achieving this objective requires that the Directive’s key concepts possess a uniform meaning across all Member

³⁰ InfoSoc Directive, art 2; Emanuela Rosati, *Originality in EU Copyright: Full Harmonization through Case Law* (Edward Elgar 2013) 45.

³¹ InfoSoc Directive, recital 21.

States.³² Divergent national interpretations of the reproduction notion would undermine the Directive's harmonisation objective and result in legal fragmentation that is incompatible with the internal market functioning. The CJEU confirmed this principle in *Infopaq International A/S v Danske Dagblades Forening*, holding that the concepts used in the InfoSoc Directive must be interpreted autonomously and uniformly throughout the European Union.³³ For the purposes of the present analysis, this principle has a significant methodological implication: the question of whether AI training constitutes reproduction cannot be resolved by reference to any particular national copyright doctrine, but must be determined exclusively by reference to EU law, specifically the case law of the CJEU. Hence, the analysis now turns to this jurisprudence.

3.1.2. Relationship to Other Rights

Article 2 of the InfoSoc Directive constitutes one of the three exclusive rights outlined in that Directive. The remaining two are the right to public information under Article 3 and the right of distribution under Article 4. It is essential to provide a brief overview of the scope of Article 2 in relation to these other rights to justify the focus of the present Chapter.

The right of communication to the public under Article 3 of the InfoSoc Directive covers acts by which a work is made available to the general public. A typical example is the transmission of content via the internet to an audience.³⁴ This right does not extend to the internal processing and copying operations involved in AI training. Model training entails the reproduction of works within a closed technical pipeline — works are copied, processed, and used to adjust the model's parameters, but are not transmitted or disclosed to the public during this process. The right to communicate works to the public may become relevant at a subsequent stage, when an AI system generates and demonstrates results that reproduce a protected expression. However,

³² InfoSoc Directive, recital 1; art 114 TFEU.

³³ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, paras 27–28.

³⁴ InfoSoc Directive, art 3(1).

this issue concerns the outputs of the AI system, not its training process, and therefore falls outside the scope of this thesis.³⁵

The distribution right established in Article 4 of the InfoSoc Directive concerns the making available to the public the original or copies of a work through sale or other forms of ownership transfer.³⁶ This right typically encompasses the physical or digital distribution of copies to end users. Similar to the right of communication to the public, the right of distribution does not extend to activities underlying AI training, as these involve internal reproduction within a technical system rather than the transfer of copies to third parties. Consequently, the right of distribution is excluded from the scope of this Chapter.

Thus, the right of reproduction provided for in Article 2 of the InfoSoc Directive constitutes an exclusive right that is primarily engaged by the activities underlying AI training. The scope and content of this right, as developed through the interpretive jurisprudence of the CJEU, are the primary focus of the subsequent analysis.

3.2. The EU-Autonomous Concept of Reproduction and the Originality Threshold

3.2.1. EU-Autonomous Interpretation

As established in the preceding Section, Article 2 of the InfoSoc Directive does not define the notion of reproduction; therefore, the CJEU was tasked with determining its substantive meaning. The Court based its approach on a fundamental methodological principle: the concepts used in the InfoSoc Directive possess an autonomous and uniform meaning across the EU, regardless of the legal definitions employed by individual Member States.³⁷ This principle of

³⁵ Peter Bernt Hugenholtz, ‘The New Copyright Directive: The EU Content Directive (Directive 2019/790)’ (2021) 7 JIPITEC 3, 4.

³⁶ InfoSoc Directive, art 4(1).

³⁷ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, paras 27–28; Case C-683/17 *Cofemel — Sociedade de Vestuário SA v G-Star Raw CV* [2019] ECLI:EU:C:2019:764, para 29.

EU-autonomous interpretation perfectly aligns with the Directive’s harmonisation objective. Allowing Member States to independently define reproduction would undermine the aforementioned objective, as divergent national standards could result in inconsistent levels of protection for identical acts of copying across jurisdictions.³⁸

The practical implications of this principle are considerable. National courts are not permitted to apply their own copyright law traditions when determining whether a specific act constitutes a reproduction under Article 2 of the InfoSoc Directive. Instead, they are bound to apply the standard established by the CJEU. This requirement holds even when a Member State’s national copyright legislation has historically imposed a different or higher originality threshold. The CJEU’s autonomous standard therefore functions as both a minimum and a maximum limit: it defines the required level of originality for copyright protection across all Member States and prohibits national courts from altering this threshold through domestic doctrine.³⁹

3.2.2. Infopaq International — The Foundational Case

The landmark judgement clarifying the scope of the reproduction right pursuant to Article 2 of the InfoSoc Directive is the decision of the CJEU in *Infopaq International A/S v Danske Dagblades Forening*.⁴⁰ Although this case involved a media monitoring service rather than artificial intelligence, it has established legal principles that still govern the evaluation of any act of partial reproduction under EU copyright law, including, as will be discussed in Section 3.3, the acts that constitute AI training.

Infopaq International A/S, a Danish company, offered media monitoring services to corporate clients. These services involved scanning printed newspaper articles, converting the scanned images into text using optical character recognition software, and extracting an eleven-word fragment surrounding each identified keyword. The resulting text snippets were temporarily

³⁸ InfoSoc Directive, recital 1.

³⁹ Rosati (n 30) 78.

⁴⁰ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569.

stored and subsequently transmitted to clients, after which the original digital files were deleted. Danske Dagblades Forening, the association representing Danish newspaper publishers, alleged that this process constituted unauthorised reproduction of its members' copyright-protected works.⁴¹

The Danish Supreme Court has referred two questions to the CJEU. Firstly, does the process of extracting and storing an eleven-word data fragment constitute 'partial' reproduction within the meaning of Article 2 of the InfoSoc Directive? Secondly, do the individual stages of the process constitute temporary acts of reproduction exempt from the reproduction right under Article 5(1) of that Directive?⁴² This section examines the Court's answer to the first question; the analysis of Article 5(1) is reserved for Section 3.4.

In response to the first question, the Court held that "partial" reproduction under Article 2 of the InfoSoc Directive encompasses the reproduction of a fragment of a protected work, provided that the fragment contains elements reflecting the author's own intellectual creation.⁴³ The Court based this interpretation on the explicit language of Article 2, which refers to reproduction "in whole or in part", thereby confirming that partial reproduction falls within the scope of the exclusive right. The decisive issue is therefore not one of quantity but of quality: whether the reproduced fragment, however short, reflects the original expression through which the author manifested creative form in his intellectual contribution.⁴⁴

The CJEU determined that the standard of originality required for reproduction is identical to that required to trigger copyright protection. A work, or a fragment thereof, is protected by copyright, and its reproduction constitutes an exclusive right, provided it reflects the author's own intellectual creation.⁴⁵ This equivalence between the standards for copyright subsistence

⁴¹ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, paras 14–22.

⁴² Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 23.

⁴³ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 37.

⁴⁴ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, paras 37–39.

⁴⁵ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 37; Case C-145/10 *Eva-Maria Painer v Standard VerlagsGmbH* [2011] ECR I-12533, para 87.

and infringement is analytically significant: any portion of a work sufficiently original to be independently protected is also subject to the right of reproduction when copied. In contrast, elements that do not meet the originality threshold, such as well-known expressions, purely factual statements, or components dictated solely by technical function, fall outside the scope of protection and may be freely reproduced.⁴⁶

The Court refrained from determining whether the specific eleven-word passages examined during the proceedings satisfied the criterion of originality, instead ruling that this assessment must be conducted by the national court, which applies the relevant EU standard to the facts of the case.⁴⁷ This allocation of responsibilities, whereby the CJEU establishes the standard and national courts implement it, is characteristic of the preliminary ruling procedure under Article 267 of the TFEU and does not diminish the significance of the established principles. The Court further clarified that eleven words may, in principle, constitute a reproduction infringing the exclusive right under Article 2 of the InfoSoc Directive, thereby rejecting the assumption that a de minimis quantity of copied material is automatically exempt from copyright protection.⁴⁸

The Infopaq International judgment articulated four key principles governing the analysis of the right of reproduction under EU law and directly relevant to the evaluation of AI training activities.

Firstly, the threshold of originality, defined as the requirement that the reproduced fragment reflects the author's own intellectual creation, determines the scope of the right of reproduction under Article 2 of the InfoSoc Directive. This right does not extend to elements of a work that lack originality.

⁴⁶ Dreier and Hugenholtz (n 27) 344.

⁴⁷ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 48.

⁴⁸ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 47.

Secondly, the length of the reproduced fragment is legally irrelevant. Whether reproduction has occurred is determined by the qualitative originality of the copied material rather than its quantitative extent. Even a brief excerpt may fall within the scope of the right of reproduction if it satisfies the originality threshold.

Thirdly, the right of reproduction is to be interpreted broadly. The Court’s reasoning affirms that the wording of Article 2 of the InfoSoc Directive, particularly the phrase “in whole or in part”, supports the expansive understanding of the right and precludes implied quantitative restrictions.

Fourthly, the concept of reproduction under Article 2 of the InfoSoc Directive constitutes an autonomous EU concept. Its content is uniformly defined by EU law, as applied and interpreted by the CJEU, regardless of national copyright traditions.⁴⁹

3.2.3. Infopaq II

The second case arising from the same proceedings, *Infopaq International A/S v Danske Dagblades Forening* (Infopaq II), was heard by the Court in 2012 and primarily addressed the conditions for the exception provided in Article 5(1) of the InfoSoc Directive.⁵⁰ For the present analysis, a key principle from that judgment is directly relevant to AI training: the Court confirmed that when a reproduction process involves several sequential steps, each step must be assessed independently according to the applicable legal standard.⁵¹ This principle has significant implications for AI training, which comprises a pipeline of technically distinct stages such as web scraping, dataset storage, pre-processing, and iterative training cycles. Each of these stages may independently constitute an act of reproduction within the meaning of Article 2 of

⁴⁹ Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, ‘The Exception for Text and Data Mining in the DSM Directive’ (2021) 52 IIC 174, 178.

⁵⁰ Case C-302/10 *Infopaq International A/S v Danske Dagblades Forening* (Infopaq II) [2012] ECLI:EU:C:2012:16, para 26.

⁵¹ Case C-302/10 *Infopaq International A/S v Danske Dagblades Forening* (Infopaq II) [2012] ECLI:EU:C:2012:16, paras 27–30.

the InfoSoc Directive. It is therefore insufficient for an AI developer to demonstrate that the process as a whole is exempt from the right of reproduction; rather, each individual stage must independently satisfy the conditions of any applicable exception. This principle is further examined and applied in Section 3.3.

3.2.4. Development of the Originality Standard in Subsequent Case Law

The standard of originality established in the Infopaq International case has been affirmed and further clarified in several subsequent CJEU judgments, each of which enhances understanding of the conditions under which reproducing a fragment of a work infringes the exclusive right under Article 2 of the InfoSoc Directive. The following four cases are particularly significant for evaluating the implications for AI training.

Painer — C-145/10

In *Eva-Maria Painer v Standard VerlagsGmbH*, the Court addressed copyright protection for portrait photographs published by Austrian and German newspapers in connection with a missing-persons case without the photographer's consent.⁵² The Court held that a photograph is protected by copyright if it demonstrates the author's intellectual creativity. This threshold is satisfied when the photographer exercises free and creative choices regarding composition, subject positioning, lighting, and framing.⁵³ This judgement is significant for two reasons. First, it affirms that the standard of intellectual creativity applies uniformly across all categories of works, including photographs, where opportunities for individual creative expression may be limited. Second, it establishes that even modest creative choices (such as those embedded in the textual, visual, and coded content of AI training data) are sufficient for copyright protection and therefore to engage the reproduction right upon copying. The consequence for AI training case is clear: creative choices embedded in literary texts, journalistic articles, source code, and

⁵² Case C-145/10 *Eva-Maria Painer v Standard VerlagsGmbH* [2011] ECR I-12533, paras 87–94.

⁵³ Case C-145/10 *Eva-Maria Painer v Standard VerlagsGmbH* [2011] ECR I-12533, para 92.

visual works within training datasets are highly likely to satisfy the originality threshold, and their reproduction during training accordingly triggers Article 2 of the InfoSoc Directive.⁵⁴

Pelham — C-476/17

In *Pelham GmbH v Ralf Hütter*, the Court examined whether the unauthorised use of a two-second rhythmic sequence from a phonogram constituted an infringement of the right of reproduction.⁵⁵ The Court determined that reproducing even a very short fragment of a phonogram constitutes reproduction within the meaning of the relevant exclusive right if the sample is incorporated into a new work in an unaltered and recognisable form.⁵⁶ Notably, the Court clarified that the right of reproduction does not permit a general exception for de minimis uses; in other words, there is no threshold below which a reproduced fragment is automatically exempt from protection solely due to its brevity.⁵⁷ At the same time, the Court acknowledged that purely technical or functional elements of a work, which do not reflect original creative expression, fall outside the scope of protection.⁵⁸ Applied to AI training, the *Pelham* decision supports the conclusion that the mass reproduction of short fragments, such as tokens, snippets, and fragments processed during model training, falls within the scope of the right of reproduction wherever these fragments contain original expression, regardless of their length.

Funke Medien — C-469/17

In *Funke Medien NRW GmbH v. Bundesrepublik Deutschland*, the Court addressed copyright issues concerning military reports prepared by the German Federal Government and published without authorisation by a news website.⁵⁹ The Court held that works of a purely factual or

⁵⁴ Martin Senftleben and Laurens Buijtelaa, ‘Robot Creativity: An Incentive-Based Neighbouring Rights Approach’ (2020) 42 EIPR 797, 799.

⁵⁵ Case C-476/17 *Pelham GmbH v Ralf Hütter* [2019] ECLI:EU:C:2019:624, para 14.

⁵⁶ Case C-476/17 *Pelham GmbH v Ralf Hütter* [2019] ECLI:EU:C:2019:624, para 29.

⁵⁷ Case C-476/17 *Pelham GmbH v Ralf Hütter* [2019] ECLI:EU:C:2019:624, para 30.

⁵⁸ Case C-476/17 *Pelham GmbH v Ralf Hütter* [2019] ECLI:EU:C:2019:624, para 31.

⁵⁹ Case C-469/17 *Funke Medien NRW GmbH v Bundesrepublik Deutschland* [2019] ECLI:EU:C:2019:623, paras 19–22.

official nature, where the author’s individual creative expression is absent due to the functional or informational character of the content, may not satisfy the originality requirement and accordingly fall outside the scope of copyright protection.⁶⁰ This judgement establishes an important limiting principle: not all content included in AI training datasets is protected by copyright. Where works primarily consist of factual information, standardised language, or content shaped by its informational function rather than creative choice, their reproduction during AI training does not infringe the reproduction right under Article 2 of the InfoSoc Directive. This clarification is practically significant, as large-scale training datasets often contain substantial factual content, such as news agency reports, government documents, and technical specifications, in addition to original literary and artistic works.⁶¹

Cofemel — C-683/17

In *Cofemel — Sociedade de Vestuário SA v G-Star Raw CV*, the Court affirmed that the standard of the author’s own intellectual creation establishes a single, uniform threshold of originality for all categories of works protected by EU copyright, excluding any additional or alternative national criteria.⁶² The Court further held that Member States may not impose supplementary requirements, such as aesthetic value or artistic purpose, as conditions for copyright protection that exceed the originality standard defined by the EU.⁶³ For the present analysis, the *Cofemel* decision serves a consolidating role by confirming that the originality standard derived from *Infopaq International* and applied in *Painer*, *Pelham*, and *Funke Medien* is not confined to specific categories of works, but applies uniformly to all protected subject matter. As a result, the determination of whether the reproduction of training data falls within the scope of Article 2 of

⁶⁰ Case C-469/17 *Funke Medien NRW GmbH v Bundesrepublik Deutschland* [2019] ECLI:EU:C:2019:623, para 25.

⁶¹ Carlini and others (n 7) 2633, 2635.

⁶² Case C-683/17 *Cofemel — Sociedade de Vestuário SA v G-Star Raw CV* [2019] ECLI:EU:C:2019:764, para 29.

⁶³ Case C-683/17 *Cofemel — Sociedade de Vestuário SA v G-Star Raw CV* [2019] ECLI:EU:C:2019:764, para 34.

the InfoSoc Directive is subject to a single, consistent legal standard, regardless of whether the works involved are literary texts, images, musical compositions, or software code.⁶⁴

3.2.5. The Operative Standard for AI Training

Based on the aforementioned case law, an operative legal standard can be formulated regarding whether the reproduction of works during AI training falls within the exclusive right established in Article 2 of the InfoSoc Directive. This standard involves three elements: (1) the work is author's own intellectual creation; (2) the notion of reproduction is uniform (applies to all categories of works) and EU-autonomous; (3) whether protection is granted is determined by a qualitative assessment of the originality of the reproduced expression rather than the quantity of material copied.⁶⁵

The standard is satisfied when a work or fragment demonstrates a free and creative choice by its author, resulting in an individual intellectual style that is distinct from purely mechanical or functional execution of a certain task.⁶⁶ Examples of content meeting this threshold include original literary expression, creative journalistic texts, individually composed source code, and artistic visual works. Such content is prevalent in the large-scale datasets used to train modern AI models, including web-scraped text corpora, digitised book collections, and image datasets. The reproduction of this content during the AI training process consequently triggers the right of reproduction under Article 2 of the InfoSoc Directive.

The standard is not satisfied when content consists exclusively of factual information, ideas, styles, methods, or common expressions that do not demonstrate individual creative choices. Raw data, statistical information, standardised technical language, and purely functional text are excluded from copyright protection under the defined EU standard and may accordingly be

⁶⁴ Matthias Leistner, 'Urheberrecht an der Schnittstelle zwischen Online-Plattformen und Nutzern' (2020) 122 GRUR 1, 5.

⁶⁵ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 37; Case C-683/17 *Cofemel — Sociedade de Vestuário SA v G-Star Raw CV* [2019] ECLI:EU:C:2019:764, para 29.

⁶⁶ Case C-145/10 *Eva-Maria Painer v Standard VerlagsGmbH* [2011] ECR I-12533, para 92.

reproduced without invoking Article 2 of the InfoSoc Directive.⁶⁷ This distinction is particularly relevant for AI training: when a training dataset contains unprotected factual content, reproducing it during training does not infringe the reproduction right. Nevertheless, in practice, the vast majority of large-scale training datasets are composed primarily of human-authored texts and other creative works that meet the originality threshold. Although the proportion of unprotected factual content is not negligible, it does not change the fundamental conclusion that current AI training practices engage the right of reproduction on a massive scale.⁶⁸

The following table summarises the application of the originality standard to the principal categories of content typically found in AI training datasets:

Content type	Originality assessment	Copyright status
Original literary and journalistic text	Reflects author's creative choices	Protected → Art. 2 engaged
Factual news reporting (purely informational)	Limited or no original expression	Partially or wholly unprotected
Raw data and statistics	No creative expression	Not protected
Source code with creative structure	Reflects programming choices	Protected → Art. 2 engaged
Standardised boilerplate language	No individual creative stamp	Not protected
Artistic images and photographs	Reflects creative choices (per <i>Painer</i>)	Protected → Art. 2 engaged

⁶⁷ Case C-469/17 *Funke Medien NRW GmbH v Bundesrepublik Deutschland* [2019] ECLI:EU:C:2019:623, para 25; Rosati (n 30) 156.

⁶⁸ Benjamin Sobel, 'Artificial Intelligence as a Threat to Copyright and How the Law Should Respond' (2023) 26 Stanford Technology Law Review 1, 18.

Official and governmental documents	Functional (no original expression)	Generally, not protected
-------------------------------------	-------------------------------------	--------------------------

The analysis demonstrates that reproducing works during AI training invokes the right of reproduction under Article 2 of the InfoSoc Directive whenever and to the extent that the reproduced content reflects the author’s own intellectual creation. Given the composition of large-scale datasets used in contemporary AI training, this condition is satisfied for the substantial majority of processed content. The lawfulness of such reproduction, which depends on the presence of a statutory exception or a contractual licence,⁶⁹ is discussed in Chapters 4 and 5. The following Section examines in detail how and at which stages of the AI training process the right of reproduction is triggered.

3.3. Application of the Findings to the AI Training Pipeline

As the CJEU has not yet issued a ruling directly addressing the copyright status of AI training activities, the following analysis applies the established EU copyright principles, particularly those derived from the *Infopaq International* case and the case law discussed in Section 3.2, to each technically distinct stage of the AI training process described in Chapter 2. Where the application of existing principles to the specific characteristics of AI training results in genuine legal uncertainty, such uncertainty is identified and assessed on its merits. The analysis proceeds in stages, consistent with the principle confirmed in the *Infopaq II* case, which requires that each individual step of a multi-stage technical process be assessed independently according to the applicable legal standard.⁷⁰

⁶⁹ Geiger, Frosio and Bulayenko (n 49) 174, 180.

⁷⁰ Case C-302/10 *Infopaq International A/S v Danske Dagblades Forening (Infopaq II)* [2012] ECLI:EU:C:2012:16, paras 27–30.

3.3.1. Web Scraping and Dataset Compilation

As outlined in Chapter 2, the initial stage of AI training involves the automated collection of content from online sources through web scraping. In this process, software tools send HTTP requests to web servers and download the returned content for storage and later processing.⁷¹ This stage is the first point at which the right of reproduction under Article 2 of the InfoSoc Directive is implicated, and the legal analysis at this point is relatively straightforward.

During web scraping, the content of a web page is downloaded, and a digital copy is created on the scraping system's storage infrastructure. This process constitutes a direct reproduction as defined by Article 2 of the InfoSoc Directive. The copy is created deliberately and in its entirety (the scraper downloads the full content of each target page rather than selectively extracting unprotected elements). If the downloaded content includes text, images, or other materials that reflect the author's intellectual creation, as established in *Infopaq International*, a right of reproduction is engaged.⁷² The subsequent storage of the downloaded content on servers, whether owned by the AI developer or managed by a third-party cloud provider, constitutes permanent reproduction: copies of the works remain indefinitely on storage media under the developer's effective control.⁷³ Such a form of permanent storage is distinct from the temporary technical copies addressed in Article 5(1) of the InfoSoc Directive. The applicability of Article 5(1) to AI training is examined separately in Section 3.4.

The scale at which web scraping is conducted for AI training amplifies the legal significance of each individual act of reproduction. Contemporary datasets for LLM training, such as Common Crawl, which forms the basis of many prominent training corpora, comprise hundreds of billions of individual documents sourced from the internet.⁷⁴ Each document that is downloaded and stored represents a distinct act of reproduction. Consequently, the aggregate effect results

⁷¹ See Chapter 2, section 2.2 of this thesis.

⁷² Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 37.

⁷³ InfoSoc Directive, art 2.

⁷⁴ Carlini and others (n 7) 2633, 2635.

in the reproduction of protected works on a scale without precedent in copyright law. While the scale does not change the legal classification of individual acts of reproduction, it is relevant to evaluating the seriousness of copyright implications and, as will be discussed in Chapter 6, to considerations regarding remedies and regulatory responses.

A comparison with search engine crawling, a technically analogous process generally regarded as lawful, clarifies the specific legal character of AI training scraping. Search engines systematically reproduce web pages to index content and direct users to original sources. In this context, reproduction functions as an intermediary mechanism that facilitates access to the original work, and the economic value generated by the search engine derives from this intermediary role rather than from the reproduced content itself.⁷⁵ By contrast, web scraping for AI training reproduces content not to redirect users to the source, but to extract and permanently embed the expressive and informational value of that content into a model that may subsequently compete with or replace the original works. The purpose, economic rationale, and impact on rightholders therefore differ fundamentally. This distinction is significant not only for analysing the reproduction right but also for determining whether web scraping activities may benefit from any implied authorisation, an issue addressed in Chapter 4 in relation to the requirement of lawful access.⁷⁶

3.3.2. Dataset Storage and Preprocessing

Following the initial collection of training data, AI developers apply a series of preprocessing operations to the compiled dataset before the training itself. These operations include deduplication, format conversion, quality filtering, and tokenisation. As outlined in Chapter 2, such

⁷⁵ Hugenholtz (n 35) 3, 5.

⁷⁶ Geiger, Frosio and Bulayenko (n 49) 174, 181.

processes transform the raw scraped content into a format suitable for the model’s training architecture.⁷⁷ Each of these operations raises distinct concerns under Article 2 of the InfoSoc Directive.

Storing a compiled dataset in its original form, prior to any pre-processing, constitutes a permanent reproduction of the works it contains and represents the most legally unambiguous action within the entire training pipeline. In contrast to temporary technical copies created during network transmission or processing, a stored copy of the dataset remains indefinitely on physical or virtual storage media.⁷⁸ This copy is accessible for repeated use throughout the learning process and subsequent activities. The indefinite retention of such copies therefore fully satisfies the “permanent” requirement of the Article 2 of the InfoSoc Directive and precludes any credible arguments to the contrary.

Format conversion operations, such as extracting plain text from HTML documents, converting PDF files into machine-readable text, or normalising character encodings, result in the creation of a transformed copy of the original content. Although such a copy differs in form, it still preserves the expressive content of the work in a new format. Article 2 of the InfoSoc Directive explicitly covers reproduction “by any means and in any form”, a formulation sufficiently broad to encompass format conversion.⁷⁹ Consequently, the conversion process constitutes a further act of reproduction in addition to the initial storage, creating a derivative copy that also engages the right of reproduction, provided the source content satisfies the originality threshold.

The legal definition of tokenisation is complex and requires separate consideration. Tokenisation is the process of decomposing text into discrete numerical units, known as tokens, which serve as the primary input elements for neural network processing. The procedure converts a natural language expression into a sequence of integers or vectors, ensuring that the original

⁷⁷ See Chapter 2, section 2.2 of this thesis.

⁷⁸ InfoSoc Directive, art 2.

⁷⁹ InfoSoc Directive, art 2.

text is no longer present in the tokenised representation in any human-readable form.⁸⁰ This raises the question of whether such a transformation constitutes “reproduction in any form” within the meaning of Article 2 of the InfoSoc Directive, or whether the level of abstraction involved is sufficient to exclude the tokenised representation from the scope of the right of reproduction. Two positions may be advanced.

A conservative interpretation holds that tokenisation constitutes reproduction within the meaning of Article 2 of the InfoSoc Directive. Although tokenised representations are encoded numerically, they preserve a complete and reversible mapping to the original textual expression. Given a specific tokenisation scheme, the original text can be accurately reconstructed from the token sequence.⁸¹ Consequently, the expressive content of the work remains fully present in tokenised form, though in a transformed encoding. Applying the broad formulation of Article 2 of the InfoSoc Directive, which covers reproduction “by any means and in any form”, the tokenised representation is best understood as a reproduction of the original work in a different form, similar to a translation or transcription into another system of notation. Under this interpretation, tokenisation constitutes a further reproduction act engaging Article 2 of the InfoSoc Directive.

A more restrictive interpretation suggests that tokenisation entails a degree of transformation that places the resulting representation outside the scope of the right of reproduction. If copyright protects only the original expression, rather than the underlying information or ideas, then purely numerical encoding, which fully abstracts from the linguistic form of the work, may cross the threshold at which the protected expression is no longer identifiable.⁸² This argument is persuasive when the tokenisation scheme is not publicly available, making reconstruction of

⁸⁰ See Chapter 2, section 2.2 of this thesis.

⁸¹ Henderson and others (n 15) 1, 8.

⁸² Sobel (n 68) 1, 22.

the original text practically impossible. Nevertheless, this position encounters significant challenges: the CJEU has not recognised any principle of “transformative reproduction” that would exclude technically transformed copies from the scope of Article 2 of the InfoSoc Directive, and the broad language of that provision militates against such implied limitations.

Deduplication, defined as the removal of duplicate documents from the training corpus, does not affect the copyright analysis of the remaining copies. While the elimination of redundant copies reduces the total number of reproduction acts, it does not retroactively legalise the original reproductions and does not alter the copyright status of the deduplicated dataset, which still comprises permanently stored copies of third-party works.⁸³ Likewise, utilising third-party cloud storage infrastructure to host training datasets does not change the copyright analysis from the AI developer’s perspective. Reproduction occurs on infrastructure under the developer’s effective control and for the developer’s commercial purposes, irrespective of the identity of the infrastructure operator. Potential issues of secondary liability for cloud storage providers are beyond the scope of the present analysis.

3.3.3. The Training Process

The training process, as described in Chapter 2, involves the repeated loading of data batches into processing memory, the execution of forward and backward propagation passes, and the iterative adjustment of model weights. This stage raises the most legally complex issues of reproduction within the entire pipeline.⁸⁴ In contrast to the web scraping and storage stages, which involve explicit and permanent acts of copying, the training process entails reproductions that are technically temporary and whose legal classification under Article 2 of the InfoSoc Directive is genuinely questionable.

⁸³ Dreier and Hugenholtz (n 27) 342.

⁸⁴ See Chapter 2, section 2.3 of this thesis.

During each iteration of the training process, a batch of training data is loaded from storage into the processing memory, specifically the graphics processing unit (GPU) or central processing unit (CPU) RAM, of the training system. This operation creates a temporary copy of the batch content in volatile memory. Such a copy is technically considered temporary in the sense that it is overwritten or cleared at the end of each processing cycle. Nevertheless, the legal significance of this temporariness requires careful consideration. As established in Section 3.2, Article 2 of the InfoSoc Directive explicitly addresses temporary reproduction: the provision's reference to both "temporary" and "permanent" reproduction precludes arguments that the temporary nature of in-memory copies would, in principle, exclude them from the scope of the reproduction right.⁸⁵ However, the temporary nature of each individual in-memory copy is directly relevant to the question of whether the exception provided in Article 5(1) of the InfoSoc Directive applies, an exception which, as demonstrated in Section 3.4, is not available for AI training on its own merits.

The primary argument advanced by AI developers and certain commentators against characterising the training process as reproduction under copyright law focuses on the nature of the output produced, rather than on the existence of copies in memory. They claim that the training process does not constitute copying in any legally relevant sense, as the model does not retain or preserve the expressive content of the training materials. Instead, it extracts statistical patterns, distributional relationships between tokens, and abstract representations of language or visual structure, none of which resemble any identifiable original expression.⁸⁶ From this standpoint, training an AI model is likened to a person learning from a book: the individual absorbs ideas, knowledge, and skills from the work without reproducing its expression. The same, as is argued, applies to a model trained on such material.

⁸⁵ InfoSoc Directive, art 2; Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 37.

⁸⁶ Sobel (n 68) 1, 19.

While this argument may initially appear persuasive, it conflates two distinct issues: the information retained by the model after training and the acts of reproduction that occur during training. According to Article 2 of the InfoSoc Directive, the right of reproduction is triggered at the precise moment copying occurs, regardless of the subsequent use or disposal of the copy.⁸⁷ The subsequent deletion of training data packets from volatile memory does not affect the legal characterisation of the reproduction that takes place when these batches are loaded. To illustrate, photocopying a protected work constitutes reproduction under Article 2 of the InfoSoc Directive, even if the copies are later destroyed. The copyright-relevant event is the creation of the copy, not its retention. Applied to AI training, loading the protected content into processing memory constitutes reproduction at the moment of loading. The question of what the model ultimately encodes in itself is distinct and does not retroactively negate the previously committed act of reproduction.⁸⁸

A further question arises in relation to gradient computation — the mathematical operation through which the model calculates the adjustments to be made to its parameters during each training iteration. Gradients, which are mathematical derivatives computed from the current parameter values and the training dataset, indicate the direction and magnitude of parameter adjustments necessary to enhance the predictive performance of a model. The central question is whether computation of gradients constitutes a legally significant act of reproduction under Article 2 of the InfoSoc Directive. It is to be argued that it does not, at least not independently. Gradients are mathematical abstractions, specifically tensors of numerical values, that do not correspond to or encode any identifiable expressive element of the training data.⁸⁹ Their computation does not result in a copy of any protected expression, but instead produces a set of mathematical instructions derived from the relationship between the model’s current state and

⁸⁷ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 37; Rosati (n 30) 80.

⁸⁸ Leistner (n 64) 1, 7.

⁸⁹ Henderson and others (n 15) 1, 12.

the training data. Since copyright does not extend to mathematical operations or their results as such, the gradient calculation stage is likely not subject to the right of reproduction. This acknowledgement reflects the necessity for intellectual rigour: not every aspect of the training process constitutes reproduction, and the limitations of copyright analysis must be recognised.

3.3.4. The Memorization Problem of Model Weights

The most legally complex and analytically original issue in this Chapter concerns the copyright status of the trained model itself. Specifically, whether the numerical parameters, or weights, of a trained AI model constitute a reproduction of the works on which the model was trained, as defined by Article 2 of the InfoSoc Directive. This question has not yet been addressed by the CJEU and remains a topic of ongoing debate within the EU copyright scholarship.⁹⁰

As outlined in Chapter 2, model weights are numerical parameters of a neural network that are iteratively adjusted during training to minimise the model's prediction error. Upon completion of training, these weights encode the model's learned representation of statistical relationships present in the training data. The weights do not contain the original training text or images in any human-readable form; the trained model does not function as a database of training documents and cannot directly retrieve them. Nevertheless, empirical research has demonstrated that LLMs can reproduce verbatim fragments from their training data under specific conditions, a phenomenon referred to as memorisation.⁹¹ The legal implications of this phenomenon are substantial and require careful analysis.

The first analytical position asserts that model weights do not constitute a reproduction of the training works within the meaning of Article 2 of the InfoSoc Directive. Under this interpretation, the weights are considered purely mathematical abstractions, specifically matrices of floating-point numbers, which do not encode or store any identifiable expressive elements of the

⁹⁰ Senftleben and Buijelaar (n 54) 797, 801.

⁹¹ Carlini and others (n 7) 2633, 2640.

training data in a manner that copyright law would recognise as reproduction. Copyright protection applies to the expression of intellectual creativity, rather than to the underlying information, ideas, or knowledge conveyed by that expression. If training causes the model to learn only the statistical properties of language without encoding any specific protected expression, the resulting weights fall outside the scope of the reproduction right. This is analogous to a human author who, after reading and being influenced by a protected work, does not infringe the copyright of the works that have shaped their knowledge and style.⁹² This perspective is further supported by general principles of statistical learning theory, which characterise model training as a process of generalisation rather than memorisation.

The second position acknowledges that the model's weights may not themselves constitute a reproduction, but instead identifies the copyright-relevant act at the inference stage, when the trained model generates output in response to user prompts. If the model produces output that reproduces, verbatim or nearly verbatim, passages from the training works, this act constitutes a new reproduction within the meaning of Article 2 of the InfoSoc Directive, distinct from and supplementary to the acts of reproduction that occur during training.⁹³ This perspective aligns with the observed phenomenon of memorisation, as the model reproduces protected expression in its output data rather than in its weights. However, this view entails a significant implication: if reproduction occurs at the output stage rather than in the weights, the logical consequence is that the training process itself may be lawful, provided it falls under the relevant exception, while the deployment of the model remains potentially infringing whenever memorised content is generated. The question of infringement of rights in the output is beyond the scope of this thesis yet is noted to ensure analytical completeness.⁹⁴

⁹² Sobel (n 68) 1, 20.

⁹³ Henderson and others (n 15) 1, 15.

⁹⁴ Geiger, Frosio and Bulayenko (n 49) 174, 182.

The third and most legally robust position holds that model weights may themselves constitute a reproduction of protected expression in a functional sense, provided that the degree of memorisation embedded in those weights is sufficient to demonstrate encoding of the protected expression in a non-random and non-generalised form. The central argument is that copyright law does not require a reproduction to be visibly identical to the original; rather, it requires that the reproduced material reflect the author’s intellectual creation.⁹⁵ To illustrate, a musical score is considered a reproduction of a musical composition even if it does not acoustically resemble the music, and a computer program is a reproduction of its source code even when compiled into binary machine instructions that are unreadable to humans. By analogy, model weights that functionally encode the expressive content of training works, as evidenced by the capacity to generate verbatim reproductions of those materials, may constitute a reproduction in a non-standard form, thereby satisfying the requirement of reproduction “by any means and in any form” as stated in Article 2 of the InfoSoc Directive.⁹⁶

A key weakness of the third position is the challenge of establishing, within current technical and legal frameworks, a precise threshold at which functional encoding transitions from general statistical representation to copyright-relevant reproduction of specific protected expression. The memorisation studies cited above indicate that this threshold is crossed in certain cases, particularly with larger models trained on smaller datasets. However, existing law lacks a mechanism for determining when the degree of memorisation in a model is sufficient to constitute reproduction.⁹⁷ This gap presents a significant obstacle to the application of current copyright doctrine to AI systems and constitutes a compelling argument for legislative clarification, as proposed in Chapter 6.

⁹⁵ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 37; InfoSoc Directive, art 2.

⁹⁶ Leistner (n 64) 1, 9.

⁹⁷ Carlini and others (n 7) 2633, 2645.

The issue of whether the weights of a trained model constitute a reproduction of the training works under Article 2 of the InfoSoc Directive remains unresolved in EU copyright law, and this paper does not provide a definitive answer. Nevertheless, the resolution of this question does not impact the central conclusion of the present Chapter. Reproduction as defined by Article 2 of the InfoSoc Directive clearly occurs during web scraping and dataset storage stages in the training pipeline, where complete and permanent copies of original works are created and stored. These acts of reproduction alone are sufficient, regardless of the status of weights, to trigger the right of reproduction and require that AI training be conducted either under a contractual licence or within a statutory exception. Whether weights constitute an additional act of reproduction introduces a separate and analytically distinct level of potential liability for copyright infringement. This issue is particularly relevant to claims concerning infringement of rights in works and to ongoing legal proceedings in various jurisdictions, but it is not required to conclude that current AI training practices infringe the reproduction right under EU law.⁹⁸

3.4. The Temporary Copying Exception — Article 5(1) InfoSoc Directive

3.4.1. Structure and Purpose of Article 5(1)

Section 3.3 established that the acts underlying AI training infringe the reproduction right under Article 2 of the InfoSoc Directive at multiple stages of the training process. It is therefore necessary to assess whether any of these acts of reproduction may be exempted from the exclusive right through the temporary reproduction exception provided in Article 5(1) of the same Directive. This exception constitutes the primary alternative legal basis available to AI developers seeking to avoid the requirement for a licence or a statutory exception under the DSM Directive. The present Section demonstrates that Article 5(1) of the InfoSoc Directive is inapplicable to

⁹⁸ Senftleben and Buijelaar (n 54) 797, 803.

AI training and that the TDM exceptions, discussed in Chapter 4, represent the only plausible legal basis for lawful AI training under EU copyright law.

Article 5(1) of the InfoSoc Directive stipulates that temporary acts of reproduction, which are transient or incidental, constitute an integral and essential part of a technological process, and are undertaken solely to enable the lawful use of a work or its lawful communication to third parties via an intermediary, and which lack independent economic significance, are exempt from the reproduction right established in Article 2 of the same Directive.⁹⁹ This provision is mandatory; unlike most exceptions in Article 5, which Member States may apply at their discretion, Article 5(1) must be incorporated into the national legislation of all Member States.¹⁰⁰ The purpose, as outlined in Recital 33 of the InfoSoc Directive, is to permit technically necessary temporary reproductions inherent to digital network processes, such as temporary copies created in browser caches, network transmission buffers, and proxy servers in the ordinary course of internet communication.¹⁰¹ The exception was intended to ensure that routine technical processes facilitating lawful digital communication are not accidentally subject to the reproduction right, rather than to establish a broad exception for copying operations with commercial significance.

The CJEU has consistently upheld a strict interpretation of Article 5(1) of the InfoSoc Directive. In the *Infopaq International* case, the Court determined that the conditions of Article 5(1) require strict interpretation, as this provision constitutes a derogation from the general principle set out by the Directive that rightholders must authorise any reproduction of their protected works.¹⁰² The principle of strict interpretation is not merely a formal methodological guideline;

⁹⁹ InfoSoc Directive, art 5(1).

¹⁰⁰ InfoSoc Directive, art 5(1); see also recital 33.

¹⁰¹ InfoSoc Directive, recital 33.

¹⁰² Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 56.

it has significant implications for the following analysis, as it precludes the extension of the exception's scope beyond what is expressly provided by its text and purpose.

A key structural feature of Article 5(1) of the InfoSoc Directive is that it functions as an enabling, rather than a primary, exception. It permits incidental reproductions that facilitate uses already lawful in their own right, rather than establishing a legal basis for uses that would otherwise infringe copyright. The appropriate analytical sequence is to first identify an independent legal basis for the primary use, and then apply Article 5(1) to any residual technical copies that are unavoidably created by the lawful primary use. Where no independent legal basis for the primary use exists, Article 5(1) cannot serve as such a basis. The implications of this structural feature for AI training, particularly the circularity problem that arises when Article 5(1) is invoked as the sole legal basis for training activities, are addressed in the analysis of Condition 4 below.¹⁰³

3.4.2. The Five Conditions — Cumulative Requirement

In the *Infopaq International* case, the CJEU confirmed that the five conditions set out in Article 5(1) of the InfoSoc Directive are cumulative; all five must be met simultaneously for the exception to apply.¹⁰⁴ Failure to meet any single condition is sufficient to render the exception inapplicable. The following analysis evaluates each condition in turn as applied to AI training.

Condition 1 — The Act Must Be Temporary

Article 5(1) of the InfoSoc Directive applies exclusively to acts of communication to the public that are temporary in nature. In *Football Association Premier League Ltd v QC Leisure*, the Court determined that a communication qualifies as temporary under Article 5(1) if it is limited in duration and is automatically deleted upon completion of its intended process.¹⁰⁵ When this

¹⁰³ Dreier and Hugenholtz (n 27) 355.

¹⁰⁴ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 54.

¹⁰⁵ Joined Cases C-403/08 and C-429/08 *Football Association Premier League Ltd v QC Leisure* [2011] ECR I-9083, para 164.

standard is applied to AI training, the principal acts of reproduction identified in Section 3.3 fail to satisfy the requirement of transience at two distinct stages.

Copies of the dataset, which constitute a permanent repository of scraped content on servers, are retained indefinitely throughout the training procedure and frequently beyond its conclusion. Such copies are not automatically deleted upon the completion of any specific process; rather, they are preserved as valuable commercial assets for repeated use. In this context, the permanence of these copies precludes any claim of temporariness under any reasonable interpretation of Article 5(1) of the InfoSoc Directive.¹⁰⁶ Similarly, trained model weights, which are the principal output of the training process, are also permanent. They represent the enduring commercial outcome of the training activity and are continuously deployed, distributed, and utilised for commercial purposes. The technical deletion of copies in RAM during individual iterations of the training batch after each processing cycle does not make the overall activity compliant: as established in *Infopaq II*, each step of a multi-stage process must independently satisfy the requirements of Article 5(1). The storage and weighting stages do not meet Condition 1, regardless of the temporary nature of the copies in RAM.¹⁰⁷

Condition 2 — The Act Must Be Transient or Incidental

The present Condition requires that reproduction must be either temporary, meaning it exists only for the duration of the technological process and is automatically deleted upon completion, or incidental, meaning it is a by-product of the process rather than its purpose.¹⁰⁸ This condition is the most decisive among the five in the context of AI training. Failure to meet this requirement alone is sufficient to defeat the exception.

¹⁰⁶ InfoSoc Directive, art 5(1); Case C-302/10 *Infopaq International A/S v Danske Dagblades Forening* (*Infopaq II*) [2012] ECLI:EU:C:2012:16, para 29.

¹⁰⁷ Case C-302/10 *Infopaq International A/S v Danske Dagblades Forening* (*Infopaq II*) [2012] ECLI:EU:C:2012:16, paras 27–30.

¹⁰⁸ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 64.

The processes of compiling and storing datasets for AI training are neither transient nor incidental. In contrast, they are intentional, systematic, and conducted on a large scale, with the explicit purpose of establishing a permanent repository of training content. Unlike incidental by-products, a dataset serves as the primary input for the training process and is actively processed, deduplicated, filtered, and maintained as a valuable commercial asset.¹⁰⁹ The deliberate and large-scale nature of dataset compilation sharply contrasts with the incidental copying addressed by Article 5(1) of the InfoSoc Directive, which occurs, for instance, in the case of automatic browser caching or temporary buffer copies created via network transmission. In these typical cases, copies are unavoidable side effects of technically necessary operations, whereas dataset compilation for AI training involves copies as the central and intentional aim of a commercial activity. This clear distinction supports the conclusion that Condition 2 is not satisfied, leaving little room for doubt.¹¹⁰

Condition 3 — Integral and Essential Part of a Technological Process

Article 5(1) of the InfoSoc Directive stipulates that reproduction must constitute an integral and essential component of the technological process. In *ITV Broadcasting Ltd v TV Catchup Ltd*, the Court clarified that a technological process may be commercially motivated and still retain the protection of Article 5(1), provided all other conditions are satisfied.¹¹¹ Applying this standard to AI training, the reproduction of training data is mechanically inseparable from the training process: without loading and processing copies of the training content, the model cannot learn, and the process cannot proceed. Reproduction is not optional or ancillary; it forms the

¹⁰⁹ Carlini and others (n 7) 2633, 2635.

¹¹⁰ Geiger, Frosio and Bulayenko (n 49) 174, 183.

¹¹¹ Case C-607/11 *ITV Broadcasting Ltd v TV Catchup Ltd* [2013] ECLI:EU:C:2013:147, para 26.

technical foundation of the entire operation.¹¹² Therefore, Condition 3 is fulfilled. This is, however, the sole condition satisfied by AI training, and its fulfilment alone is insufficient for the exception to apply, given the cumulative nature of the five requirements.

Condition 4 — Sole Purpose of Enabling Transmission or Lawful Use

Article 5(1) of the InfoSoc Directive requires that the sole purpose of reproduction must be either to facilitate online transmission between third parties via an intermediary or to enable the lawful use of the work.¹¹³ These two options are mutually exclusive; in principle, fulfilling either one would suffice. Both requirements, however, share a common defining element relevant to the analysis of AI training: the reproduction must serve a single, clearly defined purpose that falls within one of the two specified categories. AI training does not satisfy this condition in either respect.

Regarding the first scenario, which involves facilitating the transfer of content between third parties via an intermediary, AI training is plainly not an act of network intermediation. The developer reproduces training content solely for their own commercial objectives, rather than for enabling the content transfer between other parties. Transmission to third parties is neither provided for, nor intended, nor executed. Therefore, the first scenario is inapplicable.¹¹⁴

When applying the second scenario, which concerns facilitating the lawful use of a work, AI training encounters a structural issue of analytical importance: the circularity problem. Article 5(1) of the InfoSoc Directive stipulates that the use enabled by reproduction must be independently lawful, based on a legal foundation separate from Article 5(1) itself. This provision supports uses that are already lawful and does not legitimise uses that would otherwise infringe

¹¹² Henderson and others (n 15) 1, 10.

¹¹³ InfoSoc Directive, art 5(1)(a)–(b).

¹¹⁴ Dreier and Hugenholtz (n 27) 356.

rights.¹¹⁵ In the absence of such an independent basis, Article 5(1) cannot serve as a legal foundation without resulting in circular reasoning: reproduction is said to enable lawful use, yet the only candidate for that lawful use is the training process itself, the legality of which is under dispute.¹¹⁶ As this analysis assumes that no relevant licence or TDM exception has been established (discussed in Chapter 4), Article 5(1) cannot serve as a standalone legal basis for AI training. Therefore, the second scenario is also inapplicable.¹¹⁷

Even if an independent legal basis for training were established, the single-purpose requirement would persist as a distinct barrier. Reproduction activities related to AI training fulfil multiple concurrent commercial objectives: producing a ready-to-use AI product, extracting statistical and semantic value from training content, creating model weights as independent commercial assets, and generating outputs that may substitute for the original works.¹¹⁸ This multiplicity of objectives is fundamentally incompatible with the (general) single-purpose requirement. Reproduction that simultaneously serves several commercial purposes cannot be regarded as serving a single purpose within the meaning of Article 5(1), irrespective of which of the alternative purposes is cited.¹¹⁹

Condition 5 — No Independent Economic Significance

¹¹⁵ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 56.

¹¹⁶ Leistner (n 64) 1, 8.

¹¹⁷ Dreier and Hugenholtz (n 27) 358.

¹¹⁸ Sobel (n 68) 1, 24.

¹¹⁹ Geiger, Frosio and Bulayenko (n 49) 174, 183.

The fifth and final condition requires that reproduction have no independent economic significance; in other words, it should not generate economic value beyond that produced by the legitimate technological process it supports.¹²⁰ In the *Infopaq International* case, the Court established that even minimal independent economic value is sufficient to fail this condition, a standard that was rigorously enforced in the *Football Association Premier League Ltd v QC Leisure* case.¹²¹ AI training fails to satisfy this requirement in the most plain manner. The reproduction activities involved in AI training, including the scraping, storage, and processing of training data, are not incidental to the economic value of the AI model; rather, they are the source of it. The entire commercial value of a trained LLM arises directly and exclusively from the reproduction and processing of training data. A model trained without data possesses no commercial value, whereas a model trained on high-quality data at scale may be valued at billions of euros. Therefore, the independent economic significance of reproductions in the training process is not only present but forms the entirety of the value-creation process.¹²²

3.4.3. Conclusion on Article 5(1) of the InfoSoc Directive

The application of Article 5(1) of the InfoSoc Directive to AI training activities is summarised in the following table:

Condition	Result	Basis
Temporary	Fails	Dataset storage and model weights are permanent
Transient or incidental	Fails	Dataset compilation is intentional, systematic, and large-scale

¹²⁰ InfoSoc Directive, art 5(1).

¹²¹ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 69; Joined Cases C-403/08 and C-429/08 *Football Association Premier League Ltd v QC Leisure* [2011] ECR I-9083, para 175.

¹²² Senftleben and Buijtelaar (n 54) 797, 802.

Integral and essential part of technological process	Satisfied	Reproduction is mechanically inseparable from training
Sole purpose — transmission or lawful use	Fails	Multiple (commercial) purposes; circularity in lawful use sub-condition
No independent economic significance	Fails	Training reproductions constitute the entire source of model value

Article 5(1) of the InfoSoc Directive does not apply to acts of reproduction associated with AI training. Four of the five conditions are not satisfied; the only condition that is fulfilled, namely that the reproduction constitutes an integral and essential part of the technological process, is insufficient to invoke the exception due to the cumulative nature of the requirements, as confirmed by the Court in the *Infopaq International* case. Additionally, the structural character of Article 5(1) as an enabling rather than a primary exception precludes its use as a standalone legal basis for AI training activities without an independently established legitimate primary use.¹²³

This conclusion has significant structural implications for the overall analysis presented in this thesis. As Article 5(1) of the InfoSoc Directive does not apply to AI training, and no other exceptions within that Directive have been identified as relevant to the types of activities under consideration, the exceptions for text and data analysis introduced by Articles 3 and 4 of the DSM Directive represent the sole legislative basis under EU copyright law permitting AI training without a contractual licence. The scope, conditions, and limitations of these exceptions are examined in Chapter 4.¹²⁴

¹²³ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 54.

¹²⁴ DSM Directive, arts 3–4.

3.5. Answer to Research Question 1

This Section directly addresses the first research question of the thesis: whether the training of AI models constitutes copyright-relevant reproduction under EU law.

The answer is affirmative. The training of AI models constitutes a reproduction subject to copyright protection under Article 2 of the InfoSoc Directive at multiple stages of the training process. The originality standard established by the CJEU in the *Infopaq International* case, and reaffirmed in subsequent cases such as *Painer*, *Pelham*, *Funke Medien*, and *Cofemel*, indicates that acts such as web scraping, dataset storage, and iterative batch processing in AI training infringe the reproduction right whenever the processed content reflects the author's own intellectual creation.¹²⁵ Given that large-scale datasets used in contemporary AI training predominantly consist of literary, journalistic, visual, and coded content created by humans, this condition is satisfied for a significant portion of the material processed. The observation that purely factual content, standardised language, and functional elements lacking individual creative choices fall outside the threshold of originality does not undermine this conclusion; rather, it delineates the boundaries of the right without negating its application to the core activities of AI training.¹²⁶

The exception for temporary copying under Article 5(1) of the InfoSoc Directive does not affect this conclusion. As demonstrated in Section 3.4, AI training fails to satisfy four of the five cumulative conditions necessary for the exception to apply. The most significant unmet requirements are that the reproduction must be temporary or incidental and must have no independent economic significance. Furthermore, the structural design of Article 5(1) as an enabling rather

¹²⁵ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 37; Case C-145/10 *Eva-Maria Painer v Standard VerlagsGmbH* [2011] ECR I-12533, para 92; Case C-476/17 *Pelham GmbH v Ralf Hütter* [2019] ECLI:EU:C:2019:624, para 29; Case C-469/17 *Funke Medien NRW GmbH v Bundesrepublik Deutschland* [2019] ECLI:EU:C:2019:623, para 25; Case C-683/17 *Cofemel — Sociedade de Vestuário SA v G-Star Raw CV* [2019] ECLI:EU:C:2019:764, para 29.

¹²⁶ Geiger, Frosio and Bulayenko (n 49) 174, 180.

than a primary exception prevents its use as an independent legal basis for training activities conducted without a separately established lawful basis.¹²⁷ No other exception under the InfoSoc Directive appears applicable to the AI training activities under consideration.

It follows that AI training, as currently conducted by commercial developers and involving the large-scale reproduction of internet content to create commercially viable AI systems, requires either a contractual licence from rightholders or a statutory exception under EU copyright law. The InfoSoc Directive does not provide an exception for activities of this nature or scale. The only statutory exceptions under EU copyright law that may offer a legal basis for AI training are those for text and data mining, as introduced by Articles 3 and 4 of the DSM Directive.¹²⁸ The scope, conditions, limitations, and practical adequacy of these exceptions, particularly regarding their applicability to large-scale commercial AI training, constitute the subject of the following Chapter.

¹²⁷ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 54; Dreier and Hugenholtz (n 27) 355.

¹²⁸ DSM Directive, arts 3–4.

4. Text and Data Mining Exceptions under the DSM Directive

4.1. Introduction to TDM Exceptions

The conclusion reached in Chapter 3 is that AI training constitutes copyright-relevant reproduction under Article 2 of the InfoSoc Directive, and that this Directive does not provide exceptions that could justify such use. Consequently, the lawfulness of AI training under EU copyright law depends entirely on the existence of a statutory exception in other legislation or on a contractual licence. As demonstrated in Section 3.5, the only statutory exceptions that potentially could provide a legal basis are those for text and data mining introduced by Articles 3 and 4 of the DSM Directive.¹²⁹ Before these exceptions can be applied to the specific context of AI training, it is necessary to establish the legislative context in which they were adopted, the policy objectives they were intended to serve, their relationship with the pre-existing EU copyright system, and the definition of the scope of text and data mining. These foundational issues are addressed in this Section.

4.1.1. Legislative History

The exceptions to the DSM Directive concerning TDM resulted from a carefully negotiated compromise among competing interests. This historical context directly informs the interpretation and application of these provisions in relation to AI training.

The legislative process that produced the DSM Directive was initiated with an official proposal from the European Commission in September 2016.¹³⁰ Its original form included a single TDM exemption, restricted to research organisations conducting data analytics for scientific purposes. Notably, the proposal did not provide a general exemption for commercial data analytics. The Commission justified this approach by identifying a research competitiveness gap, noting

¹²⁹ DSM Directive, arts 3–4.

¹³⁰ European Commission, ‘Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market’ COM(2016) 593 final, 14 September 2016.

that EU researchers faced legal uncertainty about the permissibility of TDM under existing EU copyright law. In contrast, researchers in the United States benefited from the flexibility of the fair use doctrine under Section 107 of Title 17 of the US Code¹³¹. The accompanying impact assessment characterised this uncertainty as a structural disadvantage for EU research institutions and recommended a targeted mandatory exception to mitigate it.¹³² The original proposal was therefore designed to advance research rather than to serve as a broad authorisation for commercial data mining activities.

The legislative process in the European Parliament substantially broadened the Commission's initial proposal. The Committee on Legal Affairs introduced amendments that resulted in a second exemption for information technology, extending the personal scope.¹³³ Parliamentary debates involved significant lobbying from opposing interests: technology companies and data-intensive industries advocated for a broad exception without an opt-out mechanism, while publishers, authors, and collecting societies supported a narrow exception limited to research or, alternatively, a robust opt-out mechanism to maintain rightholders' control over the commercial use of their works.¹³⁴ The final version of the DSM Directive represents a compromise: Article 3 maintains the mandatory exception for research, and Article 4 establishes a new general exception with broad personal scope, applicable to any natural or legal person, accompanied by a rightholder opt-out mechanism.¹³⁵ The opt-out mechanism was not a supplementary addition to Article 4; rather, this concession to rightholders was essential for achieving political consensus on the broader personal scope of Article 4.

¹³¹ 17 USC § 107.

¹³² European Commission, 'Impact Assessment on the Modernisation of EU Copyright Rules' SWD(2016) 301 final, 14 September 2016, pt 2, 153.

¹³³ European Parliament, Committee on Legal Affairs, 'Report on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market' (A8-0245/2018, 2018).

¹³⁴ Geiger, Frosio and Bulayenko (n 49) 174, 176.

¹³⁵ DSM Directive, arts 3–4.

The DSM Directive entered into force on 7 June 2019, and Member States were required to transpose its provisions into national law by 7 June 2021.¹³⁶ Two aspects of the legislative history are directly relevant to the analysis of AI training in this Chapter. First, the legislative documentation does not provide substantive discussion of large-scale commercial AI training as a use case for TDM exceptions. The examples and scenarios considered during the legislative process focused on academic research, journalism, business analytics, and market analysis, which are activities conducted at a scale and with a commercial profile distinct from the training of foundation models on hundreds of billions of tokens. Second, the compromise structure of Articles 3 and 4, which combines a narrow mandatory research exception with a broad optional exception subject to opt-out, reflects a deliberate legislative calibration that was not designed with foundation model training in mind. Both observations have significant implications for the interpretation and critical assessment of the subsequent exemptions.¹³⁷

4.1.2. Policy Rationale

The TDM exemptions are supported by economic and research policy rationale, as articulated in the preamble to the DSM Directive. These justifications are essential for comprehending both the scope and limitations of the given legislative framework.

Recital 8 of the DSM Directive defines TDM as a technology with significant potential for research and innovation, emphasising its capacity to process large volumes of data and generate new knowledge. This capability provides substantial benefits for scientific research and the broader knowledge economy.¹³⁸ Recital 9 identifies the primary legislative challenge as the legal uncertainty surrounding the permissibility of TDM under existing EU copyright law. It is acknowledged that, even when reproduction for TDM purposes may be considered lawful in

¹³⁶ DSM Directive, art 29.

¹³⁷ Thomas Margoni and Martin Kretschmer, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (CREATe Working Paper 2022/1, University of Glasgow 2022) 4.

¹³⁸ DSM Directive, recital 8.

principle, the lack of an explicit legal basis discourages research and innovation involving large datasets.¹³⁹ This uncertainty was particularly pronounced in the EU, due to the absence of a flexible general exception similar to the US fair use doctrine, which enabled US courts and researchers to conduct TDM-related activities with greater legal certainty.

The broader policy context of the TDM exceptions is rooted in the EU's data economy strategy. The Commission's 2018 Communication on Artificial Intelligence identified access to data, including training data for AI systems, as a critical factor in enhancing European AI competitiveness.¹⁴⁰ The TDM exceptions were designed as part of the legal infrastructure to implement this strategy. By reducing transaction costs associated with obtaining individual licences for data-intensive research and analytics, these exceptions aim to lower barriers to entry for EU researchers and innovators engaged in data-driven activities. The economic rationale is to facilitate the market by reducing friction in the data economy, rather than to enable expropriation of rights.¹⁴¹

However, it should be noted that the political rationale articulated in the preamble to the DSM Directive and in the Commission's supporting documentation emphasises research institutions, analytics for small and medium-sized enterprises, and data-driven journalism, rather than the training of advanced AI models at commercial scale. Recital 18 of the DSM Directive, which addresses the commercial TDM exception under Article 4, specifically permits uses such as web scraping and other legitimate activities of a limited scope.¹⁴² The policy framework was therefore designed for data processing at a scale distinct from that required for the modern training of foundational models. This calibration gap between the original policy context and

¹³⁹ DSM Directive, recital 9.

¹⁴⁰ European Commission, 'Communication on Artificial Intelligence for Europe' COM(2018) 237 final, 25 April 2018, 14.

¹⁴¹ Hugenholtz (n 35) 3, 5.

¹⁴² DSM Directive, recital 18.

the current commercial reality of AI constitutes a central analytical tension in this Chapter and a key factor supporting the arguments for reform presented in Section 4.6.

4.1.3. Relationship to Existing InfoSoc Exceptions

Before the adoption of the DSM, EU copyright law lacked provisions specifically authorising TDM activities. This legislative gap was explicitly acknowledged by the Commission and served as a primary justification for the introduction of the DSM Directive.¹⁴³

The exception most relevant to TDM activities prior to the DSM Directive was Article 5(3)(a) of the InfoSoc Directive, which permits Member States to provide an exception for the use of works solely for illustration in teaching or scientific research.¹⁴⁴ However, this provision was insufficient for three reasons. First, it is limited to illustration, covering the use of works as examples or reference points in teaching and research, rather than automated analytical processing of these works as data. Second, it does not explicitly authorise automated processes; TDM does not involve human reading or illustration in the conventional sense. Third, this is an optional exception under the InfoSoc Directive, resulting in varying availability across Member States and further catalisation of the legal uncertainty identified in Recital 9 of the InfoSoc Directive.¹⁴⁵ Therefore, this pre-DSM Directive exception was both substantively inadequate and geographically inconsistent.

As a result, the TDM exceptions in the DSM Directive function as *lex specialis* in relation to the general exceptions of the InfoSoc Directive. They specifically and comprehensively regulate the legal framework for TDM activities under EU copyright law, superseding the general provisions of the InfoSoc Directive where overlap occurs.¹⁴⁶ Accordingly, Articles 3 and 4 of

¹⁴³ European Commission, COM(2016) 593 final (n 130) 3.

¹⁴⁴ InfoSoc Directive, art 5(3)(a).

¹⁴⁵ DSM Directive, recital 9; Geiger, Frosio and Bulayenko (n 49) 177.

¹⁴⁶ Martin Senftleben, 'The New EU Text and Data Mining Exceptions' (2021) 70 GRUR International 1, 3.

the DSM Directive form the applicable legal framework for assessing the copyright implications of AI training. The general exceptions of the InfoSoc Directive, including the temporary copying exception under Article 5(1) as discussed in Chapter 3, do not provide an alternative or supplementary basis for AI training activities beyond what is permitted by the DSM Directive.¹⁴⁷

4.1.4. Definition of TDM — Article 2(2) DSM Directive

Prior to applying the provisions of Articles 3 and 4 to AI training, it is necessary to determine whether the training of AI models constitutes a TDM process as defined by the DSM Directive. The resolution of this issue determines whether the TDM exceptions are even engaged as the potentially applicable legal framework.

Article 2(2) of the DSM Directive defines TDM as “any automated analytical technique aimed at analysing text and data in digital form in order to generate information such as patterns, trends and correlations”.¹⁴⁸ This definition contains four operative elements, each of which requires examination in the AI training context.

The first element, automation of the technique, presents no challenges in the context of AI training. Processes such as web scraping, dataset compilation, preprocessing, and model training are all executed by automated software systems without direct human intervention at the level of individual data elements. AI training is inherently automated, and this element is therefore unambiguously satisfied.¹⁴⁹

The second element, working with text and data in digital form, is also readily fulfilled. The training data for AI systems comprises digitally stored textual content, images, and other data

¹⁴⁷ See Chapter 3, section 3.4 of this thesis.

¹⁴⁸ DSM Directive, art 2(2).

¹⁴⁹ Margoni and Kretschmer (n 137) 6.

in machine-readable formats. Web-scraped corpora, digitised book collections, and curated datasets all constitute text and data in digital form within the meaning of this provision.¹⁵⁰

The third element, that the technique must be aimed at analysing, introduces a purpose requirement. The method must be directed toward analysis rather than simple copying or distribution. In AI training, the process aims to extract statistical and linguistic patterns from the training corpus, which the model encodes in its weights and later uses to generate output data. This analytical purpose is evident, and thus the third element is also satisfied.¹⁵¹

The fourth and most analytically significant element requires that the technique be directed toward generating information, such as patterns, trends, and correlations. At this aspect, applying the TDM definition to foundation model training presents a substantive interpretative challenge. The outcome of AI training is a commercial model capable of generating natural language text, translating documents, or creating images. This does not directly align with the “information such as patterns, trends and correlations” referenced in the definition. Legislative examples indicate results of an analytical or informational character, such as research findings, statistical analysis, or market analysis, rather than commercially deployed AI systems capable of independent content generation.¹⁵² From this perspective, training a foundation model appears to produce an outcome fundamentally distinct from what the TDM definition was designed to cover.

Nevertheless, the counterargument is more compelling from a standpoint of legal interpretation. The definition employs the phrase “information such as patterns, trends and correlations”, where the term “such as” signals an illustrative rather than exhaustive enumeration.¹⁵³ The representations, statistical associations, and distributional knowledge encoded within a trained AI

¹⁵⁰ Geiger, Frosio and Bulayenko (n 49) 178.

¹⁵¹ Sobel (n 68) 1, 25.

¹⁵² Margoni and Kretschmer (n 137) 7.

¹⁵³ DSM Directive, art 2(2).

model constitute, at a fundamental level, information derived from the patterns and correlations present in the training data, which aligns with what the definition describes. The fact that this information is embedded in the model's weights, rather than articulated as a conventional research output, does not exclude it from the definition, given the technologically neutral formulation of the provision. Moreover, the principle of legal certainty and the interpretative approach consistently adopted by the CJEU support a textual reading of EU legislative provisions, rather than a restrictive interpretation derived from legislative examples.¹⁵⁴ Where the text of a provision is capable of covering a given activity, as the TDM definition is capable of covering AI training, that coverage should be acknowledged, and the conditions for exclusion should serve as the appropriate mechanism to limit its scope.

This thesis contends that the training of AI models falls within the definition of TDM in Article 2(2) of the DSM Directive, and that the exceptions in Articles 3 and 4 accordingly constitute the operative legal framework for evaluating the copyright implications of AI training. This conclusion does not establish that AI training is lawful under these exceptions; rather, it indicates that Articles 3 and 4 are the relevant provisions for assessing lawfulness. The issue of whether the conditions of these exceptions are satisfied in the context of large-scale commercial AI training is examined in the following sections. Furthermore, the interpretative tension between the textual scope of the TDM definition and the legislative intent underlying it supports the conclusion in Section 4.6 that the TDM framework of the DSM Directive requires targeted reform to address the unique challenges posed by large-scale AI training.¹⁵⁵

4.2. Article 3 DSM — Exception for Research Organisations

Article 3 of the DSM Directive introduces the first and more robust of the two TDM exceptions, establishing a mandatory provision available exclusively to research organisations and cultural

¹⁵⁴ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 27; Emanuela Rosati, *Copyright and the Court of Justice of the European Union* (Edward Elgar 2019) 34.

¹⁵⁵ Geiger, Frosio and Bulayenko (n 49) 180.

heritage institutions. It is particularly significant for two reasons. First, it sets the conditions under which academic and research-oriented AI training may be lawfully conducted in accordance with EU copyright law. Second, it clearly delineates the category of actors, specifically commercial AI developers, that are excluded from its scope and must therefore rely solely on Article 4 of the DSM Directive. The following analysis addresses the personal and material scope of the exception, the requirement of lawful access (which is common to both TDM exceptions), as well as the mandatory nature and prohibition of contractual waivers that distinguish Article 3 from its commercial counterpart.

4.2.1. Personal Scope — Research Organisations

Article 2(1) of the DSM Directive defines a research organisation as “a university, a research institute or any other organisation the primary goal of which is to conduct scientific research or to carry out educational activities, including in the context of a public-private partnership, provided that it carries out those activities on a non-commercial basis or by reinvesting all its profits in its scientific research”.¹⁵⁶ This definition relies on two main criteria: the organisation’s primary purpose must be scientific research or education, and it must operate on a non-profit basis or reinvest all profits in scientific research. Each category of potentially eligible organisations requires separate analysis.

Universities and Research Institutes

Traditional academic institutions, such as universities, faculties, research institutes, and similar organisations, fall clearly within the scope of Article 3 of the DSM Directive and present no interpretative challenges. These entities are explicitly referenced in the legislative definition, and their primary function, by virtue of their institutional character, is to conduct research and provide education. The definition applies irrespective of the institution’s legal form, encompassing both state universities funded by public budgets and private universities operating on a

¹⁵⁶ DSM Directive, art 2(1).

non-profit basis, provided their primary purpose and non-profit status are maintained.¹⁵⁷ Recital 12 of the DSM Directive affirms that the definition is intended to cover a broad spectrum of research organisations in both the natural and human sciences, and that the legal form of the entity, whether a foundation, association, or public body, is not determinative for qualification.¹⁵⁸ For AI research conducted in a genuinely academic context, including the training of AI models as part of doctoral research, computational linguistics projects, or AI safety research programs, Article 3 offers a clear and comprehensive legal basis.

Cultural Heritage Institutions

Article 3 of the DSM Directive expressly applies to cultural heritage institutions, which are separately defined in Article 2(3) of the DSM Directive as publicly accessible libraries, museums, archives, and film or audio heritage institutions.¹⁵⁹ Extending the research TDM exception to these institutions is particularly significant for AI training, as digitisation projects undertaken by libraries, archives, and museums can produce substantial volumes of textual and graphical data suitable for use as AI training material. However, the application of Article 3 to cultural heritage institutions is restricted to research and educational goals. Training AI systems for purposes beyond the institution's research or preservation mission does not satisfy the requirements.¹⁶⁰ Consequently, the exemption supports AI research conducted by or in collaboration with cultural heritage institutions using their digitised collections, but does not authorise these institutions to pursue commercial AI development with their holdings.

Public-Private Partnerships — The Difficult Case

A significant complexity regarding the scope of application emerges in the context of public-private partnerships, which Article 2(1) of the DSM Directive explicitly identifies as a potential

¹⁵⁷ DSM Directive, art 2(1); Geiger, Frosio and Bulayenko (n 49) 174, 182.

¹⁵⁸ DSM Directive, recital 12.

¹⁵⁹ DSM Directive, arts 2(3), 3(1).

¹⁶⁰ DSM Directive, recital 11; Margoni and Kretschmer (n 137) 9.

form of qualified research organisation. According to this provision, a research organisation operating in the context of a public-private partnership may satisfy the criteria if the conditions of non-commercial activity are fulfilled. This issue is particularly relevant for AI development: if a qualified research organisation, such as a university, collaborates with a commercial company to jointly develop an AI model, it is necessary to determine whether the commercial partner benefits from the exemption provided in Article 3.¹⁶¹

The applicability of the exemption depends on the specific allocation of roles within the partnership. If the qualified research organisation conducts data management activities, such as web scraping, compiling datasets, and training, while the commercial partner only benefits from or accesses the research results, the research organisation's use of the exemption under Article 3 of the DSM Directive remains lawful, and the commercial partner does not directly benefit from it. The exception in Article 3 applies exclusively to the qualified organisation performing the data management; it does not extend to third parties who benefit from the activity but do not meet the criteria.¹⁶² Conversely, if the TDM is conducted jointly, or if the commercial partner effectively directs the mining and derives the primary benefit, the agreement's compliance with Article 3 becomes uncertain. The decisive factor is not the contractual label of the agreement but the substantive economic reality of whose activity it is and whose objectives it primarily serves.¹⁶³

Commercial AI Companies — Categorical Exclusion

Commercial entities engaged in AI, such as organisations that develop LLMs and foundational models as their principal commercial products, do not qualify as research organisations under Article 2(1) of the DSM Directive and are consequently excluded from the scope of Article 3.¹⁶⁴

¹⁶¹ DSM Directive, art 2(1).

¹⁶² Geiger, Frosio and Bulayenko (n 49) 183.

¹⁶³ Senftleben (n 146) 1, 4.

¹⁶⁴ DSM Directive, art 2(1).

This conclusion is based on the primary purpose test: the principal aim of a commercial AI company is the development and commercialisation of revenue-generating AI products. Scientific research, when conducted, serves as a means to achieve this commercial objective rather than constituting an end in itself. The distinction is not determined by the complexity or originality of the technical work involved, as the development of commercial AI may require highly advanced research, but rather by the institutional purpose that directs and structures such work. If the primary objective is the creation of a commercial product, the primary purpose test is not satisfied, irrespective of the research-intensive nature of specific activities.¹⁶⁵

Structuring Arrangements — Abuse of Law

The categorical exclusion of commercial AI companies from Article 3 of the DSM Directive raises the question of whether such entities might structure their operations to exploit this exclusion. To illustrate, companies may establish nominally non-commercial research subsidiaries, channel AI training activities through affiliated academic institutions, or adopt legal forms typically associated with research organisations while primarily pursuing commercial objectives.¹⁶⁶ These scenarios are not merely hypothetical; the AI industry has previously witnessed companies adopting non-commercial or hybrid legal structures, only to later transition to or restructure back to commercial entities.

The general principle of EU law prohibiting the abuse of rights, as recognised by the CJEU in *Halifax plc v Commissioners of Customs and Excise*, provides the legal foundation for addressing such arrangements.¹⁶⁷ According to the *Halifax* principle, a transaction or arrangement that formally complies with the conditions of an EU legal provision, but whose primary purpose is to secure an advantage contrary to the provision's objectives, constitutes an abuse of rights and may be disregarded. In the context of Article 3 of the DSM Directive, an agreement primarily

¹⁶⁵ Margoni and Kretschmer (n 137) 10.

¹⁶⁶ Geiger, Frosio and Bulayenko (n 49) 184.

¹⁶⁷ Case C-255/02 *Halifax plc v Commissioners of Customs and Excise* [2006] ECR I-1609, paras 74–75.

intended to access the TDM exemption for research, rather than to pursue genuine scientific studies, would be considered an abuse. The commercial organisation behind such an agreement would therefore not be covered by the personal scope of Article 3.¹⁶⁸ When applying this provision, national courts would evaluate the substance of the agreement rather than its formal structure, focusing on whether the main beneficiary of the TDM activities is a legitimate research organisation or a commercial entity using a research structure to obtain access to an otherwise unavailable exemption. As a result, such artificial arrangements entail significant legal risk and are not a viable strategy for commercial AI companies seeking to benefit from Article 3.

4.2.2. Material Scope — Scientific Research Purpose

The personal scope of Article 3 of the DSM Directive, as outlined in the previous sub-section, specifies the beneficiaries of the exception, while the material scope defines the permissible purposes for its application. Article 3(1) restricts the exception to reproduction and extraction conducted “for the purposes of scientific research”.¹⁶⁹ This provision accordingly introduces a purpose limitation that serves as an additional condition of control, complementing the requirement of personal scope.

The DSM Directive does not provide a definition of scientific research. Recital 12 offers limited clarification, indicating that the concept covers both the natural and human sciences. This prevents a narrow interpretation that would confine the exception to STEM disciplines and exclude fields such as the social sciences, digital humanities, and computational linguistics.¹⁷⁰ The scope of the scientific research requirement should therefore be derived from the general academic understanding of the term and a purposive interpretation of the provision. Traditional

¹⁶⁸ Case C-255/02 *Halifax plc v Commissioners of Customs and Excise* [2006] ECR I-1609, para 86; Senftleben (n 146) 5.

¹⁶⁹ DSM Directive, art 3(1).

¹⁷⁰ DSM Directive, recital 12.

academic research conducted at universities and research institutes, including AI research focused on language models, computer vision, or AI safety, clearly falls within the concept of scientific research.¹⁷¹ Applied research, which seeks the practical application of scientific knowledge, is similarly covered, provided it is undertaken by a qualifying entity and results in genuine research outcomes rather than primarily commercial products.

The distinction between research-oriented AI training and the development of commercial foundation model defines the operational boundary of the material scope of Article 3 of the DSM Directive in the context of AI. Training a model to research the linguistic properties of large text corpora, to explore the capabilities and limitations of transformer architectures, or to advance scientific understanding of machine learning (where the primary outcome is a contribution to research) meets the requirement for a scientific research objective.¹⁷² In contrast, training a model, the key outcome of which is a commercially deployable product (regardless of the technical complexity of the process or incidental research findings) does not satisfy this requirement. The decisive criterion is not the technical nature of the activity but rather the primary intended output: a contribution to scientific knowledge or the development of a commercial product.¹⁷³

A significant practical challenge arises in cases of dual-use AI training, where a model initially trained within a research-oriented framework is later commercialised, or where research institutions conduct training that produces both academic research results and commercially exploitable models.¹⁷⁴ Resolving this issue depends on the timing of the assessment of the training's goal. Under Article 3 of the DSM Directive, the purpose requirement is evaluated at the time the TDM activities are conducted, specifically during the acts of reproduction that trigger the reproduction right under Article 2 of the InfoSoc Directive. Subsequent commercialisation

¹⁷¹ Geiger, Frosio and Bulayenko (n 49) 185.

¹⁷² Senftleben (n 146) 5.

¹⁷³ Margoni and Kretschmer (n 137) 11.

¹⁷⁴ Geiger, Frosio and Bulayenko (n 49) 186.

of a model originally trained for genuine scientific research does not retroactively invalidate the application of Article 3 of the DSM Directive to the training activities; the exception is satisfied at the time of use, and later commercial exploitation does not alter this assessment.¹⁷⁵

The reverse scenario requires a different conclusion. If commercialisation of the AI model is planned or anticipated from the outset of the training activity, and where the research framing serves merely as a “better” legal framing, the objective of scientific research is not genuinely met, and Article 3 of the DSM Directive does not apply. When the purpose of training is commercial from the beginning, a later claim of a research purpose cannot retroactively transform a fundamentally commercial activity into one eligible for the research TDM exemption.¹⁷⁶ Determining which of these two situations applies in a specific case requires a factual assessment of the genuine purpose at the time of training, which is ultimately a matter for the national court applying Article 3. The external practical challenge of this assessment should be acknowledged: the internal purpose of a training project is not always apparent to rightholders or to courts lacking access to internal documentation. This evidentiary difficulty highlights a further gap in the current system, which could be addressed by implementing a robust registration or notification system for large-scale TDM events, as discussed in Section 4.6.¹⁷⁷

4.2.3. Lawful Access Requirement

Article 3(1) of the DSM Directive stipulates that, for the exception to apply, the organisation concerned must have lawful access to the works comprising the digital media content.¹⁷⁸ This requirement serves as a threshold condition, applying independently of and prior to any assessment of personal or material scope. If lawful access to a specific work cannot be established,

¹⁷⁵ DSM Directive, art 3(1); Margoni and Kretschmer (n 137) 12.

¹⁷⁶ Senftleben (n 146) 6.

¹⁷⁷ Sobel (n 68) 1, 27.

¹⁷⁸ DSM Directive, art 3(1).

the exception does not apply to that work, irrespective of whether the requirements regarding the organisation and the purpose have been satisfied.

Lawful access, as defined in Article 3(1) of the DSM Directive, requires that the accessing entity possess a legal right to access the specific work used for TDM. Recital 14 of the DSM Directive clarifies that lawful access encompasses access via subscription or licence agreement, access to open-access publications provided under Creative Commons licences or equivalent authorisations, and access to works in the public domain.¹⁷⁹ Institutional access agreements, such as those typically entered into by universities for access to academic databases and journal platforms, also constitute lawful access to the extent that the relevant licences cover the works used for TDM. It is important to note that lawful access must be established for the specific copy of the work used for analysis, rather than for the work as a whole. For example, an organisation that holds a licence to access a database via a web interface, but whose licence does not permit automated bulk downloading, does not have lawful access to the downloaded copies for the purposes of Article 3 of the DSM Directive.¹⁸⁰

Content, access to which is restricted by a paid subscription, presents a particular challenge in terms of complying with the lawful access requirement. Research organisations with active institutional subscriptions to academic databases have lawful access to the content of these databases, but this access is defined and restricted by the terms of the licence agreement. If the licence explicitly permits automated processing or TDM, the condition of lawful access is clearly met. If, however, the licence is silent on TDM, the question arises whether the subscription itself provides sufficient grounds to perform TDM activities beyond the normal scope of access. This question is not definitively resolved in the DSM Directive, although Recital 14 suggests that lawful access by subscription should, in principle, permit TDM activities without

¹⁷⁹ DSM Directive, recital 14.

¹⁸⁰ Hugenholtz (n 35) 3, 6.

an explicit contractual restriction (albeit subject to contractual override prohibition, as analysed in sub-section 4.2.4).¹⁸¹ Circumventing paid access by technical means, instead of accessing content via a valid subscription, clearly does not constitute lawful access and simultaneously is an independent copyright infringement.¹⁸²

The lawful access requirement imposes a de facto verification obligation on organisations carrying out TDM: before reproducing a work as part of a TDM dataset, the entity must be able to prove that its access to that work is lawful. On the scale of modern AI training, which involves datasets containing millions or hundreds of millions of documents from various sources, this verification obligation is practically burdensome and, in many cases, technically unfeasible.¹⁸³ The DSM Directive does not establish a standard of due diligence or a verification methodology, leaving this gap to be filled by national legislation or, in the absence of national provisions, by general principles of copyright liability. For research organisations operating under Article 3 of the DSM Directive, the practical consequence is that training datasets should be compiled from sources whose legality can be reliably verified, such as subscriptions to institutional databases, open-access repositories, and public domain corpora, rather than from indiscriminate web scraping, the legal status of which cannot be reliably assessed. The absence of a statutory standard of due diligence represents a further gap within the DSM Directive and contributes to the legal uncertainty that the TDM exceptions were intended to resolve.¹⁸⁴

4.2.4. Mandatory Nature and Contractual Override Prohibition

Article 3 of the DSM Directive establishes a mandatory exception. In contrast to the general exceptions listed in Article 5 of the InfoSoc Directive, which Member States may implement at

¹⁸¹ DSM Directive, recital 14; Geiger, Frosio and Bulayenko (n 49) 187.

¹⁸² Margoni and Kretschmer (n 137) 13.

¹⁸³ Henderson and others (n 15) 1, 16.

¹⁸⁴ Geiger, Frosio and Bulayenko (n 49) 188.

their discretion, Article 3 of the DSM Directive requires transposition into the national legislation of all Member States.¹⁸⁵ This mandatory character has two important consequences. First, qualifying research organisations benefit from this exception across all Member States, irrespective of whether a particular Member State has enacted additional or more favourable TDM provisions. Second, rightholders are unable to opt out of Article 3; no mechanism comparable to the machine-readable reservation in Article 4(3) of the DSM Directive exists to exclude research organisations from benefiting the Article 3 exception. The absence of an opt-out mechanism demonstrates a deliberate legislative position that research TDM serves public interest, enhancing EU research competitiveness, advancing scientific progress, and improving access to knowledge, which outweighs the individual economic interests of rightholders in controlling the analytical use of their works.¹⁸⁶

The mandatory character of Article 3 of the DSM Directive is further reinforced by Article 7(1) of the same Directive, which stipulates that any contractual provision conflicting with the exceptions in Articles 3 and 5 of the Directive shall be void.¹⁸⁷ This contractual override prohibition is particularly relevant for AI training, as many websites and online platforms include in their terms of service broad clauses prohibiting automated access, scraping, or systematic downloading of content. Content providers frequently rely on such provisions to restrict access to AI training crawlers. Article 7(1) renders these provisions unenforceable against relevant research organisations when they aim to restrict activities covered by Article 3. In this context, a contractual restriction cannot override the statutory right.¹⁸⁸

Article 3(2) of the DSM Directive authorises rightholders to implement measures to ensure the security and integrity of their networks and databases, provided such measures do not exceed

¹⁸⁵ DSM Directive, art 3(1); recital 9.

¹⁸⁶ Senftleben (n 146) 6.

¹⁸⁷ DSM Directive, art 7(1).

¹⁸⁸ Geiger, Frosio and Bulayenko (n 49) 189.

what is necessary for that purpose.¹⁸⁹ This provision introduces the prohibition on contractual circumvention and necessitates a clear distinction: rightholders may adopt technical measures that serve genuine security objectives, such as authentication requirements, rate limiting, or CAPTCHA verification, without these being considered impermissible restrictions on TDM. Conversely, measures that effectively block TDM under the pretext of security, thereby circumventing the Article 3 exception of the DSM Directive through technical rather than contractual means, are not permitted.¹⁹⁰ The distinction between a legitimate security measure and a de facto restriction on TDM is determined by proportionality: only measures that are necessary and proportionate to a genuine security objective are allowed, while those that exceed security requirements and effectively prevent TDM are not. This distinction is of practical significance for research organisations encountering technical barriers to dataset compilation and for national courts resolving disputes between rightholders and research TDM practitioners.

The very same Article 3(2) of the DSM Directive also stipulates that copies made for TDM purposes may be stored for scientific research purposes, including for verifying research results.¹⁹¹ This provision has significant cross-border implications, as international research collaborations among institutions in different Member States are prevalent in AI research. The ability to store and exchange TDM copies within cross-border research frameworks, without infringing reproduction rights, is essential for such collaborations. The exception applies as long as the entity has lawful access to the relevant works, irrespective of the location of servers used for storing or processing data within the EU.¹⁹² This cross-border dimension is especially pertinent for large-scale AI research projects conducted by consortia of EU universities, a model that is increasingly common in publicly funded AI research initiatives.

¹⁸⁹ DSM Directive, art 3(2).

¹⁹⁰ Margoni and Kretschmer (n 137) 14.

¹⁹¹ DSM Directive, art 3(2).

¹⁹² Senftleben (n 146) 7.

4.3. Article 4 DSM — General TDM Exception

Section 4.2 established that commercial AI developers are categorically excluded from the scope of Article 3 of the DSM Directive. The analysis now turns to Article 4 of the DSM Directive, the only remaining legislative provision in the EU copyright framework that potentially could provide a legal basis for commercial AI training. Article 4 differs from Article 3 in three fundamental aspects that shape the analysis in this Section: its personal scope is not limited, its material scope contains no purpose limitation, and its availability is conditional upon the absence of a rightholder's opt-out. The third feature, which is entirely absent from Article 3, serves as the central mechanism through which the DSM Directive seeks to balance the broad scope of Article 4 with the interests of rightholders. Accordingly, this feature receives the most extensive treatment in this Section.

4.3.1. Personal Scope

Article 4(1) of the DSM Directive requires Member States to introduce an exception to the reproduction right under Article 2 of the InfoSoc Directive, specifically for “for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining”.¹⁹³ In contrast to Article 3 of the DSM Directive, this provision does not reference research organisations, cultural heritage institutions, or any other specific categories of entities. The exception applies to any natural or legal person.

The broad personal scope of Article 4 of the DSM Directive is both intentional and comprehensive. There are no income thresholds, organisational size restrictions, or sectoral limitations that constrain the availability of the exception. As a result, individual developers, start-ups without prior commercial activity, and multinational technology corporations with substantial global revenues all fall within the scope of Article 4 on equal terms.¹⁹⁴ Commercial AI companies,

¹⁹³ DSM Directive, art 4(1).

¹⁹⁴ Geiger, Frosio and Bulayenko (n 49) 174, 190.

which are categorically excluded from Article 3 as discussed in sub-section 4.2.1, are not only included in Article 4, but their inclusion represents a significant and deliberate legislative choice.¹⁹⁵ The exclusion of commercial entities from Article 3 and their inclusion in Article 4 are interconnected legislative decisions, reflecting a compromise whereby the DSM channels commercial activities specifically to Article 4, subjecting them to an opt-out mechanism that is not available under Article 3.

This extension of the TDM exemption to commercial entities of any size constituted the most politically contentious aspect of the legislative process described in sub-section 4.1.1. The initial 2016 Commission proposal did not include a provision equivalent to Article 4 of the DSM Directive; its subsequent inclusion marked a significant concession to data-intensive industries, including but not limited to the AI sector, relative to the Commission's original research-focused approach.¹⁹⁶ However, this concession was not unconditional. The opt-out mechanism in Article 4(3) served as the condition for broadening the personal scope, thereby enabling rightholders' representatives to accept the extension of TDM rights to commercial entities.¹⁹⁷ Understanding this political economy is essential for the critical assessment presented in this Chapter. The legitimacy of the broad personal scope of Article 4, as a legislative compromise, depends on the effectiveness of the opt-out mechanism as a safeguard for rightholders. If this mechanism fails to function effectively, as explored in sub-section 4.3.4, the foundational compromise underlying the personal scope of Article 4 is consequently undermined.

4.3.2. Material Scope

Article 4 of the DSM Directive does not impose purpose-based restrictions comparable to the scientific research requirement specified in Article 3. Instead, it applies to TDM conducted “for

¹⁹⁵ Hugenholtz (n 35) 3, 7.

¹⁹⁶ European Commission, COM(2016) 593 final (n 130); Senftleben (n 146) 1, 8.

¹⁹⁷ Geiger, Frosio and Bulayenko (n 49) 191.

the purposes of text and data mining”, without further clarification regarding the intended purpose of such analysis.¹⁹⁸ Consequently, any purpose for which TDM is performed, as defined in Article 2(2) of the DSM Directive and discussed in sub-section 4.1.4, falls within the material scope of Article 4. This includes the development of commercial products, business analytics, market analysis, and, as analysed in Section 4.1.4, the training of foundational AI models.

On the face of the provision, therefore, large-scale commercial AI training is encompassed within the material scope of Article 4 of the DSM Directive. The provision does not establish quantitative limits on the scale of data analysis activities, nor does it restrict the nature or commercial character of the resulting outputs. A literal application of Article 4(1) to the training of a foundation model on hundreds of billions of documents encounters no textual obstacle.¹⁹⁹

When the wording of Article 4 of the DSM Directive is examined within the legislative context outlined in Section 4.1, a notable interpretative tension emerges. The legislative history and the recitals of the DSM Directive consistently frame TDM exceptions by reference to uses of a different nature and scale than the large-scale training of foundational models. Recital 18 of the DSM Directive, specifically addressing the commercial exception, refers to uses such as “browsing the internet [...] or other lawful uses”. This formulation suggests individual or minor analytical activities, rather than the systematic reproduction of internet-scale datasets for training commercial AI systems worth billions of euros.²⁰⁰ The Commission’s 2016 proposal and the accompanying impact assessment, as discussed in sub-section 4.1.1, focused on enhancing the research competitiveness and data-driven innovation by small and medium-sized enter-

¹⁹⁸ DSM Directive, art 4(1).

¹⁹⁹ Margoni and Kretschmer (n 137) 15.

²⁰⁰ DSM Directive, recital 18.

prises, not the development of advanced AI models. Accordingly, Article 4 of the DSM Directive should be interpreted in light of this legislative context, and its application should not extend to activities of a nature and scale that the legislator did not anticipate.²⁰¹

The textual scope argument, however, is more compelling, based on established principles of EU law interpretation. The text of Article 4(1) of the DSM Directive is unambiguous, contains no limitations regarding scale, and applies to “reproductions and extractions of lawfully accessible works [...] for the purposes of text and data mining” without any further qualification.²⁰²

The principle of legal certainty, a foundational principle of EU law, precludes reliance on legislative history to restrict the scope of a clear statutory provision. Individuals subject to EU law must be able to determine their rights and obligations from the legislation itself, rather than from the subjective intentions of legislators, which may not be apparent from the text itself.²⁰³

Additionally, the principle of strict interpretation, as affirmed by the CJEU in the *Infopaq International* case discussed in sub-section 3.4.1, requires that the conditions for an exception be interpreted strictly. However, strict interpretation does not equate to a narrow interpretation.

This principle means that the text of the provision, including its express conditions, must be respected and not extended beyond its terms, and precludes interpretation based on extrinsic considerations when the terms are clear.²⁰⁴ Where the text of Article 4 of the DSM Directive is capable of encompassing large-scale AI training (as it plainly is), this coverage should be acknowledged.

This thesis adopts the textual scope position. Article 4 of the DSM Directive, according to its wording, covers large-scale commercial AI training. This conclusion, however, requires precise

²⁰¹ European Commission, COM(2016) 593 final (n 130); Margoni and Kretschmer (n 137) 16.

²⁰² DSM Directive, art 4(1).

²⁰³ Leistner (n 64) 1, 10.

²⁰⁴ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569, para 56; see Chapter 3, sub-section 3.4.1 of this thesis.

understanding. It does not suggest that AI training conducted under Article 4 is inherently lawful. Rather, Article 4 serves as the relevant legal framework within which the lawfulness of AI training must be evaluated. The conditions of Article 4, particularly the lawful access requirement and the opt-out mechanism, must be rigorously applied to the specific circumstances of each training activity. Textual scope determines the applicable legal framework but does not establish compliance with it. The interpretative tension identified in the present Section, which exists between the broad language of the text and the narrower activities the legislator may have intended, does not disappear. Instead, it shifts to the evaluation of whether conditions of Article 4 can effectively regulate activities of a type and scale not specifically anticipated by the legislator. This ongoing tension is the central focus of the critical assessment in Section 4.6.²⁰⁵

4.3.3. Lawful Access Requirement

Article 4(1) of the DSM Directive defines the exception for lawful access to works covered by subject to TDM in terms identical to Article 3(1) of the same Directive, as detailed in sub-section 4.2.3.²⁰⁶ The legal definitions of lawful access, the handling of paywall-protected content, and the categorical exclusion of pirated sources discussed in that sub-section apply equally to Article 4 and are not reiterated here. However, two considerations unique to the commercial context of Article 4 require separate analysis.

While the legal content of the lawful access requirement is consistent across Articles 3 and 4 of the DSM Directive, its practical implications are substantially different for commercial AI developers operating at scale. If a research institution mistakenly assumes lawful access to a single database, the resulting risk is limited in both the extent of infringement and the practical capacity to remediate it. In contrast, a commercial AI company that incorporates unlawfully accessed content into a training dataset containing hundreds of millions of documents, and subsequently

²⁰⁵ Geiger, Frosio and Bulayenko (n 49) 192.

²⁰⁶ DSM Directive, art 4(1); see sub-section 4.2.3 of this thesis.

deploys this dataset in a commercial model accessed by millions of end users, encounters a categorically different risk profile. The infringement is reproduced at scale, the model may be difficult or impossible to retrain without the infringing data, and commercial exploitation amplifies the economic impact of the infringement.²⁰⁷ Although the legal standard of lawful access remains unchanged, the consequences of non-compliance and the practical incentive for rigorous verification are significantly greater for commercial operators engaged in large-scale TDM.

The scale at which commercial datasets for AI training are assembled has led to the so-called argument of practical impossibility: it is technically infeasible for an AI developer to verify the lawfulness of access to every document among the hundreds of billions of tokens comprising a training corpus.²⁰⁸ Although this argument is factually accurate, it does not reduce the legal obligation. The requirement for lawful access is a condition for exemption, not a standard adjusted to what is practically convenient for the entity seeking to rely upon it. Activities conducted at a scale that prevents compliance with legal requirements are unlawful to the extent compliance cannot be demonstrated; the scale of the activity does not justify the absence of a legal basis for it.²⁰⁹ The practical solution for commercial AI developers is not to abandon verification, but to restructure it. Training datasets should be compiled from sources where lawful access can be reviewed at the source level, such as licensed content aggregators, verified open-access repositories, or datasets with documented provenance, rather than from undifferentiated web scraping, where the legality of individual documents cannot be established. Such an approach has direct financial consequences, favouring developers with established licensing relationships and infrastructure for verifying dataset provenance over smaller developers who rely on open web scraping. This competitive dynamic is an important consideration for policy assessment in Chapter 6.²¹⁰

²⁰⁷ Sobel (n 68) 1, 29.

²⁰⁸ Henderson and others (n 15) 1, 17.

²⁰⁹ Geiger, Frosio and Bulayenko (n 49) 193.

²¹⁰ Margoni and Kretschmer (n 137) 17.

4.3.4. The Opt-Out Mechanism

The opt-out mechanism established by Article 4(3) of the DSM Directive represents the most significant provision examined in this Chapter. As detailed in sub-section 4.3.1, this mechanism renders the otherwise unconditional scope of Article 4 both politically and legally acceptable within the legislative compromise that resulted in the adoption of the DSM Directive. Consequently, the practical effectiveness of the opt-out mechanism is central to addressing Research Question 3, which concerns the extent to which the mechanism balances innovation with the protection of rightholders' interests. This sub-section provides a detailed analysis of the legislative basis of the mechanism, its legal effect, the technical means of implementation, the absence of mandatory technical standards, the practical enforceability of opt-outs, and the documented compliance practices of major AI developers.

Article 4(3) of the DSM Directive stipulates that the exception established by Article 4(1) “shall apply on condition that the use of works and other subject matter referred to in [Article 4(1)] has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online”.²¹¹ This provision therefore functions as a negative condition: Article 4 applies only in the absence of a valid reservation. Where a rightholder has validly reserved their rights in relation to a specific work, Article 4 does not apply, and any commercial entity seeking to use the work for TDM must either obtain a licence from the rightholder or exclude the work from its training dataset.²¹²

The opt-out mechanism applies to specific content rather than operating on an “all or nothing” basis. A rightholder may reserve rights in certain works while permitting TDM in others, either by making no reservations or by express licensing. To illustrate, a newspaper publisher may reserve its current news content from TDM while allowing the use of its historical archive, or

²¹¹ DSM Directive, art 4(3).

²¹² Geiger, Frosio and Bulayenko (n 49) 194.

may license selected content for AI training while reserving the remainder.²¹³ This level of granularity is, in principle, a key advantage of the mechanism, as it enables rightholders to adjust their approach according to the commercial value and strategic significance of different categories of output, rather than being limited to a binary choice between blanket permission and blanket prohibition.

Article 4(3) of the DSM Directive does not specify the technical format required for a machine-readable reservation, instead referencing “machine-readable means” as an example of an “appropriate manner”.²¹⁴ In the absence of a prescribed format, various existing and emerging technical mechanisms have been considered or implemented as potential means of expressing a reservation under Article 4(3). Each mechanism possesses distinctive characteristics and limitations.

Robots.txt. The Robots Exclusion Protocol, typically implemented via a robots.txt file placed in the root directory of a website, is a long-established web standard that enables website operators to inform automated crawlers about which parts of the site may or may not be accessed. It also allows targeting of specific crawlers by user-agent identifier.²¹⁵ Robots.txt was not designed as a copyright protection tool; rather, it is a technical convention within the web ecosystem that predated the DSM Directive by more than two decades, and there is no legal obligation to comply with it outside the DSM framework.²¹⁶ Whether the “Disallow” directive in robots.txt constitutes a valid machine-readable stipulation for the purposes of Article 4(3) has not been definitively established in EU law. However, its machine-readable nature and its function of conveying access restrictions to automated systems make it a plausible candidate. Nevertheless, robots.txt suffers from three significant limitations in this context. First, compliance is volun-

²¹³ Senftleben (n 146) 9.

²¹⁴ DSM Directive, art 4(3).

²¹⁵ Internet Engineering Task Force, *Robots Exclusion Protocol*, RFC 9309 (2022).

²¹⁶ Margoni and Kretschmer (n 137) 18.

tary due to the protocol's nature; a crawler may ignore the file, and there are no technical enforcement mechanisms to prevent this.²¹⁷ Second, there is no authoritative standard defining which robots.txt directives, if any, constitute a valid reservation under Article 4(3), as opposed to a general access preference unrelated to copyright. This creates uncertainty for both rightholders and AI developers regarding the legal significance of any specific directive. Third, the protocol's user-agent targeting mechanism depends on crawlers accurately identifying themselves. If a search robot used for AI training does not identify itself, or uses a generic user-agent not specified in the website's robots.txt configuration, the targeted disallow directive may simply fail to apply to it.²¹⁸

HTTP header tags — X-Robots-Tag. A related mechanism is the X-Robots-Tag, an HTTP response header that performs a function similar to robots.txt but operates at the level of individual server responses rather than through a single file for the entire site. Its primary advantage is its applicability to non-HTML content, such as PDF documents and images, which robots.txt cannot address in detail.²¹⁹ However, X-Robots-Tag is subject to the same limitations as robots.txt: it lacks independent legal force outside the DSM framework, compliance is voluntary, and its adoption among rightholders (particularly smaller entities without advanced web infrastructure) remains limited.²²⁰ Technically proficient rightholders, including large publishers, are more likely to implement this mechanism, but its overall coverage of content actually used in AI training datasets remains uncertain.

Metadata Standards. Embedded metadata standards, such as IPTC and XMP metadata commonly used in image files, as well as structured schema.org markup on web pages, constitute a category of mechanisms that can convey explicit rights reservation information directly within

²¹⁷ Internet Engineering Task Force (n 215) s 2.

²¹⁸ Henderson and others (n 15) 19.

²¹⁹ Margoni and Kretschmer (n 137) 19.

²²⁰ Geiger, Frosio and Bulayenko (n 49) 195.

content files or web pages, rather than as a separate instruction for the entire website.²²¹ However, metadata-based approaches are limited by their vulnerability in the data processing pipeline used to generate AI training datasets. Metadata is frequently removed, altered, or lost during web scraping, format conversion, and aggregation processes prior to inclusion in the training corpus.²²² Although a rightholder may embed a clear and explicit rights reservation in an image file’s metadata, this reservation may not reach the AI developer if the collection and pre-processing pipeline fails to preserve the metadata. As a result, the reservation becomes invisible at the intended point of enforcement, through no failure on the rightholder’s part.

Emerging Standards — C2PA. The Coalition for Content Provenance and Authenticity has developed the C2PA technical specification, which cryptographically binds provenance and rights information to content files, thereby increasing resistance to data deletion and loss commonly associated with conventional metadata.²²³ C2PA offers a more robust solution to the limitations of traditional metadata standards, as its cryptographic linking mechanism is designed to withstand format conversions and processing that typically remove metadata. Despite these advantages, C2PA remains a new standard with limited adoption in the content ecosystem. It has not been recognised as a valid Article 4(3) mechanism in any binding EU document, and its potential to serve as a standardised opt-out format depends on future regulatory recognition.²²⁴ The development of C2PA illustrates that technical solutions for implementing opt-out mechanisms are available or in progress. The absence of an effective opt-out mechanism is attributable not only to the lack of adequate technologies, but also to insufficient regulatory measures mandating or recognising such technologies.

Analysis of the technical mechanisms presented above reveals a consistent issue: Article 4(3) of the DSM Directive references “machine-readable means” as an example of an “appropriate

²²¹ Senftleben (n 146) 10.

²²² Margoni and Kretschmer (n 137) 20.

²²³ Coalition for Content Provenance and Authenticity, C2PA Technical Specification (2023).

²²⁴ Coalition for Content Provenance and Authenticity (n 223) s 1.2; Margoni and Kretschmer (n 137) 21.

manner” of reservation, yet neither the DSM Directive nor any subsequent Commission implementing act has established a mandatory technical standard specifying which mechanisms constitute a valid reservation for this purpose.²²⁵ This omission is not a minor drafting oversight. Instead, it has led to a fragmented landscape in which rightholders employ different mechanisms, or none at all, due to inaction or limited technical capacity. AI developers consequently face significant uncertainty regarding which signals, if any, constitute legally effective reservations that must be recognised. Furthermore, no authoritative body exists to resolve disputes about the validity of particular mechanisms except through individual litigation. The absence of a mandatory standard represents the first of three structural gaps within the TDM framework identified in Section 4.6. This gap is arguably the most readily addressed: a Commission implementing act that defines a recognised technical standard, potentially building on existing infrastructure such as C2PA, could resolve much of the uncertainty outlined in this sub-section without requiring amendment of the DSM Directive.²²⁶

Furthermore, even if a rightholder expresses a reservation through a mechanism that could, in principle, constitute a valid refusal under Article 4(3) of the DSM Directive, enforcing that opt-out against an AI developer who has used the reserved content presents significant evidentiary challenges. To establish a breach of a valid opt-out, the rightholder must demonstrate three elements: that the opt-out was validly expressed in a machine-readable form recognised, or at least recognisable, as an opt-out under Article 4(3); that the AI developer reproduced the content after the reservation was formalised; and that the reserved content was actually included in the training dataset used for the AI model.²²⁷ Each element poses independent difficulties, but the third one is particularly problematic. The composition of large-scale AI training datasets is typically opaque to the public; therefore, rightholders lack the means to determine whether their

²²⁵ DSM Directive, art 4(3); Geiger, Frosio and Bulayenko (n 49) 196.

²²⁶ Senftleben (n 146) 11.

²²⁷ Sobel (n 68) 30.

specific content has been used without disclosure from the AI developer.²²⁸ Article 53 of the EU AI Act, which requires providers of general-purpose AI models to publish sufficiently detailed summaries of training data, partially addresses this issue by introducing a degree of transparency that was previously absent. However, a “sufficiently detailed summary” does not equate to a comprehensive list capable of confirming or denying the inclusion of any specific work, and the evidentiary burden on rightholders remains substantial even with such disclosures.²²⁹

The preceding analysis assumed that, where a legitimate and recognisable right to opt-out exists, AI developers would comply with it. However, empirical evidence regarding compliance practices among major AI developers casts doubt on even this assumption. Documented cases indicate that AI training crawlers used by leading technology companies sometimes fail to comply with robots.txt directives intended to exclude their use. Such non-compliance may result from technical limitations, delays in implementing compliance measures after public consultation, or the use of crawler identifiers not addressed by existing robots.txt configurations.²³⁰ The cumulative effect of the legal uncertainty outlined above and these documented compliance gaps is that the opt-out mechanism, even when properly invoked, may fail to influence AI developer behaviour in practice, transforming the opt-out from an enforceable legal right into more of an aspirational protection for many rightholders.²³¹

4.3.5. Temporal Problem — Historical Training Datasets

A further dimension of the adequacy of the opt-out mechanism, distinct from yet related to the implementation and enforcement issues examined in sub-section 4.3.4, concerns its temporal

²²⁸ Henderson and others (n 15) 20.

²²⁹ EU AI Act, art 53.

²³⁰ Henderson and others (n 15) 21.

²³¹ Sobel (n 68) 31.

application to training datasets compiled prior to the legal establishment of the opt-out mechanism. This topic has received comparatively limited attention in academic literature and merits careful original treatment.

Most large-scale datasets that underpin contemporary foundational models were compiled before the 7 June 2021 — transposition deadline for the DSM Directive, and in many instances, years before the same Directive’s adoption in 2019.²³² To illustrate, the Common Crawl corpus has been continuously compiled since 2008, with many historical snapshots predating both the DSM Directive and any opt-out infrastructure by over a decade.²³³ This temporal context raises a specific legal question: if a rightholder later makes a reservation under Article 4(3) of the DSM Directive regarding content collected and included in a training dataset years before that reservation, and, in many cases, before Article 4(3) existed as a legal provision, what are the legal consequences, if any, of this subsequent reservation for the earlier reproduction?

The resolution of this issue begins with a general principle of legal assessment: the lawfulness of an act is determined by the law in force at the time the act was performed.²³⁴ For example, a reproduction act carried out through data collection in 2015 must be evaluated according to the legal framework in place in 2015. At that time, Article 4(3) of the DSM Directive did not exist, as it was adopted in 2019 and implemented by 2021. A reservation made by the rightholder in 2022 cannot retroactively alter the legal characterisation of an act committed in 2015, since the reservation did not exist at the relevant time and its subsequent existence cannot be projected backwards onto an earlier act.²³⁵ This conclusion holds regardless of whether the 2015 reproduction was lawful; it establishes that, irrespective of the legal status of that reproduction, a subsequent waiver of the right cannot alter it.

²³² Margoni and Kretschmer (n 137) 22.

²³³ Carlini and others (n 7) 2633, 2636.

²³⁴ Geiger, Frosio and Bulayenko (n 49) 197.

²³⁵ Senftleben (n 146) 12.

Having established that subsequent reservations cannot retroactively affect earlier acts of reproduction, the question arises as to the legal status at the time the act of reproduction occurred, specifically, prior to the introduction of Article 4(3) of the DSM Directive. The question is whether, if Article 4(3) had existed at that time, the act might have been subject to a valid reservation. Three analytical approaches are available for consideration.

Approach 1: Pre-opt-out scraping was lawful under Article 4 of the DSM Directive because no opt-out existed to defeat it. This approach asserts that Article 4(1) applies “on condition that” no valid reservation exists under Article 4(3), and no Article 4(3) reservation could exist before that provision was in force. Under this interpretation, any TDM activity conducted before Article 4(3) existed automatically satisfied the absence-of-reservation condition, and would therefore be lawful under Article 4, provided the lawful access requirement was independently met.²³⁶ This analysis, however, fails for a more fundamental reason. Article 4 in its entirety, not merely Article 4(3), did not exist as EU law prior to the adoption and transposition of the DSM Directive. An exception cannot retroactively legitimise an act performed before the exception existed. Article 4(1) cannot be applied to the 2015 reproduction, just as Article 4(3) cannot override a reservation that did not yet exist. Approach 1 effectively attempts to apply only one half of Article 4 (the absence-of-reservation condition) to a period before either half of the provision was in force, which is not a coherent application of the provision at all.

Approach 2: Pre-DSM scraping had no statutory TDM basis under EU law. From this perspective, prior to the DSM’s implementation, EU copyright law lacked TDM-specific exceptions, as established in sub-section 4.1.3; the existing InfoSoc exceptions were inadequate to cover TDM activities. Consequently, acts of reproduction TDM purposes prior to transposition were not covered by any statutory exception in force at the time. Their lawfulness, if any, de-

²³⁶ Margoni and Kretschmer (n 137) 23.

pendent on a contractual licence or the general application of standard copyright principles, neither of which is typically present in the case of large-scale web scraping.²³⁷ Article 26 of the DSM Directive, which extends the DSM Directive provisions to all works protected by national copyright as of 7 June 2021, including pre-existing works, does not assist in this context. Article 26 addresses the application of the Directive's main provisions, including exceptions, to works of any age from the Directive's entry into force, but does not retrospectively characterise acts of reproduction carried out before that date as lawful under an exception that did not exist at the time.²³⁸ Hence, under this view, web scraping conducted before the DSM's adoption lacked a statutory basis under EU copyright law.

Approach 3: Implied consent from online publication. The third argument advanced by certain AI developers is that publishing content on the open internet constitutes implied consent for its use in automated processing, including AI training. This claim is based on the premise that content made freely accessible online (by the rightholder's own choice) is available for any use permitted by the technical architecture of the internet.²³⁹ This argument is unpersuasive. Publishing a work online makes it accessible to human readers and, incidentally, to automated systems operating within the normal technical functioning of the internet, such as search engine indexing. As discussed in sub-section 3.3.1, such indexing operates on a fundamentally different basis and serves a fundamentally different function from AI training.²⁴⁰ Online publication does not equate to consent for any subsequent use of the content by automated systems. Legally relevant consent under copyright law requires a deliberate act or manifestation by the rightholder directed at the specific use in question, rather than a passive inference from the rightholder's inability to prevent an unforeseen use, especially where the technology enabling such use may not have existed at the time of publication. The CJEU has never recognised a

²³⁷ DSM Directive, recital 18; see Chapter 4, section 4.1.3 of this thesis.

²³⁸ DSM Directive, art 26.

²³⁹ Sobel (n 68) 32.

²⁴⁰ See Chapter 3, section 3.3.1 of this thesis.

doctrine of implied consent as broad as this argument suggests. Adoption of such a doctrine would render the opt-out mechanism in Article 4(3) of the DSM Directive redundant: if online publication alone constituted consent to AI training, rightholders would be unable to meaningfully reserve their rights under Article 4(3), since publication necessarily precedes and is independent of any subsequent reservation. An interpretation that renders an explicit statutory provision entirely redundant cannot be correct.²⁴¹

Of the three approaches considered, Approach 2(that pre-DSM TDM scraping had no statutory basis under EU copyright law at the time it occurred) offers the most legally rigorous interpretation and is adopted by this thesis. However, this conclusion raises an additional question that EU law does not currently address: if content reproduced without a legal basis years ago was subsequently incorporated into the weights of a deployed commercial AI model, as discussed in sub-section 3.3.4, what remedies are available and against what, exactly, should they be directed? An injunction against further reproduction of specific works provides little practical remedy if the relevant reproduction occurred years ago and the resulting model has already been trained and deployed. Retraining the base model to exclude certain works is, at minimum, extremely costly and may be technically impossible, depending on the model architecture.²⁴² The pre-opt-out dataset problem therefore presents not only an interpretative challenge but also a practical remedial gap. EU copyright law may characterise historical acts of reproduction as lacking a legal basis, yet fail to offer a proportionate or technically feasible remedy for the ongoing consequences of these acts in deployed AI systems. This gap, which remains largely unresolved, constitutes a strong argument for the legislative clarification proposed in Chapter 6, not because the legal characterisation of past acts is unclear, but because the legal framework

²⁴¹ Geiger, Frosio and Bulayenko (n 49) 198.

²⁴² Henderson and others (n 15) 22.

provides no adequate response to a characterisation, which, according to the analysis presented here, it is otherwise capable of reaching.²⁴³

4.4. The Pirated Database Problem

The preceding analysis of Articles 3 and 4 of the DSM Directive assumed that the lawful access requirement, which is common to both articles, could in principle be satisfied and addressed the practical challenges of verifying compliance at scale. This Section considers a more clear-cut scenario: training datasets that unequivocally fail to meet the lawful access requirement because their content was sourced from materials that are themselves products of copyright infringement. In such cases, the issue is not one of verification complexity or interpretative nuance. Rather, it constitutes a threshold violation that entirely precludes reliance on Article 3 or Article 4. Available evidence indicates that this characteristic is present in the training data of at least some large AI systems.

4.4.1. Lawful Access as Categorical Exclusion

As outlined in sub-sections 4.2.3 and 4.3.3, the lawful access requirement under Articles 3(1) and 4(1) of the DSM Directive constitutes a threshold condition. This requirement must be satisfied independently and prior to any assessment of personal scope, material scope, or opt-out mechanism. If lawful access cannot be established, the exception does not apply to the relevant work, regardless of how convincingly the entity invoking the exception meets other conditions.²⁴⁴ Lawful access is not a procedural formality that can be fulfilled retrospectively or waived based on good faith; it is an essential precondition for the application of the exception. If the source of the content is the result of copyright infringement, this threshold condition

²⁴³ Sobel (n 68) 33.

²⁴⁴ DSM Directive, arts 3(1), 4(1); see sub-sections 4.2.3 and 4.3.3 of this thesis.

is not met. No degree of good faith by the organisation conducting training data analysis, nor any subsequent processing or transformation of the content, alters this conclusion.²⁴⁵

The primary distinction when analysing pirated sources of training data is between public availability and lawful availability. These concepts are frequently, and incorrectly, treated as equivalent in discussions of AI training data, including by AI developers themselves. They are not equivalent, and this distinction is critical to the analysis presented in this section.²⁴⁶

A pirated database is inherently publicly accessible, which distinguishes it from a stolen physical archive accessible only to those with unlawful physical access. Content hosted on a shadow library or pirate repository is typically available to anyone with an internet connection. However, the origin of this public accessibility is central to legal analysis. The content is publicly available not due to the rightholder's actions, but because an unrelated third party, the operator of the pirate repository, has unlawfully reproduced and distributed the rightholder's work without authorisation.²⁴⁷ In this context, public availability does not represent an exercise of the rightholder's rights; it is a violation of them. The lawful access requirement under Articles 3(1) and 4(1) examines whether the organisation conducting TDM possesses a legal title to access the work. Such a legal title is, by definition, a right derived from or consistent with the rightholder's own authorisation. The practical fact of accessibility, standing alone, is not a legal title and cannot substitute for one.²⁴⁸

This conclusion is not merely an inference; it is explicitly confirmed by the text of the DSM Directive. Recital 14 clarifies that, regarding online content, TDM may be considered authorised if rightholders “have already authorised the relevant use, for example by making their works available to the public online” under an open licence or similar agreement.²⁴⁹ The critical

²⁴⁵ Geiger, Frosio and Bulayenko (n 49) 174, 199.

²⁴⁶ Margoni and Kretschmer (n 137) 24.

²⁴⁷ Geiger, Frosio and Bulayenko (n 49) 200.

²⁴⁸ DSM Directive, recital 14; Margoni and Kretschmer (n 137) 25.

²⁴⁹ DSM Directive, recital 14.

term in this context is “authorised”. The Recital does not state that TDM is permitted wherever content is available on the internet. Instead, it specifies that such mining is allowed only where the rightholder has authorised the use. Public availability on the internet serves as an example of authorisation, not as a substitute for it.²⁵⁰

When applied to a pirated database, Recital 14 provides a clear outcome. The works in such a database are made available to the public online, but without the copyright holder’s authorisation or consent. The operator of the pirate repository, not the rightholder, made the works available, acting without the rightholder’s authorisation and contrary to it. Authorisation that Recital 14 identifies as the basis for lawful access is entirely absent in this context. It is not merely undocumented or uncertain; it is explicitly absent, as the rightholder’s rights were infringed by the act that resulted in public availability.²⁵¹ Therefore, the pirated database does not satisfy the requirement of lawful access set out in Articles 3(1) and 4(1) of the DSM Directive. This is not due to evidentiary difficulties or interpretative doubts, but rather the direct application of Recital 14 own language: there is no authorisation, since the act itself constituted an infringement of the rightholder’s rights.

4.4.2. Known Pirated Datasets in AI Training

The analysis in sub-section 4.4.1 might initially seem to address a purely theoretical scenario: a category of datasets that, while categorically excluded from TDM exceptions in principle, are assumed not to be present in the training data of major AI systems. However, available evidence contradicts this assumption. Multiple large datasets with documented connections to pirate repositories have been incorporated into training corpora used by leading AI developers. Their use has been revealed through academic research and, increasingly, legal proceedings.²⁵²

²⁵⁰ Geiger, Frosio and Bulayenko (n 49) 201.

²⁵¹ Margoni and Kretschmer (n 137) 26.

²⁵² Carlini and others (n 7) 2633, 2638.

Books3 is a dataset comprising approximately 196,000 books collected from the Bibliotik shadow library, a pirate repository specialising in unauthorised digital copies of published works.²⁵³ This dataset has been identified as part of the training corpora for several prominent LLMs, including Meta’s LLaMA models.²⁵⁴ The provenance of Books3 is well-documented and has been explicitly examined in ongoing legal proceedings, such as *Kadrey v Meta Platforms Inc* in the US District Court for the Northern District of California. In this case, plaintiffs allege that Meta’s use of Books3 to train its LLaMA models constitutes copyright infringement on the basis of the dataset’s piracy-derived composition.²⁵⁵

Library Genesis (LibGen) is a mass-scale shadow library that provides access to millions of books and academic articles without the authorisation of copyright holders.²⁵⁶ In contrast to Books3, which is a discrete and identifiable dataset of defined scope, LibGen is a continuously updated repository. It has been identified as the source of numerous AI training datasets for various developers and models, as documented in academic research on large-scale training corpora.²⁵⁷ The vast scale of LibGen, which includes millions of academic and commercial publications, means that its inclusion in a training corpus, even in part, represents a substantial volume of content for which the lawful access requirement under Articles 3(1) and 4(1) of the DSM Directive cannot be satisfied.

LAION-400M and *LAION-5B* are large-scale datasets composed of image-text pairs, compiled by extracting image links and associated text from the Common Crawl corpus, which is widely used for training text-to-image generation models.²⁵⁸ These datasets do not contain the images themselves; rather, they provide links to images hosted elsewhere on the internet along with corresponding text descriptions. Many of the referenced images, however, have been sourced

²⁵³ Henderson and others (n 15) 1, 24.

²⁵⁴ Sobel (n 68) 1, 35.

²⁵⁵ *Kadrey v Meta Platforms Inc*, No 3:23-cv-03417 (ND Cal 2023).

²⁵⁶ Henderson and others (n 15) 25.

²⁵⁷ Carlini and others (n 7) 2639.

²⁵⁸ Henderson and others (n 15) 26.

from platforms and websites without the authorisation of the relevant rightholders. In documented instances, this includes images from stock photography platforms that rely on licensing for their business models.²⁵⁹ In *Getty Images (US) Inc v Stability AI Ltd* in the US District Court for the District of Delaware, the LAION datasets are specifically cited in allegations that Stability AI's image generation models were trained on Getty Images' copyrighted photographs without authorisation.²⁶⁰

The aforementioned cases are all adjudicated under US law and do not serve as precedent for the application of EU copyright law.²⁶¹ Their relevance to this analysis is factual rather than legal, as they provide court-admitted evidence regarding the composition of training datasets used by major AI developers. This evidence demonstrates that the inclusion of pirated content in AI training datasets is a documented reality rather than a hypothetical concern. The legal conclusions reached by US courts under the fair use doctrine in Section 107 of Title 17 of the US Code²⁶², which has no counterpart in EU copyright law, cannot be directly applied to the EU context. What is transferable is the factual scenario: if a dataset such as Books3 was used to train a model deployed in the EU, the categorical exclusion of that dataset's contents from Articles 3 and 4 of the DSM Directive, as outlined in sub-section 4.4.1, would apply to the extent that EU copyright law governs the activity, irrespective of the outcomes of parallel US legal proceedings.²⁶³ As the litigation landscape develops, similar proceedings are likely to arise under the laws of EU Member States concerning these same training datasets.

²⁵⁹ Sobel (n 68) 36.

²⁶⁰ *Getty Images (US) Inc v Stability AI Ltd*, No 1:23-cv-00135 (D Del 2023).

²⁶¹ *Kadrey v Meta Platforms Inc*, No 3:23-cv-03417 (ND Cal 2023); *The New York Times Co v Microsoft Corp*, No 1:23-cv-11195 (SDNY 2023); *Getty Images (US) Inc v Stability AI Ltd*, No 1:23-cv-00135 (D Del 2023).

²⁶² 17 USC § 107.

²⁶³ Sobel (n 68) 37.

4.4.3. Legal Consequences

The combination of the categorical analysis in sub-section 4.4.1 with the documented dataset evidence in section 4.4.2 leads to several legal consequences under EU copyright law that arise directly once the threshold issue of lawful access has been resolved against the AI developer.

If training data includes content obtained from a pirated database, Articles 3 and 4 of the DSM Directive are inapplicable to that content. This outcome results directly from the threshold requirement of lawful access described in sub-section 4.4.1. Since lawful access is not established, further analysis of other exception condition, such as personal scope, material scope, or the absence of a valid opt-out, is unnecessary and irrelevant.²⁶⁴ The AI developer cannot assert qualification as a research organisation under Article 3, nor can they rely on the absence of a rightholder opt-out under Article 4(3) for pirated content. These arguments pertain only to conditions that become relevant after lawful access is established, which is not the case here. The absence of TDM exceptions for pirated content is absolute and cannot be remedied by fulfilling any other conditions.

The operator of a pirated database, defined as the organisation that compiles, hosts, and distributes unauthorised content, commits a primary copyright infringement under EU law. By reproducing works contained in the database and making them available to the public, the operator infringes both the right of reproduction under Article 2 of the InfoSoc Directive and the right of communication to the public under Article 3 of the same Directive.²⁶⁵ This infringement occurs independently of any subsequent use of the database by an AI developer and constitutes a fundamental unlawful act from which the categorical exception analysed in sub-section 4.4.1 is derived.

²⁶⁴ DSM Directive, arts 3(1), 4(1); Geiger, Frosio and Bulayenko (n 49) 202.

²⁶⁵ InfoSoc Directive, arts 2–3.

The actions of the AI developer, specifically, downloading content from a pirated database and incorporating it into a training dataset, constitute an additional and independent act of reproduction within the meaning of Article 2 of the InfoSoc Directive, as established in Chapter 3.²⁶⁶ It is essential to characterise this accurately: the AI developer is not merely a secondary infringer whose liability arises from or depends on an infringement by the dataset operator. The act of downloading and storing content from a pirated database constitutes a direct act of reproduction, performed by the AI developer, of works to which the developer has no lawful access. The AI developer, therefore, is a direct infringer; their liability does not depend on any relationship with the dataset operator, other than reproducing the operator's unlawfully stored content.²⁶⁷

An AI developer may contend that, at the time of training, they were unaware that a particular dataset contained content obtained from a pirated repository. However, under EU copyright law, such good faith does not constitute a defence against liability for copyright infringement, which is assessed objectively. The relevant question is whether the reproduction occurred without a valid exception or licence, rather than whether the individual reasonably believed they had the right to do so.²⁶⁸ Article 13 of the IPRED clarifies the role of good faith: it is relevant to the calculation of damages, allowing a court to distinguish between an infringer who acted knowingly and one who acted unwittingly when determining compensation owed to the rightholder, but it does not negate the fact of infringement.²⁶⁹

Although good faith does not preclude liability, the practical question of whether an AI developer can plausibly claim good faith for the purposes of assessing damages under Article 13 depends on the developer's due diligence regarding the sources of training data. Commercial

²⁶⁶ See Chapter 3, sub-section 3.3.1 of this thesis.

²⁶⁷ Geiger, Frosio and Bulayenko (n 49) 203.

²⁶⁸ Sobel (n 68) 38.

²⁶⁹ Enforcement Directive, art 13.

organisations operating at the scale and with the commercial significance of major AI developers are reasonably expected to exercise a level of diligence concerning the origin of the data used that matches the scale of their operations and commercial interests.²⁷⁰ The use of datasets such as Books3 or LibGen, both of which are well-known, widely discussed in academic literature, and the subject of legal proceedings concerning their use in AI training, is difficult to reconcile with any standard of due diligence that a commercial organisation of significant scale and sophistication could reasonably be expected to meet. This is particularly relevant for any AI developer who continued to use or failed to remove such datasets from their training pipeline after the pirated nature of these datasets became public knowledge through legal proceedings or academic publications. For such developers, the good-faith defence relevant to compensation for damages under Article 13 is, at minimum, substantially weakened and may, in some cases, be entirely unavailable.²⁷¹

4.5. Common Provisions of Articles 3 and 4

The analysis thus far has treated Articles 3 and 4 of the DSM Directive largely as parallel but distinct frameworks. Article 3 applies to research organisations without an opt-out provision, while Article 4 is available to any person, provided they may opt out. Three further issues arise concerning both provisions, necessitating analysis of their shared characteristics and differences: the permissible duration and scope of retention of copies made for TDM purposes; the interaction between the TDM exceptions and the sui generis database right; and the relationship between the TDM exceptions and contractual restrictions, including the differential treatment of Articles 3 and 4 under the contractual override prohibition of Article 7. Each of these issues directly affects the practical operation of the TDM framework for AI training and will be examined in turn.

²⁷⁰ Henderson and others (n 15) 27.

²⁷¹ Sobel (n 68) 39.

4.5.1. Retention and Storage of Copies

Article 3(2) of the DSM Directive expressly provides that copies of works made for the purposes of TDM under Article 3(1) may be stored for scientific research purposes, including verification of research results.²⁷² In contrast, Article 4 lacks an explicit provision regarding the storage of copies made under its exception. This asymmetry is not simply a drafting oversight; rather, it presents a significant issue of legal interpretation concerning the temporal scope of the Article 4 exception, which the Directive does not explicitly resolve.²⁷³

For beneficiaries of Article 3 of the DSM Directive, the situation is relatively clear: copies made for TDM may be stored as long as necessary to verify the research results derived from them. This reflects a widely recognised principle of scientific methodology: reproducibility of research results depends on researchers, including third parties, having access to the raw data underlying the conclusions.²⁷⁴ Therefore, a research organisation conducting TDM under Article 3 may retain its training dataset beyond the immediate analytical process, provided retention serves the verification function.

The absence of a comparable provision in Article 4 of the DSM Directive raises the issue of whether an AI developer may retain a training dataset compiled under this exception, and for what duration, following completion of the TDM process. Two competing arguments emerge in this context.

The first argument maintains that the exception under Article 4 of the DSM Directive applies solely to the act of TDM, specifically the reproduction of works for analytical purposes, and that retaining copies beyond what is necessary for this process exceeds the scope of the exception. According to this perspective, once the model has been trained, the exception's purpose is fulfilled, and continued storage of the training corpus constitutes reproduction outside the scope

²⁷² DSM Directive, art 3(2).

²⁷³ Geiger, Frosio and Bulayenko (n 49) 174, 204.

²⁷⁴ Margoni and Kretschmer (n 137) 27.

of Article 4 of the DSM Directive, requiring an independent legal basis, such as a licence, to be lawful.²⁷⁵ This argument is grounded in the general principle that exceptions to exclusive rights should not extend beyond what is necessary to achieve their intended purpose.

The second argument emphasises the practical realities of AI development and deployment. Retraining, fine-tuning, auditing, and rectifying identified issues in a deployed model, including addressing memorisation issues such as those discussed in sub-section 3.3.4 or responding to a valid opt-out exercised after initial training, may require ongoing access to the original training corpus.²⁷⁶ From this view, retaining the training dataset is not a separate activity distinct from the TDM process, but rather a continuous practical necessity for the responsible operation of the trained model. Accordingly, retention should be regarded as an integral part of TDM activities, which Article 4 is intended to facilitate.

This thesis does not consider either argument to be clearly resolved by the text of the DSM Directive and regards this as a genuine interpretive gap rather than a question with a settled answer. It can be stated with greater certainty that the absence of an express provision, in contrast to the explicit provision contained in Article 3(2) of the DSM Directive, is itself a source of legal uncertainty for the beneficiaries of Article 4, as they cannot point to an equivalent textual basis for retention and must instead rely on arguments by analogy or on the general purpose of the exemption. This uncertainty is yet another, comparatively minor, illustration of the broader issue, examined in more detail in Section 4.6, of provisions designed with the research context of Article 3 in mind but overlooked in the commercial context of Article 4.²⁷⁷

A comparable asymmetry exists in respect of security obligations. Article 3(2) of the DSM Directive expressly permits rightholders to implement measures to ensure the security and in-

²⁷⁵ Geiger, Frosio and Bulayenko (n 49) 205.

²⁷⁶ See Chapter 3, sub-section 3.3.4 of this thesis; see Chapter 4, sub-section 4.3.5 of this thesis.

²⁷⁷ Margoni and Kretschmer (n 137) 28.

tegrity of their networks and databases, provided these measures comply with the proportionality restrictions discussed in sub-section 4.2.4. Such measures must not exceed what is necessary to achieve the security objective and may not serve as a pretext for blocking lawful TDM.²⁷⁸ In contrast, Article 4 of the DSM Directive does not contain an explicit provision regarding security measures. Its absence, however, does not necessarily preclude rightholders from applying security measures to content within the scope of Article 4. General principles of lawful protection of network infrastructure likely apply regardless of the specific TDM exception. Yet it should be noted that the proportionality framework explicitly governing Article 3 has no direct textual counterpart in Article 4, again reflecting the comparative absence of provisions addressing the operational detail of the commercial TDM context.²⁷⁹

Another practical issue concerns the role of third-party infrastructure providers, primarily cloud computing platforms, whose servers often store training datasets and whose computational infrastructure is frequently used for the training process. Neither Article 3 nor Article 4 of the DSM Directive explicitly addresses this scenario, but analysis can be derived from the general structure of the exemptions.²⁸⁰

When a cloud infrastructure provider processes training data on behalf of an AI developer, storing datasets or performing computations for training, strictly at the direction of and for the purposes of the AI developer's TDM activity, it is reasonable to conclude that the TDM exception available to the AI developer extends to processing conducted on the cloud service provider's infrastructure, on the basis that the cloud provider acts as a processor executing the AI developer's TDM activities, rather than as an independent entity.²⁸¹ This analysis draws on the distinction between data controllers and data processors established in EU data protection law,

²⁷⁸ DSM Directive, art 3(2); see sub-section 4.2.4 of this thesis.

²⁷⁹ Geiger, Frosio and Bulayenko (n 49) 206.

²⁸⁰ Margoni and Kretschmer (n 137) 29.

²⁸¹ Margoni and Kretschmer (n 137) 30.

applied here by analogy. The entity determining the purposes of TDM, namely the AI developer, is the relevant party for classification under Article 3 or Article 4, while the infrastructure provider's role is instrumental to the developer's activities.

This extension, however, is subject to certain limitations. If a cloud infrastructure provider engages in activities that exceed the processing role described above, such as independently compiling, curating, or using a dataset beyond the scope of services provided to the AI developer, the provider's (own) activities require an independent legal basis under Article 3 or Article 4 of the DSM Directive. This assessment must consider the provider's individual qualifications and purposes, which may differ from those of the AI developer on whose behalf the initial processing was conducted.²⁸² The TDM exception does not apply to data or infrastructure in themselves; rather, it applies to the relevant activities of a qualified entity. An infrastructure provider that assumes a role beyond a purely instrumental one in data processing is therefore excluded from the protection afforded by the exception for the AI developer exception.

4.5.2. Interaction with the Database Right

The Database Directive establishes, alongside any copyright that may exist in relation to the contents of a database, a separate *sui generis* right protecting databases in which substantial resources have been invested in the collection, verification or presentation of the contents.²⁸³ This right is conceptually distinct from copyright: it protects investment (the resources expended in compiling and maintaining the database) rather than creativity or originality of expression, and grants the database maker the right to prevent the extraction or re-utilisation of the whole or a substantial part of the database's contents.²⁸⁴ Thus, a database may be protected by the *sui generis* right, even if its individual contents or structure do not meet the threshold of originality for copyright protection discussed in Chapter 3 — and conversely, individual works

²⁸² Geiger, Frosio and Bulayenko (n 49) 207.

²⁸³ Database Directive, art 7(1).

²⁸⁴ Peter Bernt Hugenholtz, 'The Database Right after Ryanair' (2020) 51 IIC 1019, 1021.

within a database may be independently protected by copyright regardless of whether the database as a whole meets the criteria for protection by a specific right.

The relevance of database rights to AI training is evident: a training dataset may itself be composed of a database or constitute a database within the meaning of the Database Directive, and the reproduction or extraction of substantial parts of such a database for TDM purposes may constitute an infringement of the *sui generis* right, irrespective of any infringement of copyright in individual elements of the database.

The DSM Directive directly addresses this potential gap. Both Article 3(1) and Article 4(1) refer to reproductions and extractions made “pursuant to” the Database Directive, in addition to their primary reference to the right of reproduction under Article 2 of the InfoSoc Directive.²⁸⁵ This wording extends the TDM exceptions to cover not only copyright infringement but also infringement of the *sui generis* database right: a TDM operation which would otherwise constitute unauthorised extraction or re-utilisation of a substantial part of a protected database, whilst complying with the conditions of Article 3 or Article 4 of the DSM Directive, is exempt from the database right in the same way as it is exempted from the copyright reproduction right.²⁸⁶

This extension is a deliberate and significant feature of the TDM framework. Without it, a rightholder could potentially circumvent the TDM exceptions by relying on the database right rather than copyright: even if an AI developer satisfied all the conditions of Article 4 of the DSM Directive in relation to copyright in the content of a database, the rightholder could still object to the extraction of that content on the basis of a separate *sui generis* right, which would render the copyright exception ineffective in respect of any content organised as a protected database.²⁸⁷ By extending the exceptions to cover the database right as well, the DSM Directive

²⁸⁵ DSM Directive, arts 3(1), 4(1).

²⁸⁶ Geiger, Frosio and Bulayenko (n 49) 208.

²⁸⁷ Hugenholtz (n 284) 1023.

ensures that the TDM exceptions provide a comprehensive framework for the relevant acts of reproduction and extraction, covering both of the fundamental rights that might otherwise be engaged.

4.5.3. Article 7 — Contractual Override Prohibition

Article 7(1) of the DSM Directive stipulates that any contractual provision conflicting with the exceptions in Articles 3, 5, and 6 is unenforceable.²⁸⁸ The omission of Article 4 from this list is intentional and signifies a deliberate structural distinction. Article 3 is fully protected against the nullification of contractual provisions, whereas Article 4 is subject to a different regulatory mechanism: the possibility of refusal under Article 4(3).²⁸⁹

For entities that benefit from Article 3 of the DSM Directive, the application of Article 7(1) is clear, as discussed in sub-section 4.2.4. Any clause in the terms of service that seeks to prohibit scraping, automated access, or TDM is unenforceable against a qualified research organisation if it conflicts with the exception provided in Article 3.²⁹⁰ Even if a contractual provision attempts to prohibit such activities, the rightholder's remedy is limited to proportionate security measures permitted under Article 3(2), as outlined in sub-sections 4.2.4 and 4.5.1, and does not extend to contractually excluding the exception itself.²⁹¹

The situation for beneficiaries of Article 4 of the DSM Directive is more complex, since Article 7(1) does not apply to protect Article 4 in the same blanket manner. This complexity prompts the question of whether a rightholder may use contractual provisions, such as terms of service prohibiting TDM, to restrict TDM under Article 4 in ways that Article 7(1) would preclude for Article 3.²⁹²

²⁸⁸ DSM Directive, art 7(1).

²⁸⁹ Senftleben (n 146) 1, 13.

²⁹⁰ DSM Directive, art 7(1); see sub-section 4.2.4 of this thesis.

²⁹¹ DSM Directive, art 3(2); see sub-sections 4.2.4 and 4.5.1 of this thesis.

²⁹² Senftleben (n 146) 14.

The most persuasive interpretation is that rightholders cannot, or should not, circumvent the specific mechanism established by Article 4(3) of the DSM Directive for this purpose. The structure of the DSM Directive demonstrates a deliberate policy choice: instead of fully shielding Article 4 from contractual restrictions as Article 7(1) does for Article 3, the Directive provides rightholders with a dedicated tool, namely a machine-readable opt-out under Article 4(3), through which any restriction on TDM under Article 4 must be expressed.²⁹³ A general contractual prohibition on TDM in a website's terms of service, applicable to all users regardless of compliance with the machine-readable opt-out requirement, does not utilise this designated tool. Instead, it seeks to impose a restriction through ordinary contract law that the Directive intends to regulate through a different and more demanding mechanism.

The present analysis distinguishes between two categories of restrictions that may initially appear similar but are treated differently under the DSM Directive. A machine-readable opt-out under Article 4(3) constitutes an established and lawful mechanism by which a rightholder may render Article 4 inapplicable to specific content; it operates within the Directive's framework, is content-specific as discussed in sub-section 4.3.4, and its legal effect is expressly provided by Article 4(3). In contrast, a blanket prohibition in the terms of service, such as a statement that "no text or data mining of any kind is permitted on this website", operates outside this framework. Such a prohibition does not take the machine-readable form required by Article 4(3), lacks the content-specificity permitted by the opt-out mechanism, and its legal effect, if any, is not addressed in Article 4, since Article 4(3) specifically and exhaustively defines the means by which rightholders may restrict text and data mining in accordance with Article 4.²⁹⁴

²⁹³ Geiger, Frosio and Bulayenko (n 49) 210.

²⁹⁴ DSM Directive, art 4(3); Margoni and Kretschmer (n 137) 32.

Accordingly, a blanket terms-of-service prohibition on TDM should be considered legally ineffective to the extent that it seeks to restrict activities covered by Article 4 of the DSM Directive. This is not because Article 7(1) extends protection to Article 4, which it does not, but because Article 4(3) prescribes the manner in which restrictions on Article 4 TDM are to be expressed. A general contractual term that fails to meet the requirements of Article 4(3) does not affect the availability of the Article 4 exception. The rightholder seeking to restrict Article 4 TDM in respect of their content must use the mechanism the Directive provides for that purpose. If a rightholder has not done so, whether by relying solely on a blanket contractual prohibition or by omitting any mechanism, Article 4 remains applicable regardless of what the terms of service may state.²⁹⁵

4.6. Are TDM Exceptions Sufficient for AI?

The preceding Sections of this Chapter have examined the TDM exemptions from the DSM Directive in both descriptive and analytical terms, including what they say, who they cover, what conditions they impose, and how those conditions apply to the specific characteristics of AI training. At several points, that analysis has flagged gaps, tensions, and asymmetries: a calibration gap between the political context of 2016–2019 and the current reality of AI training (sub-section 4.1.2); the absence of a mandatory technical standard for opt-out in accordance with Article 4(3) of the DSM Directive (sub-section 4.3.4); the practical impossibility of enforcing opt-outs against AI developers (sub-section 4.3.4); the lack of provisions addressing pre-opt-out training data (sub-section 4.3.5); and the asymmetrical calibration of Articles 3 and 4 regarding storage and security measures (sub-section 4.5.1). This section synthesises these findings into a structured critical assessment focused on the central question: whether the TDM

²⁹⁵ Senftleben (n 146) 15.

exceptions, in their current form, provide a sufficient legal basis for training AI in accordance with EU copyright law.

A fundamental observation underlying this critical assessment concerns the concept of scale. It is necessary to define this precisely, as scale, considered solely as a quantitative indicator, does not affect the legal analysis. As shown in sub-section 4.3.3, conducting an activity on a large scale does not justify non-compliance with legal requirements; scale alone does not modify the obligations imposed by law. The scale argument presented here instead addresses what the law was intended to consider when formulating its requirements for AI training, rather than what the law currently mandates.²⁹⁶

The provisions of the DSM Directive on TDM originate from the Commission's 2016 proposal and were adopted in 2019 within the policy context discussed in sub-section 4.1.2, which primarily addresses academic researchers working with scientific journal databases, business analysts processing market data, and journalists analysing publicly available records.²⁹⁷ These activities are generally conducted by identifiable institutional entities using datasets measured in thousands or millions of documents, for purposes that are relatively transparent and often publicly documented in advance, such as research projects, market analyses, or investigative reports. In contrast, modern training of foundational models operates at a vastly different scale: datasets comprising hundreds of billions of tokens, collected from across the internet, are processed by organizations whose training activities are not disclosed publicly in advance, and whose resulting models, sometimes valued at tens of billions of euros, are deployed for commercial purposes by hundreds of millions of users.²⁹⁸ Recital 8 of the DSM Directive presents

²⁹⁶ See Chapter 4, sub-section 4.3.3 of this thesis.

²⁹⁷ European Commission, COM(2016) 593 final (n 130); see Chapter 4, sub-section 4.1.2 of this thesis.

²⁹⁸ Carlini and others (n 7) 2633, 2640.

TDM as a tool for research and innovation; there is no indication in the legislative documentation that the legislator anticipated activities of the scale and nature described above.²⁹⁹

This analysis describes a description of a mismatch rather than an accusation of bad faith. Legislators cannot be expected to anticipate the precise development of technologies that do not yet exist in their eventual form. The TDM provisions of the DSM Directive were not inadequately drafted for the issues they were intended to address; rather, they were designed to solve a different problem.³⁰⁰ However, when these provisions are applied, as noted in sub-section 4.1.4, to activities unforeseen by their drafters, the alignment between the legal framework and the activities it now governs is, at best, imperfect. The remainder of this section identifies three specific areas where this imperfection constitutes a structural inadequacy, before proceeding to assess the implications of this mismatch for the lawfulness of AI training under the current legal framework.

Gap 1 — The Absence of a Compensation Mechanism Under Article 4

The first structural shortcoming concerns compensation, specifically its complete absence under Article 4 of the DSM Directive. Neither Article 3 nor Article 4 of the DSM provides for any payments to rightholders whose works are used for TDM.³⁰¹ In the case of Article 3, this absence is justified by the underlying rationale: research organisations conducting non-commercial scientific research do not derive commercial benefit from rightholders' works in a manner that has traditionally required compensation under the broader copyright system. Furthermore, the public benefit of research has long been recognised as a basis for non-remunerative exceptions in both EU and national copyright legislation.³⁰²

²⁹⁹ DSM Directive, recital 8.

³⁰⁰ Senftleben (n 146) 1, 16.

³⁰¹ DSM Directive, arts 3, 4.

³⁰² Geiger, Frosio and Bulayenko (n 49) 174, 211.

The situation under Article 4 of the DSM Directive differs significantly, and this distinction is the central point. Article 4 extends the same uncompensated structure to commercial entities, including, as examined in this thesis, large AI developers that use training data to develop commercially deployed foundational models.³⁰³ The rationale for non-remunerated use in research (non-commercial purpose, public benefit, and absence of commercial gain) does not apply in the same manner to Article 4. Commercial AI developers training foundational models are, by definition, engaged in activities that generate commercial gain. The resulting models are commercial products, often of considerable value, and the training data constitutes a necessary input for their creation, comparable to other inputs such as computing infrastructure, energy, or skilled labour, for which developers are required to pay.³⁰⁴ Thus, training data is the only input for AI development for which developers are not obliged to compensate the supplier, despite that input being, in many respects, the most distinctive and valuable contribution to the capabilities of the final product.

A relevant comparison is the private copying exception found in the legislation of most EU Member States. This exception permits individuals to make copies of copyright-protected works for personal use without infringing copyright, but in most Member States, it is accompanied by a compensation mechanism, typically a levy on media or recording devices, which is collected and distributed to rightholders through collective management societies.³⁰⁵ The EU legal tradition is therefore familiar with the “exception plus compensation” model, where the activity is permitted without individual licensing, but rightholders collectively receive financial compensation reflecting the aggregate value of the permitted use. Article 4 of the DSM Directive diverges from this established model by providing an exception without compensation

³⁰³ DSM Directive, art 4(1); see Chapter 4, sub-section 4.3.2 of this thesis.

³⁰⁴ Sobel (n 68) 1, 40.

³⁰⁵ Margoni and Kretschmer (n 137) 33.

in the context of commercial AI training, where the aggregate economic value of the permitted use significantly exceeds that of private copying, which itself justified compensation.

The absence of compensation is not an oversight in the drafting of the DSM Directive. As demonstrated in sub-sections 4.1.1 and 4.3.1, it represents a conscious legislative decision, made as part of the political compromise that resulted in the adoption of Article 4, with the objective of fostering the data economy by reducing transaction costs for TDM activities.³⁰⁶ Although this choice was intentional, its consequences in the specific context of large-scale commercial AI training (a context the legislature did not have centrally in view, as established earlier) produce an outcome that is difficult to justify on grounds of fairness. AI developers are able to obtain substantial commercial benefit from the copyrighted works of others on a large scale, while those whose works are used receive no compensation. This result arises not because the legislature determined it to be fair, but because, in making the relevant decision, the legislature was addressing a different issue.³⁰⁷

Gap 2 — The Absence of a Binding Technical Standard for the Opt-Out

The second structural gap was examined in detail in sub-section 4.3.4 and is restated here for the purposes of the overall critical assessment. Article 4(3) of the DSM Directive makes the existence of an exception to Article 4 conditional upon the absence of a reservation by the rightholder, “expressly reserved [...] in an appropriate manner, such as machine-readable means”, but neither the Directive itself nor any subsequent EU document specifies the form such a reservation must take in order to be legally effective.³⁰⁸ The consequence is a fragmented landscape in which multiple technical mechanisms exist, none of which has been confirmed as

³⁰⁶ See Chapter 4, sub-sections 4.1.1 and 4.3.1 of this thesis.

³⁰⁷ Geiger, Frosio and Bulayenko (n 49) 212.

³⁰⁸ DSM Directive, art 4(3); see Chapter 4, sub-section 4.3.4 of this thesis.

constituting a valid Article 4(3) reservation, and in which rightholders, AI developers, and ultimately courts lack an authoritative benchmark for determining whether any given mechanism satisfies the statutory requirement.³⁰⁹

The significance of this gap for the present assessment lies not so much in the fragmentation itself as in its distributional consequences. Legal uncertainty of this kind is rarely neutral in its practical effects; it tends to work in favour of the party for whom inaction or the continuation of existing behaviour constitutes a failure to fulfil obligations, and to the detriment of the party who must act positively to remedy that failure.³¹⁰ In the event of the opt-out under Article 4(3) of the DSM Directive, Article 4 applies by default — TDM is permitted unless otherwise provided for by law. If the validity of a particular reservation mechanism is unclear, the AI developer may reasonably act on the basis that the exception applies, as the burden of proving that a valid reservation existed and was not complied with falls on the rightholder seeking to challenge such conduct. The rightholder, by contrast, bears the costs and risks of this uncertainty: they must choose a reservation mechanism without knowing whether it will be recognised as legally effective, and, if it is not complied with, they bear the burden of litigation to prove that their mechanism was valid and that it was not complied with. In other words, the absence of a mandatory standard does not affect both parties symmetrically; instead, it systematically favours the party whose conduct continues by default.³¹¹

As identified in Section 4.3.4, this is, among the three gaps examined in this Section, the most readily remediable: the technological building blocks for a standardised, machine-readable, cryptographically secure format for reserving rights already exist, as demonstrated by the C2PA specification, and it is not technology that is lacking, but regulatory recognition.³¹² A Commission implementing act defining a recognised technical standard for the reservation under Article

³⁰⁹ See Chapter 4, sub-section 4.3.4 of this thesis.

³¹⁰ Senftleben (n 146) 17.

³¹¹ Henderson and others (n 15) 1, 28.

³¹² See Chapter 4, sub-section 4.3.4 of this thesis.

4(3) of the DSM Directive, whether based on C2PA or another existing or future standard, would not require amendments to the text of the DSM Directive and could relatively quickly resolve a significant portion of the uncertainty described in this Section, compared to the legislative reform that other gaps might require.

Gap 3 — The Absence of an Enforcement Mechanism

The third structural gap pertains to enforcement. Even if an effective and technically well-defined opt-out mechanism existed, addressing Gap 2, and even assuming compensation had been introduced, addressing Gap 1, the DSM Directive still lacks a mechanism to verify or ensure compliance with these provisions at the scale required for AI training.³¹³

The DSM Directive does not require AI developers to register TDM activities, disclose the sources of training data, or document steps taken to verify lawful access or compliance with opt-outs.³¹⁴ Consequently, as discussed in sub-section 4.3.4, rightholders seeking to determine whether their work was used to train a specific AI model (a prerequisite for bringing any claim, including for breach of opt-out or use of pirated source material as stipulated in Section 4.4) have no means of doing so except through voluntary disclosure by the AI developer or through a burdensome and uncertain litigation process.³¹⁵

Article 53 of the EU AI Act partially addresses this gap by requiring providers of general-purpose AI models to publish sufficiently detailed descriptions of the content used for training.³¹⁶ However, as noted in sub-section 4.3.4, such descriptions are insufficient for rightholders to determine with certainty whether their specific works have been used. The transparency obligation under the AI Act serves a different purpose, focusing on the public interest in understanding AI training data for reasons of safety, accountability, and public trust, rather than

³¹³ Geiger, Frosio and Bulayenko (n 49) 213.

³¹⁴ DSM Directive, arts 3, 4.

³¹⁵ See Chapter 4, sub-section 4.3.4 and Section 4.4 of this thesis.

³¹⁶ EU AI Act, art 53.

providing rightholders with the information necessary to enforce their rights under the DSM Directive.³¹⁷ These instruments were not designed to function together for this purpose. Consequently, even with the AI Act's transparency provisions, the enforcement gap identified here remains largely unaddressed.

The practical consequence of Gap 3 is that the TDM framework, even where its substantive conditions could be breached, such as through the use of content covered by a valid opt-out or unlawfully obtained content (Section 4.4), relies for its enforcement entirely on rightholders' ability to detect such breaches and bring individual claims. Such an approach is inadequate for activities involving datasets so large that individual verification is impractical, and for AI developers who are not subject to disclosure obligations tailored to this context.³¹⁸ Implementing a registration or notification system specifically for large-scale commercial content data management, which would require developers exceeding a certain threshold to document, in an accessible form, the categories of sources used and the steps taken to verify lawful access and compliance with exceptions, would address this gap in a manner that Article 53 of the AI Act, as currently framed, does not.

This analysis of this Chapter is grounded in the foundations established in sub-sections 4.1.4 and 4.3.2, specifically that the wording of Article 4 of the DSM Directive is sufficiently broad to encompass large-scale commercial AI training. This textual interpretation must be applied even though the legislative intent, as discussed in this Chapter, did not target activities of this nature or scale.³¹⁹ Having identified three distinct structural gaps resulting from the application of this broad yet contextually misaligned provision to AI training, it is necessary to consider how the combination of broad statutory language, narrower legislative intent, and structural gaps affects the overall assessment of the adequacy of the TDM framework.

³¹⁷ Henderson and others (n 15) 29.

³¹⁸ Sobel (n 68) 41.

³¹⁹ See Chapter 4, sub-sections 4.1.4 and 4.3.2 of this thesis.

In fact, this combination results in a legal basis that is formally valid but substantively hollow. According to the analysis presented, Article 4 of the DSM Directive provides commercial AI developers with a legal foundation for TDM, thereby fulfilling the formal requirement to identify an applicable exception. However, when applied to AI training, it lacks essential safeguards such as compensation, a functional opt-out mechanism, and effective enforcement, which would typically accompany a provision specifically designed for such activities. While the legal form exists, the practical mechanisms that would protect the interests of rightholders are largely absent. This is not a scenario in which AI training is simply unlawful, as sub-sections 4.1.4 and 4.3.2 confirm the existence of a statutory framework. Nor does the TDM framework directly and adequately regulate AI training, as demonstrated by the structural gaps identified. Instead, the law applies but does not function as would be expected for legislation intended to regulate activities of this scale and nature.³²⁰

It is important to reiterate that this shortcoming does not result from bad faith or poor drafting by the EU legislator, for the reasons stated above. The legislator addressed this issue between 2016 and 2019, and the TDM provisions represent a reasonable response to the challenges identified at that time. The gap identified in this chapter is best understood as a consequence of technological advancement outpacing legislative cycles. This phenomenon is not unique to copyright law but is particularly significant in the context of AI, given the rapid evolution of AI capabilities and training methods compared to the slower pace of EU legislative reform.³²¹ Thus, the DSM Directive was not designed to fail in the context of AI training; rather, it was created for a context that has since been surpassed by the current scale and form of AI training.

³²⁰ Hugenholtz (n 35) 3, 9.

³²¹ Margoni and Kretschmer (n 137) 34.

4.7. Answer to Research Question 2

The analysis conducted across this chapter permits a direct answer to the second research question posed by this thesis: to what extent do the TDM exceptions of the DSM Directive permit the use of copyright-protected works for AI training?

The answer varies according to both the subject category and the type of training data considered. Article 3 of the DSM Directive, as demonstrated in Section 4.2, establishes a robust and mandatory exception for research-oriented activities conducted by research organisations and cultural heritage institutions, contingent upon lawful access. For AI research at universities, research institutes, and similar entities, Article 3 offers substantial legal certainty, provided the activity aligns with the primary purpose and non-commercial requirements of Article 2(1) (sub-section 4.2.1) and is genuinely directed toward scientific research rather than commercial product development (sub-section 4.2.2). Commercial AI developers are categorically excluded from Article 3, as their primary purpose is product development and commercial deployment. This exclusion applies regardless of the technical complexity of their research or any structural arrangements intended to leverage the exemption.³²²

For commercial AI developers, Article 4 of the DSM Directive, discussed in Section 4.3, constitutes the sole legal basis under EU copyright law for large-scale training of foundational models. Sub-section 4.3.2 demonstrates that Article 4 encompasses commercial AI training, despite a divergence between its textual scope and the narrower legislative intent. This scope is subject to two independent conditions that limit its practical applicability. First, lawful access to training data must be ensured (sub-section 4.3.3); this requirement is not satisfied if data is sourced from pirate repositories, regardless of other Article 4 conditions (sub-section 4.4). Second, rightholders must not have exercised a valid machine-readable opt-out under Article 4(3) for the content used. The practical impact of this condition, detailed in sub-section 4.3.4, is

³²² DSM Directive, arts 2(1), 3(1); see Chapter 4, sub-sections 4.2.1 and 4.2.2 of this thesis.

diminished by the lack of a mandatory technical standard, significant compliance challenges, and documented gaps in AI developers' compliance practices.³²³

Collectively, these findings provide a comprehensive answer to Research Question 2. The TDM exceptions of the DSM Directive offer a partial legal framework for current AI training practices. Article 3 establishes a strong foundation for academic AI research but excludes commercial development. Article 4 presents a broad textual basis for commercial AI training, yet its scope is limited by conditions: it does not cover training data from pirated sources, which are entirely excluded (Section 4.4), nor, in principle, content subject to a valid opt-out that is not respected. The practical relevance of this clarification depends on the effectiveness of the opt-out mechanism (sub-section 4.3.4). This effectiveness, as shown in Section 4.6, is significantly compromised by the absence of a mandatory technical standard and documented enforcement gaps. Overall, the TDM framework formally encompasses a substantial portion of current AI training activities, but leaves significant categories (particularly training on pirated sources and, to a lesser extent, post-reservation training) without sufficient legal basis under EU copyright law.³²⁴

The answer to Research Question 2 reached in this chapter identifies the opt-out mechanism under Article 4(3) of the DSM Directive as the pivotal feature of the TDM framework's adequacy as applied to commercial AI training. This mechanism is intended to balance the broad TDM rights granted to commercial entities under Article 4 of the DSM Directive against the interests of rightholders. However, as concluded in Section 4.6, its practical effectiveness is considerably below what is required for this balancing function. Chapter 5 will examine the opt-out mechanism in detail, considering both its technical implementation across the mechanisms discussed in Section 4.3.4 and its broader structural adequacy, including its relationship

³²³ DSM Directive, arts 4(1), 4(3); see Chapter 4, sub-sections 4.3.2, 4.3.3, 4.3.4 and Section 4.4 of this thesis.

³²⁴ See Chapter 4, Section 4.6 of this thesis.

with the emerging market for direct and collective licensing of training data.³²⁵ This examination directly addresses the Research Question 3: to what extent does the opt-out mechanism effectively balance these competing interests?

³²⁵ See Chapter 4, sub-section 4.3.4 and Section 4.6 of this thesis.

5. The Opt-Out Mechanism and Licensing Practices

5.1. The Opt-Out as Policy Instrument

Chapter 4 examined Article 4(3) of the DSM Directive as a condition of statutory application, specifically whether the exception applies in particular circumstances. This Chapter addresses a different issue: whether the opt-out mechanism, as part of the regulatory framework, is capable of fulfilling its intended purpose. As established in sub-section 4.3.1, the opt-out mechanism constituted a central concession that rendered the broad personal scope of Article 4 politically acceptable. This chapter evaluates the adequacy of such a concession.

The selection of a default opt-out option reflects the market facilitation logic discussed in sub-section 4.1.2: TDM is permitted without prior individual negotiations, thereby avoiding the transaction costs that an opt-in requirement would impose by necessitating consent in each case.³²⁶ A consent-based approach, in which TDM is allowed only with the rightholder's explicit consent, would reintroduce these costs by requiring AI developers to seek and obtain permission before mining each individual work. The drafters of the DSM Directive chose the opposite allocation: permission is the default, while objection requires affirmative action.³²⁷

This default allocation is a deliberate policy choice. Ronald Coase's analysis of transaction costs demonstrates that when negotiations are costless, the initial allocation of rights does not affect the final outcome, as parties can negotiate freely.³²⁸ However, transaction costs are never zero, and when they are significant, the initial allocation determines which party bears the costs of achieving an alternative outcome. Under Article 4(3) of the DSM Directive, a rightholder seeking to prevent TDM must bear the costs of expressing this objection, including the identification, implementation, and maintenance of the reservation mechanism. In contrast, under a

³²⁶ See Chapter 4, sub-section 4.1.2 of this thesis.

³²⁷ DSM Directive, art 4(3); see Chapter 4, sub-section 4.3.1 of this thesis.

³²⁸ R H Coase, 'The Problem of Social Cost' (1960) 3 *Journal of Law and Economics* 1, 15–19.

voluntary participation model, such costs would fall on the AI developer seeking permission. The DSM Directive assigns these costs to the market side with the largest number of dispersed and individually under-resourced participants, a point examined further in section 5.3.1.³²⁹

A brief comparative observation is useful. US copyright law does not contain any TDM-specific exceptions; instead, AI training is evaluated under the general fair use doctrine of Section 107 of Title 17 of the US Code.³³⁰ Fair use operates as a default authorisation subject to ex post judicial review, functioning as an analogue to opt-out, except here the “opt-out” takes the form of litigation rather than a technical signal. Two conclusions follow. First, the EU’s adoption of a default authorisation model is not an outlier; major jurisdictions converge on it. Second, the EU’s mechanism is, in principle, more efficient, since a preliminary technical signal is less costly than ex post litigation, provided it operates effectively. Whether it does is the question this Chapter examines.³³¹

The remainder of this chapter tests that conditional. Sections 5.2 and 5.3 evaluate whether the implementation of the opt-out mechanism, considering both its technical form and the structural position of the parties who must to use it, achieves the efficiency promised in principle, or whether the disparity between design and practice undermines its effectiveness.

5.2. Technical Implementation of the Opt-Out

Sub-section 4.3.4 addressed the issue of legal validity, specifically whether robots.txt, the X-Robots-Tag, embedded metadata, and C2PA can constitute a valid opt-out under Article 4(3) of the DSM Directive, concluding that the lack of a mandatory standard introduces uncertainty for each mechanism. This Section shifts focus to a different issue: assuming any of these mechanisms could be recognised, how extensively are they actually used, by whom, and with what

³²⁹ Geiger, Frosio and Bulayenko (n 49) 174, 215.

³³⁰ 17 USC § 107.

³³¹ Sobel (n 68) 1, 42.

outcomes? A mechanism that may satisfy Article 4(3) in principle offers little practical value if the intended rightholders do not or cannot use it. Therefore, the following analysis examines implementation and practical application, rather than legal validity, which is considered resolved in Chapter 4 and is not revisited here.

5.2.1. Robots.txt — From Niche Signal to Default Battleground

The general-purpose use of robots.txt is nearly universal and has persisted for several decades. In contrast, AI-specific directives are a recent and rapidly evolving development, emerging primarily since 2023 as the AI training controversy gained prominence.³³² The scale of adoption since then has been considerable. By mid-2025, more than 2.5 million websites had configured their robots.txt files to disallow AI training crawlers entirely, and approximately one in five sites specifically blocked OpenAI’s GPTBot.³³³ Cloudflare, which provides infrastructure for a substantial portion of the web, made blocking of declared AI crawlers the default setting for newly registered domains from July 2025, and reported that its customers blocked over 400 billion requests from AI bots in the subsequent five months.³³⁴

Adoption, however, does not equate to effect. Compliance with robots.txt directives remains voluntary, as established in sub-section 4.3.4, and empirical evidence demonstrates that this voluntary nature results in significant non-compliance. A study conducted in November 2025 found that 72% of UK business websites that were tested experienced at least one instance of an AI crawler disregarding an explicit disallow directive in the robots.txt file. Cloudflare also

³³² Internet Engineering Task Force, *Robots Exclusion Protocol*, RFC 9309 (2022); see Chapter 4, sub-section 4.3.4 of this thesis.

³³³ Coronium, ‘The Closing Web in 2026: How AI Crawler Blocking and Pay-Per-Crawl Changed Web Scraping’ (2026); Digital Applied, ‘AI Crawler & Bot Traffic 2026: The Data’ (2026).

³³⁴ Coronium (n 333); Digital Applied (n 333).

documented that a major AI-based search provider employed undeclared crawlers that intentionally modified user agents and IP address ranges to bypass exclusion rules.³³⁵ Reports indicate that instances of robots.txt circumvention increased by approximately 400% between the second and fourth quarters of 2025 alone.³³⁶ Consequently, even rightholders who have correctly implemented the robots.txt disallow directive, having addressed the legal uncertainty in the most favourable way by using the most widely recognised mechanism, cannot be assured of compliance. This issue is directly related to the enforcement gap discussed in sub-section 5.3.3.

A further complication is what might be called the moving-target problem. The number of named AI crawlers (such as GPTBot and OAI-SearchBot (OpenAI), ClaudeBot (Anthropic), Google-Extended and Googlebot (Google), Meta-ExternalAgent (Meta), PerplexityBot (Perplexity), Applebot-Extended (Apple), Amazonbot (Amazon), and Bytespider (ByteDance)), among others, continues to expand, with new agents as well as renamed agents appearing on a rolling basis.³³⁷ A rightholder's robots.txt file is a static document that requires manual updates for each new entrant; a configuration that was comprehensive in 2024 may be outdated by the time it is deployed. Cloudflare's own recommendations acknowledge that robots.txt directives are only evaluated after server- or firewall-level rules, and that for at least one major crawler, robots.txt, according to Cloudflare, "not a reliable defense" at all.³³⁸ In summary, while opting out via robots.txt is increasingly popular, there is a documented and apparently widening gap between the intentions of rights holders and the practices of AI developers.

³³⁵ See Chapter 4, sub-section 4.3.4 of this thesis; SoftwareSeni, 'How to Govern AI Crawler Access to Your Website in 2026' (2026); Soar, 'AI Bots robots.txt Guide: GPTBot, ClaudeBot, PerplexityBot' (2026).

³³⁶ SoftwareSeni (n 335).

³³⁷ No Hacks, 'The AI User-Agent Landscape in 2026: A Complete Reference' (2026); Citevera, 'The Complete robots.txt Guide for AI Crawlers in 2026' (2026).

³³⁸ Digital Applied, 'AI Crawler Access Control: The 2026 Decision Matrix' (2026); Soar (n 335).

5.2.2. Metadata and Embedded Rights Information — A Two-Tier System

The fragility of metadata-based signals within the data pipeline was established in sub-section 4.3.4 and is not revisited here.³³⁹ The prior analysis, however, did not examine which parties embed such signals in practice, and the answer varies considerably depending on the category of rightholder.

Large institutional rightholders, such as news agencies, stock photography platforms, and major publishers, have historically used IPTC and XMP metadata for licensing tracking, attribution, and rights management, primarily for purposes unrelated to AI training.³⁴⁰ For these organisations, adding or repurposing rights-reservation metadata represents a gradual extension of existing infrastructure. In contrast, individual creators, including independent authors, photographers, bloggers, and open-source developers, typically lack comparable infrastructure, are often unaware of the legal significance of such metadata under Article 4(3) of the DSM Directive, and do not have accessible tools to embed it. Consequently, a de facto two-tier system emerges: rightholders most capable of signalling rights reservation through metadata are also, as discussed in Section 5.4, those with the greatest capacity for direct licensing, while those least prepared to use metadata collectively contribute the largest share of works to typical educational corpora.³⁴¹ Schema.org markup provides a parallel example: although widely adopted, it is used almost exclusively for search engine optimisation and not as a signal of rights reservation.³⁴²

Even where institutional rightholders implement rights metadata, the fragility of the processing pipeline identified in sub-section 4.3.4 means that such signals may not survive the scraping, format-conversion, and aggregation processes that precede inclusion in a training corpus.³⁴³

Therefore, the two-tier system described above understates a practical challenge: even signals

³³⁹ See Chapter 4, sub-section 4.3.4 of this thesis.

³⁴⁰ Margoni and Kretschmer (n 137) 35.

³⁴¹ Geiger, Frosio and Bulayenko (n 49) 174, 216.

³⁴² Margoni and Kretschmer (n 137) 36.

³⁴³ See Chapter 4, sub-section 4.3.4 of this thesis.

generated by the best-equipped rightholders are not assured to reach the stage in the pipeline where they are required to function.

5.2.3. Emerging Standards and Initiatives — From Technical Promise to Regulatory Process

Sub-section 4.3.4 identified the C2PA mechanism as technically sound, with its primary limitation being the absence of regulatory recognition. The analysis concluded that regulatory measures, rather than technological solutions, were lacking.³⁴⁴ Subsequent developments necessitate an updated assessment, not due to the resolution of the main issue, but because the regulatory process outlined in sub-section 4.3.4 has now commenced.

In December 2025, the European Commission, supported by the EU Intellectual Property Office and acting pursuant to Measure 1.3 of the Copyright section of the General-Purpose AI Code of Practice under the AI Act, initiated a stakeholder consultation on protocols for reserving rights from text and data mining.³⁴⁵ The consultation aims to identify, from the opt-out technical solutions listed in the EUIPO study of generative AI, a set of protocols that are state-of-the-art, technically implementable, and widely adopted within the creative sectors. Following stakeholder workshops, the Commission intends to publish a list of generally accepted machine-readable opt-out solutions, subject to review at least every two years.³⁴⁶ Signatories to the General-Purpose AI Code of Practice, which the Commission has approved as an adequate mechanism for general-purpose AI providers to meet copyright obligations, have committed to comply with the protocols developed through this process, as well as with existing agreements such as robots.txt, in accordance with the Copyright section of the Code.³⁴⁷

³⁴⁴ See Chapter 4, sub-section 4.3.4 of this thesis.

³⁴⁵ European Commission, ‘Commission Launches Consultation on Protocols for Reserving Rights from Text and Data Mining under the AI Act and the GPAI Code of Practice’ (Shaping Europe’s Digital Future, 2025).

³⁴⁶ European Commission (n 345); Society for Computers and Law, ‘European Commission Consults on Protocols for Reserving Rights from Text and Data Mining under the AI Act and the GPAI Code of Practice’ (2025).

³⁴⁷ Civil Society Alliances for Digital Empowerment, ‘Commission Opens Consultation on Machine-Readable Copyright Opt-Out Protocols under the AI Act’ (2025).

This development is significant for the analysis presented in this thesis for two reasons. First, it confirms from a regulatory perspective the conclusion regarding Gap 2 identified in Section 4.6: the Commission's consultation document explicitly aims to address the missing standard implied, but not specified, by Article 4(3) of the DSM Directive, which aligns with the gap identified in this thesis.³⁴⁸ Second, this suggests that addressing Gap 2 may increasingly depend on the governance framework established by the EU Artificial Intelligence Act, specifically the GPAI Code of Practice and its associated obligations, rather than on a separate implementing act directly targeting Article 4(3). Whether protocols developed within this AI Act-related process will be considered authoritative for legal proceedings under Article 4(3) of the DSM Directive remains uncertain. The language of the consultations indicates an expectation of convergence, although the two instruments formally remain distinct, as discussed in Section 4.6.³⁴⁹

As of the time of writing, no opt-out mechanism, whether robots.txt, metadata-based signals, or C2PA, has been established as a mandatory, legally recognised format for opt-out under Article 4(3) of the DSM Directive. The Commission's consultations have not yet produced the harmonised list they were intended to create.³⁵⁰ The implementation-focused analysis in this Section confirms the doctrinal findings of Section 4.6: Gap 2 is not merely a theoretical legislative omission, but one that results in a fragmented environment. The most widely used mechanism (robots.txt) faces a growing compliance gap, while the mechanism best suited for integration into training pipelines (embedded metadata) is primarily used by rightholders who are least dependent on it. The ongoing regulatory process has not yet closed this gap. Although the speed at which this process will resolve Gap 2 cannot be determined within this thesis, the

³⁴⁸ See Chapter 4, Section 4.6 of this thesis; European Commission (n 345).

³⁴⁹ See Chapter 4, Section 4.6 of this thesis (on EU AI Act, art 53 and its relationship to the DSM Directive).

³⁵⁰ Inside Privacy, 'European Commission Launches Consultations on the EU AI Act's Copyright Provisions and AI Regulatory Sandboxes' (2025).

available data most clearly indicate that the gap identified in Chapter 4 has been acknowledged at the regulatory level, but its closure remains a future objective rather than a current reality.³⁵¹

5.3. Structural Problems with the Opt-Out Mechanism

The opt-out mechanism, examined in Section 5.1 as a policy instrument and in Section 5.2 from a technical implementation perspective, encounters four additional challenges that influence its effectiveness as a balancing tool. Two of these challenges, the historical data problem and the enforcement gap, were primarily addressed in Chapter 4. In this chapter, they are briefly re-stated and repositioned within the context of policy balancing, rather than being re-examined in detail. The remaining two challenges, burden allocation and the territorial mismatch with the global AI training geography mechanism, are developed here for the first time.

5.3.1. Burden Allocation Asymmetry

Section 5.1 demonstrated, using Coasean analysis, that the opt-out default places the cost of changing the default on rightholders. This sub-section makes that abstract point concrete.

The number of rightholders whose works may be included in AI training datasets is, in practice, unlimited. Every photographer, blogger, independent musician, open-source software developer, and author who has ever published content online is a potential rightholder in this context. In contrast, the number of organisations conducting large-scale commercial AI training is limited to several dozen. For the opt-out mechanism to operate as intended, each rightholder must individually be aware of the existence of a machine-readable reservation mechanism with legal significance under Article 4(3) of the DSM Directive, understand the required technical format, implement it correctly, and maintain this implementation as AI crawlers evolve, as outlined in sub-section 5.2.1.³⁵² If any rightholder fails to complete any of these steps, their works remain

³⁵¹ See Chapter 4, Section 4.6 of this thesis.

³⁵² See Chapter 5, sub-section 5.2.1 of this thesis.

subject to the default conditions, meaning TDM is permitted. The members of the second group, meanwhile, need do nothing at all to benefit from any such failure.

This imbalance is further intensified by the resource asymmetry, which follows the same trajectory. Individual creators typically lack legal departments and technical staff responsible for rights management, and in many cases are unaware that concepts such as “machine-readable reservation” have legal significance. In contrast, major AI developers possess substantial legal and technical resources, a comprehensive understanding of the regulatory framework, and a direct commercial interest in the resolution of any regulatory uncertainty. The consequences extend beyond abstract distribution. As established in Section 4.6, the validity of any reservation mechanism under Article 4(3) of the DSM Directive remains uncertain without a binding standard.³⁵³ When uncertainty arises, the party with greater resources and a direct interest in the outcome is better positioned to adopt a favourable interpretation, while the less-resourced party bears the costs in terms of time, expertise, and potential litigation associated with any subsequent dispute. Thus, Gap 2 does not affect both sides of the opt-out relationship equally; it disproportionately impacts the party least able to absorb the costs of the uncertainty it generates.

A comparison with the EU’s approach in a related regulatory area demonstrates that this allocation of the burden was not the only option available to the EU legislator, even at a similar point in time. The General Data Protection Regulation establishes an opt-in structure for personal data processing: processing requires a lawful basis, often the consent of the data subject, and the responsibility for establishing this basis rests with the party seeking to process the data, rather than with the data subject, who would otherwise need to object to unauthorised processing.³⁵⁴ This comparison is not intended as a direct analogy. Copyright and data protection pursue distinct objectives, and implementing an opt-in regime for TDM would reintroduce the

³⁵³ See Chapter 4, section 4.6 of this thesis.

³⁵⁴ General Data Protection Regulation, art 6.

transaction costs that, as discussed in Section 5.1, motivated the adoption of the opt-out in the first place.³⁵⁵ Nevertheless, the comparison illustrates an important point: at the time of the DSM Directive’s adoption, the EU legal framework did not consider the “the burden falls on the party whose resource is at stake” model as the sole available approach. Alternative models, including intermediate approaches between pure opt-out and pure opt-in, have precedent in EU regulatory practice and are directly relevant to the reform options discussed in Chapter 6.³⁵⁶

5.3.2. Historical Data Problem

In sub-section 4.3.5, based on the most stringent of the three approaches considered, it is established that training data reproduced prior to the implementation of the DSM Directive had no legal basis under EU copyright law at the time of their reproduction; that the subsequent decision not to use such data cannot retrospectively alter this conclusion; and that, nevertheless, EU law does not provide an adequate legal remedy against the consequences of such reproductions, as they remain in models already in use.³⁵⁷ This analysis is not repeated here.

For the purposes of this Chapter, this implies a temporal limitation on the possibility of opting out as a policy tool, which no improvement to the mechanism itself can remedy. More sophisticated technical standardisation (see sub-section 5.2.3), concerns how opt-outs will be expressed in the future. A more favourable distribution of the burden (see sub-section 5.3.1), concerns who bears the costs of expressing them in the future. Neither of these approaches concerns data that was reproduced before any mechanism for expressing reservations came into existence. A hypothetically ideal opt-out mechanism (universally standardised, easily implemented and fully complied with) introduced today would still leave the training data embedded in every base model trained up to that point exactly where it was found in sub-section 4.3.5: beyond the

³⁵⁵ See Chapter 5, Section 5.1 of this thesis.

³⁵⁶ Geiger, Frosio and Bulayenko (n 49) 174, 217.

³⁵⁷ See Chapter 4, sub-section 4.3.5 of this thesis.

reach of any opt-out mechanism for reasons that have nothing to do with whether that opt-out works.

For the purposes of Research Question 3, this means that any assessment of how effectively the opt-out mechanism currently balances AI innovation with the protection of rightholders' rights must take into account the substantial existing stock of AI capabilities built during a period when this mechanism did not exist at all — a stock the balance of which, according to the present analysis, has never been disrupted by the opt-out mechanism and cannot be restored by it now. This is a fixed component of the answer to Research Question 3, not a variable: it does not depend on how Section 5.2 or sub-sections 5.3.1, 5.3.3 or 5.3.4 are ultimately resolved, and it is carried forward into the balancing assessment of Section 5.5 on that basis.³⁵⁸

5.3.3. Enforcement Gap

Section 4.6 established that, as the third structural gap identified in the chapter, the DSM Directive lacks a mechanism to verify an AI developer's compliance with lawful access requirements or the exceptions in Article 4(3). Article 53 of the EU AI Act, which addresses general transparency regarding training data rather than the enforcement of copyright, only partially addresses this gap.³⁵⁹ The analysis of this issue is not repeated here.

This subsection draws a distinction that Section 4.6 did not need to draw, because it was concerned with whether a mechanism for verification exists at all, and demonstrated its absence. The present sub-section is concerned with effectiveness, and requires separating two distinct scenarios. In the first one, referred to as Failure Mode A, the situation Section 4.6 addressed, there is either no valid opt-out or its validity is uncertain, and the rightholder lacks any means to establish that their work has been used or that their reservation was intended to prevent such

³⁵⁸ See Chapter 5, Section 5.5 of this thesis.

³⁵⁹ See Chapter 4, Section 4.6 of this thesis.

use. In the second scenario, Failure Model B, the rightholder has fulfilled all reasonable requirements by implementing an explicit reservation targeting the named AI crawler via the user-agent, as detailed in sub-section 5.2.1. Despite this, the reservation, as the compliance data presented in that sub-section documents, simply been disregarded.³⁶⁰

Failure Model B undermines the legitimacy of the opt-out mechanism to a greater extent than Failure Model A. A mechanism with uncertain legal validity may become effective once the uncertainty is resolved (see sub-section 5.2.3), which outlines the current regulatory process addressing this issue. In contrast, a mechanism that is unambiguously and correctly triggered but still ignored provides no such prospect, leaving the rightholder without further recourse. At this stage, the opt-out, as described in sub-section 4.3.4, ceases to be a legal right with practical effect and instead becomes a largely desirable protection. The compliance data in sub-section 5.2.1 confirm that this is not merely a hypothetical scenario but a documented and worsening reality.³⁶¹

The practical implication for Research Question 3 is clear. An opt-out that can be validly made but has no effect is, from the rightholder's perspective, functionally equivalent to the absence of an opt-out. The formal existence of Article 4(3) of the DSM Directive and the practical absence of consequences for disregarding a reservation made under it are distinct issues, and Research Question 3 addresses the latter concern.

5.3.4. Global Mismatch

The preceding analysis examined so far concern the opt-out's operation within the framework that governs it, specifically its technical form, burden allocation, temporal scope, and enforcement. This final subsection considers a more fundamental issue: whether the regulatory framework's geographical scope aligns with that of the activities it regulates.

³⁶⁰ See Chapter 5, sub-section 5.2.1 of this thesis.

³⁶¹ See Chapter 4, Section 4.3.4 of this thesis.

Article 4(3) of the DSM Directive, consistent with that Directive as a whole, constitutes EU legislation governing activities within the territorial scope of EU copyright.³⁶² In practice, however, the training activities of major AI developers frequently occur outside the EU. Organisations are often established abroad, and the servers used for scraping, storage, and computation are typically located outside the EU, predominantly in the United States.³⁶³ When the act of reproduction (specifically, the collection and storage of training data) occurs outside the EU and is conducted by an entity not established within the EU, the direct applicability of EU copyright law to such acts becomes doctrinally uncertain. Consequently, even if an EU rightholder's reservation is implemented in full compliance with the relevant conditions (as discussed in Section 5.2), it may have no factual effect on reproduction acts that fall outside the scope of EU copyright law.

EU law possesses mechanisms for regulating acts committed outside its territory. In various contexts, it has been applied based on effects produced within the EU, irrespective of where the underlying acts occurred; this approach is often referred to as the “Brussels effect”.³⁶⁴ Accordingly, if a model trained outside the EU is subsequently used, sold, or applied within the EU, such use may establish a jurisdictional basis for imposing certain EU legal consequences on the AI developer. Nonetheless, a significant gap remains in this reasoning. While EU law may regulate the deployment of a model based on its impact within the EU, this does not necessarily extend to the training process that created the model, particularly when that process occurred entirely outside the EU and prior to any deployment decision. A deployment-based mechanism addresses actions distinct from those governed by the opt-out mechanism. Furthermore, even if EU legal consequences could, in principle, apply to deployment, sub-section 4.3.5 has demon-

³⁶² DSM Directive, art 4(3).

³⁶³ Sobel (n 68) 1, 44.

³⁶⁴ Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (Oxford University Press 2020) 25–40.

strated that practical remedies for training data embedded in an already trained model are limited.³⁶⁵ As a result, the challenges of non-compliance and historical data compound one another at exactly this point, leaving EU rightholders with, at best, remedies targeting model deployment and only limited recourse regarding the training activity, which the opt-out mechanism was not intended to address.

A final consideration is more predictive than doctrinal, and should be read accordingly. While training activities within the EU fall under Article 4(3) of the DSM Directive, subject to the practical limitations outlined in Section 5.2 and sub-sections 5.3.1–5.3.3, training activities conducted outside the EU do not fall directly within its scope. This framework creates a structural incentive to relocate the training-related stages of AI development outside the EU, regardless of the eventual deployment location of the resulting model. Such relocation does not diminish the commercial appeal of deploying models for EU users, as this depends on factors unrelated to the training location; it merely shifts the activity most directly regulated by the opt-out mechanism beyond its reach. If this incentive operates as described, the opt-out will be most effective where it is least necessary, namely, against EU developers who, due to their general subjection to EU law, may already have strong incentives to comply. Conversely, it will be least effective against the largest global developers, who have both the practical capacity to relocate operations and, as discussed in Section 4.6, the greatest influence on rightholders in cases of non-compliance.³⁶⁶

The conclusion of this sub-section is therefore more limited than those of sub-sections 5.3.1–5.3.3 and is stated as follows: the territorial assumptions underlying the opt-out mechanism may not align with the actual territorial scope of the activities it seeks to regulate. If such misalign-

³⁶⁵ See Chapter 4, sub-section 4.3.5 of this thesis.

³⁶⁶ See Chapter 4, Section 4.6 of this thesis.

ment exists, it represents a limitation on the mechanism's effectiveness that persists independently of, and cannot be addressed by, any improvements in implementation discussed above. Whether this limitation currently operates in practice or remains a structural possibility is a question that this thesis identifies rather than resolves.

5.4. Licensing Practices and Market Developments

Section 4.5.2 establishes that the existence of a valid licence renders the question of the applicability of the TDM exception irrelevant: if an AI developer has obtained authorisation directly from the rightholder, neither Article 3 nor Article 4 applies at all.³⁶⁷ Therefore, licensing constitutes an alternative route to legality that operates independently of the opt-out mechanism discussed in this Chapter. Its relevance to Research Question 3 is evident: a sufficiently developed licensing market could collectively compensate the shortcomings of the opt-out mechanism identified in Sections 5.2 and 5.3. Even an imperfect opt-out mechanism becomes less significant if rightholders can secure compensation through licensing instead. This Section examines the current state of this market and evaluates whether it can fulfil this compensatory function.

5.4.1. Emerging Direct Licensing Markets

The market for direct licensing of content to AI developers has grown significantly and changed in character since the period addressed in Chapter 4 analysis. By mid-2025, news and journalistic content accounted for around 48 publicly documented licensing deals (significantly more than music and audio (16) and images and video (12)), and the overall volume of such deals has risen from zero in 2022 to 28 in 2024, with further growth forecast for the future.³⁶⁸ By

³⁶⁷ See Chapter 4, sub-section 4.5.2 of this thesis.

³⁶⁸ Rob Kelly, 'AI Content Licensing Deals: June 2026 Update' (Media and the Machine, 2026).

mid-2026, OpenAI alone had entered into 24 publicly announced licensing agreements, which is reportedly almost double that of Microsoft or Meta.³⁶⁹

The structure of named arrangements reflects the composition of the licensing market. In journalism, OpenAI's licensing partners include the Associated Press, Axel Springer (publisher of Bild and Business Insider), Le Monde, Prisa Media, the Financial Times, Vox Media, Time, The Guardian, and News Corp. Reportedly, a five-year agreement valued at over \$250 million was signed with News Corp, covering publications such as the Wall Street Journal and The Times of London.³⁷⁰ Beyond traditional publishing, Reddit has entered into licensing agreements with OpenAI and Google, reportedly worth approximately \$60 million annually for access to user-generated content. Stack Overflow has also licensed its corpus of technical questions and answers to OpenAI.³⁷¹ In book publishing and the stock image sector, similar agreements have emerged, although in smaller numbers compared to the journalism sector.³⁷²

A further development, not fully evident during the analysis in Chapter 4, concerns the evolving nature of licensing agreements. Early licensing deals were framed principally around training data access. More recent agreements increasingly emphasise “live access”, including the ongoing feeds, attribution, and integration into AI-powered search and retrieval-augmented generation, rather than one-time access to historical archives. The number of deals covering attribution and real-time access increased from 2 in 2023 to 18 in 2025, with additional growth projected.³⁷³ This trend indicates that, for major AI developers, ongoing commercial relationships with content providers, valuable for reasons that include but extend beyond the original TDM

³⁶⁹ Kelly (n 368).

³⁷⁰ Will Scott, ‘How AI Licensing Deals Determine Search Visibility in 2025’ (2025); Authors Alliance, ‘What Happens When Your Publisher Licenses Your Work for AI Training?’ (2025); eMarketer, ‘OpenAI to Fund Four Axios Newsrooms in Content Licensing Deal’ (2025).

³⁷¹ Contently, ‘How Do I Get into Training Datasets Used by ChatGPT or Claude?’ (2025).

³⁷² Authors Alliance (n 370).

³⁷³ Kelly (n 368).

training use, have become at least as significant a licensing consideration as the discrete question of training-data lawfulness.

A separate development with more direct implications for copyright is the settlement, in late 2025, of the *Bartz v Anthropic PBC* case, a class action lawsuit in the US concerning Anthropic's use of approximately 500,000 books obtained from pirate libraries, including Library Genesis, precisely the kind discussed in sub-section 4.4.2 of this thesis.³⁷⁴ Anthropic agreed to pay a minimum of \$1.5 billion, amounting to approximately \$3,000 per work, and also undertook to destroy the pirated source files and any copies derived from them.³⁷⁵ The settlement, again, is a matter of US law and does not set a precedent in the EU, the same caveat applies here as to the court proceedings discussed in sub-section 4.4.2.³⁷⁶ Its significance for this thesis is twofold. Firstly, this is the first instance (indeed, the only one) where a specific, court-ordered monetary sum has been awarded for the unauthorised use of a specific copyright-protected work in AI training: \$3,000 per book. Secondly, the same US court that approved this settlement had previously ruled in the same case that training on legally acquired copyrighted books could constitute fair use, meaning that the \$1.5 billion figure relates specifically to the piracy component of Anthropic's actions, rather than to training on copyrighted materials per se.³⁷⁷

The \$3,000-per-work figure from the Anthropic settlement provides, for the first time, a concrete benchmark for the issue of compensation raised as Gap 1 in Section 4.6.³⁷⁸ It does not translate directly into an EU compensation mechanism, it is an amount of damages arising from litigation concerning unlawfully sourced material, rather than a licence fee for lawfully obtained material, and there is no equivalent class-action mechanism in EU law through which such a

³⁷⁴ See Chapter 4, sub-section 4.4.2 of this thesis; *Bartz v Anthropic PBC*, No 3:24-cv-05417-WHA (ND Cal, settlement preliminarily approved 25 September 2025).

³⁷⁵ Susman Godfrey LLP, 'Susman Godfrey Secures \$1.5 Billion Settlement in Landmark AI Piracy Case' (2025); Fieldfisher, 'Anthropic's \$1.5 Billion Copyright Settlement with Authors Could Set Precedent in AI Industry' (2025).

³⁷⁶ See Chapter 4, sub-section 4.4.2 of this thesis.

³⁷⁷ Ropes & Gray LLP, 'Anthropic's Landmark Copyright Settlement: Implications for AI Developers and Enterprise Users' (2025).

³⁷⁸ See Chapter 4, Section 4.6 of this thesis.

figure could arise.³⁷⁹ Nevertheless, it is a first point of reference, wherever it may be, for what “compensation per work” might mean in the context of AI training — an issue which, as discussed in sub-section 5.4.2, any EU collective licensing scheme will also need to address, in some form, for lawfully accessed material.

These developments are real, substantial and were not fully apparent when the doctrinal analysis of Chapter 4 was conducted. They should not be dismissed. They are, however, concentrated within a particular segment of the rightholder population: major news organisations, large publishers, and platforms (such as Reddit and Stack Overflow), which themselves aggregate vast volumes of individually-authored content and can license it on behalf of, or independently of, their individual contributors.³⁸⁰ The ability to enter into a multi-year \$250 million agreement or to be one of 24 counterparties in OpenAI’s licensing programme requires precisely the level of legal literacy, bargaining power and institutional scale which, as noted in sub-section 5.3.1, are systematically lacking among the vast majority of individual rightholders. The content that makes up the bulk of large-scale training corpora (individual websites, blogs, forums, open-source repositories, and works by independent creators) remains, on the evidence available, effectively outside this market. It is impossible to give an exact figure for the proportion of a typical training corpus covered by systems such as those discussed above, but given the scale of the corpora discussed in sub-section 4.1.2 and the number of counterparties (even at 24, even at 48), this share almost certainly constitutes a small fraction of the total volume, concentrated precisely in those content categories (news archives, major platforms) that are most visible and most capable of self-organisation.³⁸¹

³⁷⁹ DSM Directive, arts 3–4.

³⁸⁰ See Chapter 5, sub-section 5.3.1 of this thesis.

³⁸¹ See Chapter 4, sub-section 4.1.2 of this thesis.

5.4.2. Collective Licensing Models

If direct licensing applies only to those rightholders who are able to negotiate individually, the question arises as to whether EU legislation already provides for a mechanism to extend licence agreements to rightholders who are not in such a position. Article 12 of the DSM Directive provides for such a mechanism in the form of extended collective licensing.³⁸²

Under Article 12 of the DSM Directive, a collective management organisation that is sufficiently representative of rightholders in a given category of works may conclude a licence agreement covering a specific use, and Member States may provide for the extension of this licence, by operation of law, to rightholders in the same category who are not members of the society, subject to safeguards, including the right of any rightholder who is not a member of the society to opt out of the extension.³⁸³ A structural parallel with Article 4(3) is noteworthy: both mechanisms operate on a default-plus-opt-out logic. The difference lies in what the default consists of. Under Article 4(3), the default is authorisation without compensation, whilst a rightholder's opt-out completely excludes the use. Under Article 12 extended licence, the default would be a licensed and compensated use, whilst the rightholder's refusal to participate will exclude only that rightholder from the collective agreement, leaving them free to negotiate individually, just as the rightholders discussed in sub-section 5.4.1 already do.³⁸⁴

The EU has no shortage of institutional infrastructure for collective licensing of this kind. Copyright collective management organisations operating in related fields: performance rights, mechanical reproduction rights and their equivalents in all Member States — have decades of experience precisely in the task that would be required for an AI-based extended licensing scheme: aggregating the interests of large numbers of individual rightholders, negotiating on their behalf with large commercial users, and distributing the resulting revenue.³⁸⁵ Historical

³⁸² DSM Directive, art 12.

³⁸³ DSM Directive, arts 12–13.

³⁸⁴ Geiger, Frosio and Bulayenko (n 49) 174, 218.

³⁸⁵ Margoni and Kretschmer (n 137) 38.

precedents for the underlying “permitted use plus compensation” structure are also well known. Cable retransmission rights, where individual licensing by every rightholder to every retransmitting operator was impractical at scale, were resolved through collective licensing; private copying levies, already cited in Section 4.6 as a model for compensating for Gap 1, which is absent from Article 4, are in practice administered through the same collecting-society infrastructure.³⁸⁶ The question is, therefore, not whether the EU possesses the institutional tools to administer a compensation scheme for AI training, but whether it has the legal authority to apply these tools to this activity — and, as matters currently stand, it does not.

A scheme built on this infrastructure could function as follows: collective management organisations, authorised depending on the category of work (literary, visual, musical, audiovisual and, possibly, software or code as a separate category), would grant AI developers licences for large-scale TDM on behalf of both represented and, through the extension mechanism provided for in Article 12 of the DSM Directive, unrepresented rightholders, whilst unrepresented rightholders would also retain the individual right to exclude their works.³⁸⁷ AI developers would contribute to the scheme in accordance with a specified metric: the scale of training data obtained from the relevant category, the revenue generated by the resulting models, or a combination thereof — the precise metric being a design question not addressed in this thesis, leaving it for reform discussion in Chapter 6.³⁸⁸ Such a scheme would directly address Gap 1 and do so on a scale that the direct licensing agreements discussed in sub-section 5.4.1 are structurally unable to achieve, regardless of how much they continue to grow, since the extension mechanism provided for in Article 12 of the DSM Directive, unlike bilateral negotiations, does not depend on any individual rightholder being able to be found, contacted or negotiated with.

³⁸⁶ See Chapter 4, Section 4.6 of this thesis.

³⁸⁷ DSM Directive, art 12(3); Senftleben (n 146) 1, 19.

³⁸⁸ See Chapter 6 of this thesis.

5.4.3. Critical Assessment

The preceding sub-sections converge on a single conclusion, which can be examined from two perspectives. Direct licensing (sub-section 5.4.1) is an established and evolving practice, and in certain cases, most notably the Anthropic settlement, it has begun to set specific monetary values for the use of copyrighted works in AI training. However, this practice is largely limited to rightholders with sufficient bargaining power, such as major publishers, platforms, and, in Anthropic's case, a well-organised and well-resourced class action. Collective licensing (sub-section 5.4.2) possesses the institutional infrastructure and legal precedent to include rightholders who have no factual access to direct licensing. Nevertheless, there is currently no legal mandate to extend this infrastructure to AI training, and the development of the direct licensing market described in sub-section 5.4.1 does not indicate that AI developers, who currently benefit from an uncompensated exemption under Article 4 of the DSM Directive, would voluntarily adopt such a scheme.³⁸⁹

The asymmetry identified in sub-section 5.3.1 for the opt-out reappears here in mirror form. There, the concern was that individual rightholders bear the burden of acting under a default, from which AI developers benefit without active participation. In the current context, the concern is that individual rightholders lack sufficient leverage to participate as counterparties in the licensing market shaped by the same default rule. Individually-authored content on the open web, which constitutes the majority of training corpora, simultaneously represents the group least able to exercise an effective opt-out and least able to secure a licensing deal. Collective licensing in adjacent fields addresses this second asymmetry; its absence in this context leaves individual authors doubly unprotected.³⁹⁰

³⁸⁹ See Chapter 4, Section 4.6 of this thesis (on the absence of compensation under art 4 of the DSM Directive as a deliberate legislative choice).

³⁹⁰ See Chapter 5, sub-section 5.3.1 of this thesis.

The central question of the present Section can now be addressed directly: Can voluntary market developments substitute for reform of the opt-out mechanism? Based on the available data, the answer is negative. The segment of the market reached by licensing largely comprises entities that, due to their scale and complexity, have never been highly dependent on the opt-out mechanism for protection. Large rightholders possess the capacity to negotiate, litigate, or pursue both routes, irrespective of Article 4(3) of the DSM Directive. In contrast, the segment most reliant on the ability to opt out of licensing remains largely unaddressed by current licensing practices. The evolution of the licensing market, as described in sub-section 5.4.1, toward direct access and attribution agreements with major platforms, does not indicate that market forces alone will resolve this gap. Collective licensing, as discussed in sub-section 5.4.2, is the most appropriate mechanism to address this issue, as its effectiveness does not depend on the factors that limit direct licensing for most rightholders. However, implementing such a mechanism requires a legal framework that is currently absent. This necessity for legislative action is the focus of Chapter 6, and it cannot be presumed that such a framework will arise solely from existing market dynamics.³⁹¹

5.5. Balancing Innovation and Rightholder Protection

The findings of Sections 5.1–5.4 have so far been evaluated against policy objectives, specifically transaction-cost reduction and market facilitation, which Section 5.1 identified as the opt-out’s rationale. This section assesses those conclusions against a more rigorous benchmark: the framework of fundamental rights established by the CFREU. Within this framework, the TDM provisions of the DSM Directive, as legislative measures of EU law, must themselves be capable of justification.

³⁹¹ See Chapter 6 of this thesis.

Article 17(2) of the CFREU protects intellectual property, thereby establishing copyright and other forms of IP as fundamental rights rather than mere statutory entitlements.³⁹² Article 16 of the CFREU safeguards the freedom to conduct a business, a right directly relevant to the interests of AI developers engaged in TDM and the development of commercial AI products without disproportionate restrictions.³⁹³ Articles 11 and 13 of the Charter, which guarantee freedom of expression and information, and freedom of the arts and sciences, respectively, are also engaged. TDM and AI development may be framed as exercises of these freedoms, particularly in the context of research activities protected by Article 3 of the DSM Directive.³⁹⁴ None of these rights is absolute. The CJEU has consistently held, across contexts ranging from intermediary liability to data retention, that when the exercise of one fundamental right by one party restricts the exercise of another fundamental right by another party, a fair balance between them must be achieved. Legislative measures adopted by the EU legislature represent the legislator's primary attempt to strike this balance, an attempt that remains subject to judicial review for compliance with the requirement of fair balance.³⁹⁵

The doctrine of fair balance is most prominently recognised in the Court's case law on intermediary liability and data protection. The *Promusicae v Telefónica* and *SABAM v Netlog* cases both concern the balance between the protection of IP under Article 17(2) of the CFREU and competing rights, such as the freedom to conduct a business under Article 16 of the CFREU, particularly in the context of measures targeting internet intermediaries.³⁹⁶ Until recently, CJEU had not specifically addressed this balance in relation to the TDM exceptions of the DSM Directive. That situation is now evolving. A reference from the Budapest Regional Court, registered as *Like Company v Google Ireland*, is currently pending before the CJEU and concerns

³⁹² CFREU, art 17(2).

³⁹³ CFREU, art 16.

³⁹⁴ CFREU, arts 11, 13; Geiger, Frosio and Bulayenko (n 49) 174, 219.

³⁹⁵ Case C-275/06 *Productores de Música de España (Promusicae) v Telefónica de España SAU* [2008] ECR I-271, paras 61–68; Case C-360/10 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85, paras 42–46.

³⁹⁶ *Promusicae* (n 395); *SABAM v Netlog* (n 395).

precisely the question of what constitutes an “appropriate” rights reservation under Article 4(3) of the DSM Directive where AI training is at issue, and examines how this preservation interacts with the reproduction right under Article 2 of the InfoSoc Directive.³⁹⁷ Separately, the Hamburg Regional Court’s decision in *Kneschke v LAION*, although a national first-instance ruling rather than a decision of the CJEU, applied the research exception of Article 3 of the DSM Directive to the compilation of an AI training dataset. The court also offered, in obiter dictum, observations on the machine-readability requirement in Article 4(3) of the DSM Directive; however, as noted in subsequent commentary, these remarks were not essential to the court’s decision, which relied exclusively on Article 3 of the DSM Directive.³⁹⁸

The pendency of *Like Company v Google Ireland* is significant for this thesis in two respects. First, it confirms that the question addressed in this Section (specifically, whether the balance Article 4 of the DSM Directive strikes between Articles 16 and 17(2) of the CFREU is fair in light of the actual operation of the opt-out mechanism) is neither abstract nor premature. Instead, it has reached the EU’s highest court through a preliminary ruling from a national court confronting the question on its facts.³⁹⁹ Second, the analysis presented in this Section must be regarded as preliminary, in contrast to the doctrinal analysis in Chapter 4. The Court’s forthcoming decision may clarify, in unforeseen ways, what constitutes an “appropriate” reservation under Article 4(3) of the DSM Directive, and how it interacts with the reproduction right. Such clarification may directly influence the conclusions drawn in Sections 5.2 and 5.3. This thesis

³⁹⁷ Case C-250/25 *Like Company v Google Ireland Ltd* (reference lodged 3 April 2025); Oxford Law Blogs, ‘AI and the Redefinition of Law in Code: The Future of Legal Intention under Article 4 DSM’ (15 December 2025).

³⁹⁸ *Robert Kneschke v LAION e.V.*, Landgericht Hamburg, 27 September 2024; RAILS Blog, ‘Landmark Court Decision in the EU: Copyright Permissibility of Text and Data Mining for the Purpose of AI Training’ (5 January 2025); Morrison Foerster, ‘To Scrape or Not to Scrape? First Court Decision on the EU Copyright Exception for Text and Data Mining in Germany’ (4 October 2024).

³⁹⁹ *Like Company v Google Ireland Ltd* (n 397).

relies on the legislation and evidence available at the time of writing, while acknowledging that, in this domain, the highest judicial authority is, quite literally, pending.⁴⁰⁰

Applying the fair balance framework to the two TDM exceptions in turn, Article 3 of the DSM Directive represents the less complex scenario. This mandatory, uncompensated exception is limited to qualified research organisations engaged in non-commercial scientific research and is subject, as established in Chapter 4, to compliance with lawful access requirement and proportionate security measures under Article 3(2). This framework aligns with the long-standing pattern of exceptions for research and education recognised in both EU legislation and national copyright law.⁴⁰¹ The resulting balance prioritises Articles 11 and 13 of the CFREU (research, expression, the arts and sciences) over Article 17(2) of the same Charter within a specific and narrowly defined set of cases, consistent with established legal practice. The analysis in this Chapter supports this assessment.

Article 4 of the DSM Directive is where the findings of this Chapter converge, and where the fair balance question becomes genuinely difficult. The balance established by Article 4 encompasses a broad personal scope extending to commercial entities (as detailed in sub-section 4.3.1), lacks a compensation mechanism (identified as Gap 1 in Section 4.6), and relies on the opt-out mechanism under Article 4(3) as the primary safeguard for rightholders. The effectiveness of this opt-out mechanism as a meaningful restriction is questionable. This Chapter demonstrates that the mechanism is insufficient for several independent and cumulative reasons: technical implementation remains fragmented and, even with ongoing standardisation efforts, incomplete (Section 5.2); the default allocation of responsibility places the burden on those least able to bear it (sub-section 5.3.1); it cannot address training data that predates its introduction, a limitation that cannot be remedied by future improvements (sub-section 5.3.2); even when

⁴⁰⁰ Like *Company v Google Ireland Ltd* (n 397); 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 *Journal of Intellectual Property Law & Practice* 453, 462.

⁴⁰¹ See Chapter 4, sub-sections 4.2.3 and 4.2.4 of this thesis.

properly implemented, evidence shows that the principle is frequently disregarded (sub-section 5.3.3); and its territorial assumptions may not align with the actual geography of relevant activities (sub-section 5.3.4). The licensing market discussed in Section 5.4, although expanding, primarily serves rightholders least dependent on the opt-out in the first place, and cannot adequately compensate for these deficiencies for those most dependent on such protection.

A precise definition of the emerging imbalance is essential, as imprecision risks exaggeration. The imbalance identified here does not stem from the existence of Article 4 of the DSM Directive, its application to commercial actors, or the adoption of an opt-out rather than a default opt-in mechanism. Each of these choices, when considered independently, falls within the spectrum of outcomes that a legislator aiming to balance Articles 16 and 17(2) of the CFREU could reasonably select. The imbalance instead arises from the combination of these choices with the implementation deficiencies documented in this Chapter: a framework whose fairness depends on a safeguard that does not, in practice, perform the function that its existence is relied upon to justify. This represents the constitutional analogue of the doctrinal conclusion reached in Section 4.6, that Section 4, as applied to AI training, is formally valid but substantively hollow.⁴⁰² A safeguard that exists only on paper may meet the formal requirement for a balancing mechanism; whether it fulfills the substantive requirement of a fair balance is a separate issue. On the evidence of this Chapter, the answer to that separate question is negative.

This framing has a direct consequence for how reform, which is addressed in Chapter 6, should be perceived. If the imbalance results from the gap between the formal existence of a safeguard and its substantive effect, then bridging this gap requires ensuring the functionality of the existing protection, the possibility of refusal, rather than introducing new protective measures outside the structure of the DSM Directive. Addressing the absence of a technical standard, identified in Section 4.6 as Gap 2, would enable rightholders who formally possess the opt-out

⁴⁰² See Chapter 4, Section 4.6 of this thesis.

right under Article 4(3) of the DSM Directive to exercise it effectively. Remedying the lack of control (Gap 3) would ensure that a properly exercised opt-out is respected in practice. Furthermore, expanding collective licensing in line with the principles discussed in sub-section 5.4.2 would address Gap 1 (compensation) for a significant proportion of rightholders, for whom an opt-out remains the “wrong” tool, because their interest is not in preventing use but in being paid for it. In this context, the reform reinforces the internal logic of the DSM Directive rather than departing from it.⁴⁰³

The fair balance intended by the drafters of the DSM Directive for Article 4, as analysed in this Chapter, remains unachieved. This outcome, however, does not result from a flaw in the legislative framework. As noted in Section 4.6, the underlying inconsistency was attributed to technological progress outpacing legislative cycles, rather than legislative error, and the conclusions of this Chapter uphold that characterisation.⁴⁰⁴ The primary issue is that the implementation of this framework (the technical, administrative, and market infrastructure it requires) has not matched the scale and complexity of the activities it now governs. Furthermore, as highlighted by the historical data problem in sub-section 5.3.2, this gap could not have been closed, regardless of the speed of implementation. Chapter 6 addresses the resulting question: what concrete legislative and regulatory steps could bring the implementation of Article 4 into alignment with the balance its design already presupposes, and how feasible are those steps within the existing structures of EU law and policy-making, including the regulatory processes, such as the consultation discussed in sub-section 5.2.3, that are already underway?

⁴⁰³ See Chapter 4, Section 4.6 of this thesis; see Chapter 5, sub-section 5.4.2 of this thesis.

⁴⁰⁴ See Chapter 4, Section 4.6 of this thesis.

5.6. Answer to Research Question 3

The analysis conducted across this Chapter permits a direct answer to the third research question posed by this thesis: how effectively does the opt-out mechanism under Article 4(3) of the DSM Directive balance AI innovation against rightholder protection?

The concise answer is that it does not balance these interests effectively. This conclusion is supported by five distinct shortcomings, none of which can be remedied by the others: the lack of a binding technical standard for reservations; a default allocation of compliance costs to those least able to bear them; the inherent incapacity to access training data predating the mechanism's creation; routine disregard of even properly formulated reservations; and a territorial scope that may not align with the geography of relevant activities.⁴⁰⁵ Voluntary licensing primarily benefits rightholders who needed the opt-out least, while failing to protect those who require it most.⁴⁰⁶ In comparison with Articles 16 and 17(2) of the Charter, the resulting framework is formally valid but does not, in practice, strike the fair balance mandated by these provisions.⁴⁰⁷

This is not, as Chapter 4 and the present Chapter have both maintained, a conclusion that the approach of the DSM Directive was inherently flawed. Rather, implementation has not caught up with design, in the face of an activity whose scale the design did not anticipate.⁴⁰⁸ Chapter 6 addresses potential solutions to this gap and assesses whether current EU legislative and regulatory tools, including ongoing processes, are adequate for this purpose.

⁴⁰⁵ See Chapter 5, Sections 5.2 and 5.3 of this thesis.

⁴⁰⁶ See Chapter 5, Section 5.4 of this thesis.

⁴⁰⁷ See Chapter 5, Section 5.5 of this thesis.

⁴⁰⁸ See Chapter 4, Section 4.6 of this thesis; see Chapter 5, Section 5.5 of this thesis.

6. Liability Issues and Future Regulatory Perspectives

6.1. Copyright Liability for AI Training Under Current EU Law

Chapters 3 to 5 identify circumstances in which AI training is not protected under Articles 3 or 4 of the DSM Directive. This Section examines the implications of this finding, specifically addressing liability, its legal basis, and resulting consequences. The analysis builds upon the provisions previously outlined in this thesis and applies them to the issues of liability and remedies, which were only briefly addressed in the aforementioned Chapters.

6.1.1. Direct Infringement

Where neither Article 3 nor Article 4 of the DSM Directive applies, and no licence has been obtained, the AI developer is a direct infringer of the reproduction right under Article 2 of the InfoSoc Directive. This conclusion synthesises the findings in Chapter 3 that AI training involves reproduction and the conclusions in Chapters 4 and 5 that, in the scenarios examined, no exception is applicable.⁴⁰⁹ Furthermore, copyright infringement in the EU is assessed objectively; good faith influences the available remedies but does not determine whether infringement has occurred, as established in sub-section 4.4.3 for the specific case of pirated datasets.⁴¹⁰

This generalisation is genuinely new and represents the first instance where the thesis's various findings (about why Article 4 of the DSM Directive fails to cover a given act of training) are shown to converge on a single liability consequence. Sub-section 4.4.3 established strict liability specifically for scenarios involving pirated datasets. The same reasoning applies wherever Article 4 does not apply, including situations where the source was not pirated but a valid exception under Article 4(3) existed and its own terms were unmet. A developer who, due to the uncertainty of standardisation identified as Gap 2 in Section 4.6, reasonably but mistakenly

⁴⁰⁹ See Chapter 3, Section 3.5; Chapter 4, Section 4.7 of this thesis.

⁴¹⁰ See Chapter 4, sub-section 4.4.3 of this thesis.

believed that no valid reservations had been made, faces the same liability as one who used a pirated dataset: both are direct infringers. The distinction lies not in the occurrence of infringement, but in the subsequent consequences, as further discussed in sub-section 6.1.3.⁴¹¹

Chapter 3 delineated scraping, storage, the training process, and the resulting model weights as sequential stages within the AI training pipeline. The first two stages clearly constitute reproduction, while the latter raises unresolved questions, as noted in sub-section 3.3.4.⁴¹² For liability assessment, this sequence reflects two distinct forms of disclosure. Scraping and storage are discrete, time-stamped actions, each either within or outside the scope of an applicable exception at the time of occurrence. This status is fixed at that moment and remains unaffected by subsequent events, including a later (even if valid) opt-out, as established in sub-section 4.3.5 for historical-data context.⁴¹³ In contrast, the trained model represents a continuous state; while in use, it embodies the consequences of prior actions, regardless of whether the issue of “weights” in Section 3.3.4 is resolved in favour of reproduction. Therefore, liability from scraping and storage is retrospective and time-limited, whereas liability associated with the model’s continued existence is ongoing.

This distinction is central to the primary question of this Section: can a deployed model be enjoined? Typically, injunctions addressing copyright infringement target future acts; however, the act most clearly constituting reproduction, namely scraping and storage, is usually completed before legal proceedings begin. An injunction against a model already in use would restrict not the repetition of that act, but the model’s continued existence or use. EU law does not clearly specify the form such an injunction might take or whether it can be issued in these circumstances at all.⁴¹⁴

⁴¹¹ See Chapter 4, Section 4.6 of this thesis; see Chapter 6, sub-section 6.1.3 of this thesis.

⁴¹² See Chapter 3, Section 3.3 of this thesis.

⁴¹³ See Chapter 4, sub-section 4.3.5 of this thesis.

⁴¹⁴ IPRED, art 11.

Three distinct objects for such an injunction can be identified, each with a different relationship to the infringement it seeks to address. (1) The first, an injunction restraining *further training* using the same unlawfully-sourced material is the most conventional in form. It addresses a genuinely future act but does not affect models that have already been trained. (2) The second, an injunction restraining deployment of the existing model directly targets the model itself. This approach, however, presents a significant challenge: the capabilities of a base model are determined by its entire training corpus, and there is currently no method to isolate the contribution of any specific subset, including unlawfully obtained data, to the model's overall function. Consequently, an injunction restraining deployment is inevitably over-inclusive, as it suppresses the model's lawful capabilities along with whatever unlawful contribution the infringing data made, in proportions that cannot be specified. (3) The third option, an injunction requiring re-training or deletion of the model, so as to remove the infringing contribution specifically, is the only one of the three that targets the infringement without over-inclusiveness problem. Nevertheless, as noted in sub-section 4.3.5, its technical feasibility remains uncertain. It has not been established that removing the contribution of a specific dataset from an already trained model is possible by any means other than training a new model from scratch.⁴¹⁵

This taxonomy provides concrete substance to the proportionality requirement that EU law imposes on remedies, including injunctions, in this context.⁴¹⁶ An injunction of the second type, which restricts the deployment of a commercially significant system due to unlawfully obtained training data, is, as analysed above, inherently disproportionate since its effect cannot be confined to the specific infringement. While such an injunction is not categorically precluded, as proportionality involves balancing interests, and the scale of infringement may, in principle, justify the remedy, it does require a different analytical approach. The proportionality analysis in this context cannot proceed as usual by simply asking whether the remedy is proportionate

⁴¹⁵ See Chapter 4, sub-section 4.3.5 of this thesis; Sobel (n 68) 1, 46.

⁴¹⁶ IPRED, art 3(2); Promusicae (n 395) para 68.

to the infringement. Instead, it must first address the unresolved question of whether the remedy can be linked to the infringement at all. Where it cannot (which is typically the case as indicated above), the proportionality analysis collapses into an all-or-nothing decision. This outcome is inconsistent with EU legislation on remedies, which is designed around proportionate counter-measures.

A useful, though imperfect, point of comparison exists outside EU law. In the settlement agreement in *Bartz v Anthropic PBC*, discussed in section 5.4.1, the remedy included both a damages payment and the destruction of pirated source files and any derivative copies. This remedy, equivalent to an injunction, more closely aligns with object (1) than object (3): the court ordered destruction of the source material, but the settlement did not require retraining or deletion of models already trained on it.⁴¹⁷ If this outcome is considered indicative of the practical limits of even well-justified remedies, addressing unlawfully held source material but not the trained model, it supports the concern outlined above. Object (3), the remedy that would most directly address infringement within the trained model, may be the least likely to be implemented in practice, not due to a lack of justification, but because it may simply be unavailable given current technical constraints.

6.1.2. Liability for Use of Pirated Datasets

Sub-section 4.4.3 establishes liability for training AI systems on pirated datasets in full: it identifies primary infringement by the dataset operator and direct infringement by the AI developer through independent scraping and storage. The sub-section further clarifies that good faith is irrelevant to liability as such, and that commonly used notorious pirated databases do not meet the required standard of due diligence.⁴¹⁸ These points are not reiterated here.

⁴¹⁷ See Chapter 5, sub-section 5.4.1 of this thesis.

⁴¹⁸ See Chapter 4, sub-section 4.4.3 of this thesis.

This sub-section is included purely for the sake of argument. The scenario involving pirated datasets provides the most illustrative example of direct infringement analysed in sub-section 6.1.1. Since, in accordance with sub-section 4.4.3, no good-faith-based mitigation is realistically available, the remedies discussed in sub-section 6.1.3 (particularly the recovery of unlawfully obtained profits and the increase in damages for wilful infringement) apply in full. The Anthropic settlement, again, serves as an illustration: the US court, which considered this scenario, determined both the amount of damages and a remedy equivalent to an injunction, similar to that discussed in sub-section 6.1.1. Therefore, in this scenario, both parts of the remedy analysis considered in this Section (injunctions in sub-section 6.1.1 and damages in sub-section 6.1.3) are simultaneously and most clearly engaged.⁴¹⁹

6.1.3. Remedies Available to Rightholders

Articles 11 and 13 of the IPRED, which address injunctions and damages respectively, serve as sources of legal remedies at the EU level, as discussed in sub-sections 4.4.3 and 6.1.1.⁴²⁰ This sub-section organises these remedies, examining both the substance of each remedy and the extent to which their practical availability is shaped by the gaps in Article 4 of the DSM Directive identified in Chapters 4 and 5.

Damages may be calculated based on the rightholder's actual losses, typically lost licensing revenue, or on a hypothetical licence fee for the relevant use.⁴²¹ The first approach presents a complex challenge in the context of Article 4 of the DSM Directive. Section 4.6 identified, as Gap 1, that Article 4 does not provide for compensation even when applied as intended. In a scenario where the rightholder's right to opt out is respected, no licensing revenue would be due.⁴²² If the underlying right is valued at zero, it becomes unclear what "loss" a claim for actual

⁴¹⁹ See Chapter 5, sub-section 5.4.1 of this thesis; see Chapter 6, sub-sections 6.1.1 and 6.1.3 of this thesis.

⁴²⁰ IPRED, arts 11, 13; see Chapter 4, sub-section 4.4.3 of this thesis; see Chapter 6, sub-section 6.1.1 of this thesis.

⁴²¹ IPRED, art 13(1).

⁴²² See Chapter 4, Section 4.6 of this thesis.

damages would reference, even if the AI developer's actions were unlawful for unrelated reasons, such as the standardisation failure of Gap 2 resulting in a valid reservation being unrecognised. Thus, Gap 1 may not only preclude compensation under Article 4, but also restrict the damages available when Article 4 is breached by removing the baseline against which "actual loss" is measured.

A hypothetical licence fee measure is less affected by this problem, as it considers the plausible market cost of a licence rather than relying on the provisions of Article 4 of the DSM Directive itself. In this context, the \$3,000-per-work figure cited in sub-section 5.4.1 of the *Bartz v Anthropic* settlement provides the first concrete real-world benchmark for what such a fee might look like in the AI training context. This figure, however, originates from US litigation over piracy specifically, not from the EU licensing market, and its transferability is accordingly limited.⁴²³ Nonetheless, this example demonstrates that a hypothetical-fee measure, unlike actual loss calculations, does not require resolving the fundamental issue presented by Gap 1. Instead, it focuses on what the market would charge, not what Article 4 of the DSM Directive would have provided.

Recovery of unlawfully obtained profits under Article 13 of the IPRED is available in cases of intentional infringement. Here, the challenge of identifying the source of infringement, as discussed in sub-section 6.1.1 regarding injunctions, arises in a different context.⁴²⁴ Similar to the inability to limit an injunction to the infringing party's capabilities within the model, the recovery of unlawfully obtained profits cannot clearly attribute specific portions of the base model's profits to any particular work or to the infringing party's entire body of work. Both issues reflect the same underlying difficulty: the inability to isolate a distinct contribution within an integrated model, whether in the context of injunctions or monetary compensation.

⁴²³ See Chapter 5, sub-section 5.4.1 of this thesis.

⁴²⁴ IPRED, art 13(1)(a); see Chapter 6, sub-section 6.1.1 of this thesis.

The determination of whether a breach is committed “knowingly” for the purposes of the extended remedies under Article 13 of the IPRED is closely connected to the distinction concerning the Failure Modes outlined in sub-section 5.3.3.⁴²⁵ Failure Model A, in which there was either no valid opt-out or its validity was uncertain given Gap 2, encompasses situations where the developer’s belief in the applicability of Article 4 of the DSM Directive, even if mistaken, could plausibly be justified in good faith. In such cases, available remedies are generally limited to compensation for actual losses or hypothetical payments, as previously discussed. In contrast, Failure Model B, which involves a clear and unambiguous reservation that is disregarded, is less amenable to such justification. A developer who proceeds despite an unambiguous reservation faces significant difficulty in asserting that the breach was unintentional, thereby making enhanced damages more readily available.

This dynamic produces a significant asymmetry that merits specific attention, as it challenges the Chapter’s broader conclusions. Section 5.6 finds that Failure Regime B, which involves the disregard of a legitimate mechanism, is more damaging from a prevention standpoint since it leaves a rightholder who has complied fully without further recourse.⁴²⁶ Conversely, from the perspective of damages, the situation is inverted: Failure Mode B also presents the circumstances in which the most robust remedies, such as confiscation of unlawfully obtained profits and increased damages for wilful infringement, are most accessible. This is due to the clarity of the ignored provision, which makes it difficult to argue that the infringement was unintentional. Failure Regime A, which Chapter 4 identified as the most prevalent due to the uncertainty of Gap 2, is also the scenario in which remedies are most likely restricted to compensation for actual losses or hypothetical levies. These remedies are themselves limited, as previously discussed, by the fundamental issue of Gap 1. Consequently, the Failure Regime that inflicts the greatest practical harm (B, as indicated in sub-section 5.3.3) is also the one that provides

⁴²⁵ See Chapter 5, sub-section 5.3.3 of this thesis.

⁴²⁶ See Chapter 5, Section 5.6 of this thesis.

the greatest leverage in litigation. In contrast, the most common regime (A, reflecting the frequency of Gap 2) is characterised by the weakest remedies for rightholders, even if pursued successfully. While this does not alter the conclusion in Section 5.6 (an opt-out that depends on litigation to have any effect has already failed as a preventive measure, regardless of how strong the eventual remedy might be), it does represent a significant refinement of the liability framework developed in this Chapter.

6.2. EU AI Act — Relevance for Copyright Law

Article 53 of the EU AI Act has appeared at three points in this thesis so far: as a partial response to the law enforcement gap identified as Gap 3 in Section 4.6, as the legal basis for the Commission’s consultations on refusal protocols discussed in sub-section 5.2.3, and as relevant context for the significance of the pending *Like Company v Google Ireland Ltd* case noted in Section 5.5.⁴²⁷ Each reference addressed Article 53 for specific, localised purposes. This section examines a broader question: do the EU AI Act and the DSM Directive together constitute a single, coherent regulatory framework for AI training, or do they represent two parallel systems with only incomplete points of contact between them?

The requirements of the EU AI Act for the present analysis can be summarised as follows. Providers of general-purpose AI models must comply with EU copyright law, including, where applicable, identifying and adhering to reservations made under Article 4(3) of the DSM Directive, and must publish a sufficiently detailed description of the content used for training.⁴²⁸ Equally significant are the limitations of the EU AI Act. It does not determine the lawfulness of any specific training activity; a provider may fulfill the AI Act’s documentation and policy requirements while the training described remains, as analysed in Chapters 3–5, unlawful under the DSM Directive. The EU AI Act does not create direct legal remedies for rightholders; a

⁴²⁷ See Chapter 4, Section 4.6 of this thesis; see Chapter 5, sub-section 5.2.3 and Section 5.5 of this thesis.

⁴²⁸ EU AI Act, art 53(1)(c)–(d).

finding of inadequate policy constitutes a breach of the EU AI Act, addressed through its own enforcement mechanisms, but does not itself establish copyright infringement. Furthermore, it does not specify the necessary content of the summary description of training data for copyright purposes, in contrast to the broader transparency objectives of the EU AI Act itself, which extend beyond copyright compliance.⁴²⁹

These three gaps are not isolated phenomena but reflect a fundamental characteristic of the EU's regulatory architecture. The EU AI Act assigns supervision of service providers' obligations under the GPAI provisions, including Article 53, to the AI Office.⁴³⁰ In contrast, copyright enforcement remains the responsibility of rightholders, who must initiate private actions before national courts under the IPRED, as stipulated in sub-section 6.1.3. These are not two components of a unified system, but rather two distinct systems with different institutions, procedures, bodies of law, and remedies, which only partially overlap in subject-matter. A rightholder who obtains, whether from an AI Office finding or from a published Article 53 summary, clear evidence that their work was used in training cannot simply use it in a copyright claim as though the two systems were integrated. While the evidence may be relevant, the rightholder must still independently prove in a national court, applying the DSM Directive as analysed in Chapters 3–5, that the use was unlawful. The structure of the EU AI Act does not make this step automatic.

The consultations scheduled for December 2025, as stated in sub-section 5.2.3, establish these matters in concrete terms. Assuming the consultations are successful, an agreed list of protocols will be published with an opt-out option, and signatories to the GPAI Code of Practice will, as the Commission's framing already anticipates, honour them.⁴³¹ A provider that complies with these protocols fulfils an obligation under the EU AI Act. Nevertheless, whether compliance

⁴²⁹ See Chapter 4, Section 4.6 of this thesis; see Chapter 5, sub-section 5.2.3 of this thesis.

⁴³⁰ EU AI Act, recitals on AI Office establishment and GPAI oversight; see Chapter 6, sub-section 6.1.3 of this thesis (IPRED framework).

⁴³¹ See Chapter 5, sub-section 5.2.3 of this thesis.

with a protocol agreed through the EU AI Act Code of Practice equates to recognising a valid reservation under Article 4 (3) for the separate purposes of DSM Directive litigation remains unresolved. The consultations do not address this question, and, as noted in sub-section 5.2.3, it remains open.⁴³² This situation is not a hypothetical example of institutional fragmentation; rather, it is an ongoing development in the area of standardisation with an opt-out option, where the EU AI Act and the DSM Directive most directly address the same issue.

The unresolved nature of this issue represents a reform challenge that must be addressed in Section 6.4. Resolving this challenge will require an explicit linking provision, whether in the EU AI Act, the DSM Directive, or both, stipulating that conformity decisions taken within one system entail specific legal consequences in the other. In the absence of such a provision, even a fully successful standardisation process, such as the one underway since December 2025, would only address Gap 2 within the scope of the EU AI Act. Its relevance to legal proceedings under the DSM Directive would remain subject to case-by-case determination, specifically through judicial interpretation as represented by *Like Company v Google Ireland* case and discussed in Section 6.3.⁴³³

6.3. Current Litigation Landscape

The cases examined in this section have previously been referenced in this thesis: sub-section 4.4.2 discussed *Kadrey v Meta*, *NYT v Microsoft/OpenAI*, and *Getty Images v Stability AI* as evidence regarding the composition of training datasets; sub-section 5.4.1 addressed *Bartz v Anthropic* in relation to damages; and Section 5.5 considered *Kneschke v LAION* and the pending *Like Company v Google Ireland* reference within the context of the fair balance doctrine.⁴³⁴ These cases are not plainly re-examined here. Instead, this section addresses a new

⁴³² See Chapter 5, sub-section 5.2.3 of this thesis.

⁴³³ See Chapter 6, Section 6.3 of this thesis.

⁴³⁴ See Chapter 4, sub-section 4.4.2 of this thesis; see Chapter 5, sub-section 5.4.1 and Section 5.5 of this thesis.

question: what does this body of litigation, considered as a whole, suggest about how the unresolved questions this thesis has identified are likely to be addressed in practice?

6.3.1. US Litigation — From Factual Record to Legal Signal

The US cases previously discussed in Chapter 4 served to establish that pirate repositories such as LibGen and Books3 were actively utilised by the major AI developers.⁴³⁵ When these cases are reconsidered from the standpoint of legal reasoning rather than factual confirmation, a notable aspect of the Anthropic case emerges. The court's distinction, that training on lawfully acquired books may constitute fair use, while training on pirated copies may not, aligns doctrinally with the multi-factor fair use test.⁴³⁶ Structurally, this mirrors the distinction identified in sub-section 4.4.1 under EU law: a separate, preliminary, and potentially decisive inquiry into the manner in which data was obtained, regardless of whether training on lawfully stored data is permissible in itself.⁴³⁷

This convergence deserves further consideration. Despite following distinct legal pathways, the EU through an explicit textual provision in Article 4 of the DSM Directive, and the US through judicial development of an open standard of fair use, both systems have established the same structural distinction between the lawfulness of acquisition and the admissibility of use. This parallel suggests that the distinction is not merely a product of Article 4's drafting, but rather reflects a broader shared legal intuition regarding the boundary between lawful and unlawful AI training. For the EU, this convergence provides certainty on a single point: notwithstanding the various shortcomings identified in Chapters 4 and 5, the threshold for lawful access is not a unique or arbitrary aspect of the DSM Directive. A common law system reasoning from first principles reached for the same threshold.

⁴³⁵ See Chapter 4, sub-section 4.4.2 of this thesis.

⁴³⁶ Ropes & Gray LLP (n 377); see Chapter 5, sub-section 5.4.1 of this thesis.

⁴³⁷ See Chapter 4, sub-section 4.4.1 of this thesis.

The aspects that are not transferred are as significant as those that are. The central arguments in *NYT v OpenAI* (memorisation and market substitution) address whether the model’s outputs compete with or reproduce the works on which it was trained.⁴³⁸ Sub-section 3.3.4 extends this inquiry beyond the scope of the present thesis, which is primarily concerned with the input side of training.⁴³⁹ The US legal system predominantly emphasises outputs, whereas the EU litigation landscape, as discussed below, focuses on the input side, which aligns with the issues examined in this thesis.

6.3.2. EU Developments — What the Case Law Suggest About Sequencing

The *Kneschke v LAION* and *Like Company v Google Ireland* cases are both introduced in Section 5.5 in relation to the “fair balance” doctrine.⁴⁴⁰ Collectively, they indicate a likely sequence of steps that EU courts may follow when considering the issues analyzed in this thesis.

In the *Kneschke* case, the Hamburg court considered Article 4(3) of the DSM Directive; the parties argued it and the court examined it in detail. However, the judgment was based exclusively on Article 3 of the DSM Directive, with comments on Article 4(3) treated as obiter dictum.⁴⁴¹ The court resolved the case on the simpler issue, as LAION, a non-profit research organisation, fell within the scope of Article 3. Such an approach reflects a broader pattern rather than an isolated decision. When defendants can credibly invoke Article 3 (such as research institutions, organisations dealing with cultural heritage, or entities clearly covered by Article 3 as outlined in Section 4.2), courts have a straightforward means to avoid applying Article 4(3). Consequently, the complex issues discussed in Chapters 3 and 4 of this thesis may remain

⁴³⁸ *The New York Times Co v Microsoft Corp*, No 1:23-cv-11195 (SDNY 2023); see Chapter 4, sub-section 4.4.2 of this thesis.

⁴³⁹ See Chapter 3, sub-section 3.3.4 of this thesis.

⁴⁴⁰ See Chapter 5, Section 5.5 of this thesis.

⁴⁴¹ *Kneschke v LAION* (n 398); *RAILS Blog* (n 398); see Chapter 5, Section 5.5 of this thesis.

unaddressed in court in cases where they are least contentious, and may arise only in proceedings against entities, such as commercial AI developers, for whom Article 3 is inapplicable from the outset.

Judgment *Like Company v Google Ireland* is significant as it does not permit such avoidance.⁴⁴² The reference arises in proceedings against Google Ireland, which is not a research organisation, and is directly linked to Article 4(3) of the DSM Directive, specifically regarding the meaning of “appropriate manner” and the interaction between the exception and the reproduction right.⁴⁴³ These issues were identified as unresolved not only under Gap 2 in sub-section 4.3.4, but also in practice in Section 5.2. While it is not possible to predict how the CJEU will address these matters, the two opposing interpretations highlight the central issue. A narrow reading of “appropriate manner” (requiring something close to the standardisation that Gap 2 found absent) would result in many existing reservation attempts failing to meet the requirements. This would significantly limit the practical applicability of Article 4 and support the criticism of Gap 2 expressed in this thesis. Conversely, a broad interpretation (accepting established conventions such as robots.txt as sufficient regardless of standardisation) would address Gap 2 through judicial interpretation, but would not affect Gap 3 (enforcement) or Gap 1 (compensation).⁴⁴⁴

The limit shared by both readings is the point worth emphasising. Gap 1, the absence of compensation, concerns what Article 4 of the DSM Directive provides once its conditions are met, rather than the meaning of “appropriate manner”. No interpretation of the reservation requirement in Article 4(3), regardless of the Court’s resolution, offers an answer beyond what is stated in the text. The judgment in *Like Company v Google Ireland* will clarify the scope of Gap 2, potentially in a significant way. However, given the nature of the question, it cannot address

⁴⁴² *Like Company v Google Ireland Ltd* (n 397); see Chapter 5, Section 5.5 of this thesis.

⁴⁴³ Oxford Law Blogs (n 397); see Chapter 4, sub-section 4.3.4 of this thesis (Gap 2); see Chapter 5, Section 5.2 of this thesis.

⁴⁴⁴ See Chapter 4, sub-section 4.3.4 of this thesis (Gap 2).

Gap 1, and the enforcement deficit (Gap 3) is, at best, only indirectly addressed by clarifying the requirements. The pending reference is accordingly best understood not as a substitute for the legislative reform discussed in Section 6.4, but as a development that may address only one of the three gaps identified in Section 4.6, leaving the other two, including the one (Gap 1) most directly tied to the fair-balance concerns of Section 5.5, for legislation to address regardless of the Court's decision.⁴⁴⁵

6.4. Future Regulatory Perspectives

The preceding Sections of this Chapter have examined the implications of the gaps identified in Chapters 4 and 5: liability and its limits (Section 6.1), the interaction with the AI Act, which does not address these gaps (Section 6.2), and the litigation landscape which may narrow one of them but cannot resolve the others (Section 6.3). This concluding Section shifts from diagnosis to proposal. Each option considered below is evaluated in relation to the specific gaps it addresses and those it does not, using the vocabulary Gaps 1, 2, and 3 from Section 4.6, as well as the issues of burden allocation, historical data, and global mismatch from section 5.3, employed in this thesis.

6.4.1. The Existing Reform Landscape

Reform does not begin from a blank page. Three processes are already underway, and any proposal must be assessed in relation to these ongoing developments rather than in isolation.

Firstly, Article 25 of the DSM Directive requires the Commission to review the functioning of the Directive, including TDM exemptions, no earlier than 7 June 2026, a date which, at the time of writing, has just passed.⁴⁴⁶ Secondly, the consultations in December 2025, discussed in subsection 5.2.3 and noted in Section 6.2, already constitute a form of such a review, conducted

⁴⁴⁵ See Chapter 4, Section 4.6 of this thesis; see Chapter 5, Section 5.5 of this thesis; see Chapter 6, Section 6.4 of this thesis.

⁴⁴⁶ DSM Directive, art 25; European Parliament, Answer to Parliamentary Question E-000528/2025 (2025).

through the GPAI Code of Practice under the EU AI Act rather than directly through Article 25 of the DSM Directive. This process exemplifies the institutional fragmentation identified in Section 6.2.⁴⁴⁷ The chapter of the Code of Practice on copyright was published in July 2025 alongside a template for training data summaries in accordance with Article 53(1)(d) of the EU AI Act, and signatories to the Code have already begun to fulfil obligations regarding the documentation of datasets and opt-outs.⁴⁴⁸ Thirdly, *Like Company v Google Ireland* case represents a judicial track running parallel to both.⁴⁴⁹

A fourth development has recently been drafted and now stands alongside these three. On 10 March 2026, the European Parliament adopted, by an overwhelming majority, a resolution entitled “Copyright and Generative Artificial Intelligence — Opportunities and Challenges”, based on a report by its Legal Affairs Committee.⁴⁵⁰ Although the resolution is not legally binding and is viewed by its supporters primarily as a political signal rather than as law, its content is directly linked to the conclusions of this thesis. The resolution calls for opt-out mechanisms to be machine-readable and standardised, while explicitly rejecting a centralised opt-out register in the EU. It proposes that failure by AI providers to comply with transparency obligations should give rise to a rebuttable presumption that protected works were used in training. Most notably, the rapporteur’s accompanying text records the view that even in cases where Article 4 of the DSM Directive is deemed inapplicable, AI developers should have access to a straightforward licensing route, with the rationale of Article 17 of the DSM Directive specifically cited as a starting point.⁴⁵¹

⁴⁴⁷ See Chapter 5, sub-section 5.2.3 of this thesis; see Chapter 6, Section 6.2 of this thesis.

⁴⁴⁸ Taylor Wessing, ‘Copyright in 2026: Clarification, Review and Reform’ (2026).

⁴⁴⁹ See Chapter 5, Section 5.5 of this thesis; see Chapter 6, sub-section 6.3.2 of this thesis.

⁴⁵⁰ European Parliament, ‘Report on Copyright and Generative Artificial Intelligence — Opportunities and Challenges’ A10-0019/2026; European Parliament, Resolution of 10 March 2026 on Copyright and Generative Artificial Intelligence — Opportunities and Challenges (2025/2058(INI)), P10_TA(2026)0066.

⁴⁵¹ European Parliament, A10-0019/2026 (n 450); Nomos, ‘European Parliament Resolution: MEPs Want to Strengthen Copyright Protection to Meet the Challenges of Generative Artificial Intelligence’ (2026); European Federation of Journalists, ‘EFJ Welcomes European Parliament’s Call for an EU Regulatory Framework that Protects Copyright in the Age of AI’ (2026).

These four strands (the current obligation to review under Article 25 of the DSM Directive, the consultations conducted under the EU AI Act, *Like Company v Google Ireland*, and the resolution of March 2026) are not competing alternatives to the reform options considered below. Rather, they constitute the channels through which any of these options must proceed. A proposal that builds upon or develops one or more of these strands is likely to receive greater political support than a proposal presented as an independent initiative. This is the standard against which the feasibility assessment in sub-section 6.4.4 is conducted.

6.4.2. Legislative Reform Options

Four options are under consideration, each assessed against the gap-map developed in Chapters 4 and 5.

Option A — Targeted Amendment of Article 4 of the DSM Directive: Standardisation and Compensation

A Commission implementing act that establishes a mandatory technical standard for Article 4(3) of the DSM Directive reservations would address Gap 2, building on the C2PA discussions in sub-sections 4.3.4 and 5.2.3 as well as ongoing consultations.⁴⁵² When combined with a mandatory compensation requirement for commercial TDM, this approach would also address Gap 1. Option A is narrowly focused; it proposes to amend the terms of Article 4 of the DSM Directive without creating new institutions and is closely linked to the consultation process initiated in December 2025. However, it only indirectly addresses Gap 3: while a clearer standard facilitates the identification of infringements, identification alone does not ensure compliance. Furthermore, it does not address the problem of historical data (sub-section 5.3.2) or global inconsistency (sub-section 5.3.4), both of which extend beyond the scope of amending the prospective provisions of Article 4.

⁴⁵² See Chapter 4, sub-section 4.3.4 of this thesis; see Chapter 5, sub-section 5.2.3 of this thesis; see Chapter 4, Section 4.6 of this thesis (Gap 1).

Option B — Mandatory Collective Licensing

The ECL mechanism developed in sub-section 5.4.2 would fully eliminate Gap 1 and, since it does not depend on the actions of individual rightholders, would empower the group identified in sub-section 5.3.1 as least able to exercise their right to opt out of the compensation scheme, regardless of their capacity to act.⁴⁵³ This mechanism leaves Gaps 2 and 3 unaffected, although it reduces their significance: a rightholder receiving compensation through the ECL scheme is protected even if their individual attempt to opt out is unsuccessful. The infrastructure for Option B is already in place, as established in sub-section 5.4.2; however, a legal mandate to apply it to AI training does not yet exist and would require an extension of Article 12 of the DSM Directive. Resistance from the AI industry to such an extension is possible, but the Anthropic agreement (sub-section 5.4.1) suggests that the industry’s de facto position may be shifting from “no compensation” to “compensation through litigation”, as discussed in sub-section 6.4.4.⁴⁵⁴

Option C — A New Sui Generis Right for AI Training Data

A purpose-built right, distinct from copyright and analogous to the sui generis database right stipulated in sub-section 4.5.2, could in principle address all three gaps through its initial design rather than through subsequent amendment.⁴⁵⁵ This also constitutes a limitation for assessment: as no such right exists in any draft, its evaluation remains necessarily abstract, unlike Options A and B. Sub-section 4.5.2 demonstrates that EU law can support parallel regimes, as copyright and database rights coexist, though not always without friction. However, their interaction raises interpretive questions. A sui generis AI-training right would raise the same challenge in a new form: it is unclear whether Article 4 of the DSM Directive would continue to apply to the non-AI TDM, and what would be the relationship between the two regimes. The advantages

⁴⁵³ See Chapter 5, sub-sections 5.3.1 and 5.4.2 of this thesis.

⁴⁵⁴ See Chapter 5, sub-section 5.4.1 of this thesis; see Chapter 6, sub-section 6.4.4 of this thesis.

⁴⁵⁵ See Chapter 4, sub-section 4.5.2 of this thesis.

of Option C are genuine yet largely theoretical; its drawbacks include the scale of the legislative effort required (a new Directive rather than an amendment) and the practical challenges, as outlined in sub-section 4.5.2, associated with introducing a second regime alongside the existing one.

Option D — A Harmonised Liability Safe Harbour

Modelled on Article 17 of the DSM Directive’s structure for online content-sharing platforms, would introduce a liability shield, rather than an exemption, conditional upon compliance with recognised standards of opt-out and transparency obligations. This approach could potentially address Gaps 2 and 3 simultaneously.⁴⁵⁶ Gap 1 remains unaddressed unless compensation is built into the shield’s conditions, in which case the option converges with a hybrid of Options A and B. Article 17 serves as a precedent for such conditional defenses in EU copyright law, but it is also widely recognised as one of the most controversial provisions of the DSM Directive. Transposing its structure risks importing its contradictions. The risk of manipulation is particularly acute here in relation to Gap 2: a safe harbour conditional on recognition of standardised opt-out options is only as reliable as the standard itself. Empirical data in Section 5.2 on compliance gaps, even for widely used mechanisms such as robots.txt, suggest that conditioning a liability shield on recognition of a standard that is unevenly complied with in practice may result in a defense that protects formal rather than substantive compliance.

Notably, the March 2026 Parliament resolution’s invocation of Article 17 of the DSM Directive as a reference point for a post-Article-4 licensing route, as discussed in sub-section 6.4.1, provides Option D with a degree of institutional support not originally anticipated.⁴⁵⁷ Although it is not possible to determine from the resolution’s text whether the authors intended to reference

⁴⁵⁶ DSM Directive, art 17.

⁴⁵⁷ See Chapter 6, sub-section 6.4.1 of this thesis.

Option D specifically or to justify Article 17 more generally, this coincidence indicates that Article 17 is already part of the vocabulary used by EU institutions in this context.

6.4.3. A Boundary Case — The Japanese Model

In 2018, Japan amended Article 30-4 of its Copyright Act to introduce a TDM exception covering the commercial AI development, without an opt-out mechanism. This amendment constitutes the most liberal approach to AI training among major jurisdictions currently in force.⁴⁵⁸

Rather than serving as a model for implementation, the Japanese example is best understood as a borderline case that positions existing EU options and previously discussed alternatives along a single spectrum. Japan represents the “permission, no opt-out, no compensation” endpoint, whereas the current Article 4 of the DSM Directive (opt-out, no compensation) is situated further along this spectrum. Options B and D would shift the EU even further, toward “permission conditioned on compensation or compliance”.

The rationale for the EU not adopting the Japanese end of this spectrum is grounded in constitutional structure rather than policy preference. The analysis in Section 5.5 is predicated on the recognition of the status of Article 17(2) of the CFREU as a fundamental right, which must be balanced against the freedom to conduct a business as provided in Article 16 of the same Charter.⁴⁵⁹ A TDM exception with neither opt-out nor compensation would disrupt this balance by disregarding one side of the equation. According to the fair balance analysis applied in this thesis, such a regime would be less justifiable than the current Article 4 of the DSM Directive. While Japan’s approach may be coherent within a legal order that does not grant IP the same constitutional status, it cannot be assumed to be applicable to the EU without first addressing the relationship between Articles 16 and 17(2) of the CFREU, which serves as the evaluative framework from Section 5.5 onward.

⁴⁵⁸ Copyright Act (Japan), art 30-4, as amended by Act No. 30 of 2018.

⁴⁵⁹ See Chapter 5, Section 5.5 of this thesis.

6.4.4. Preferred Reform Proposal: A Combination of Options A and B

The gap-map presented in sub-section 6.4.2 points toward a combination rather than a single option. Option A alone leaves Gap 3 largely unresolved and fails to achieve the burden-allocation asymmetry outlined in sub-section 5.3.1. Option B alone does not address Gaps 2 and 3. When implemented together, these options are mutually reinforcing: standardisation under Option A supplies the technical foundation required for the collective licensing scheme in Option B to determine, at scale, which uses fall within its scope. The entity administering the TDM compensation scheme depends on such standards to identify both the nature and the users of relevant materials, a function fulfilled by Option A. In turn, compensation under Option B addresses the needs of the group identified in sub-sections 5.3.1 and 5.4.3, for whom even a fully standardised and effectively implemented opt-out mechanism is inadequate, as their primary interest is in remuneration rather than prevention of use. This combined approach was previously outlined at the conclusion of Section 5.5, where it was framed as a logical extension of the DSM Directive's existing trajectory rather than a deviation; it is at this point that the preliminary description is advanced as a formal proposal.⁴⁶⁰

Two significant limitations must be articulated as clearly as the proposal itself. First, the historical-data problem (sub-sections 4.3.5 and 5.3.2) remains unaffected by any prospective amendments to Article 4 of the DSM Directive, as the training data embedded in existing models were generated under conditions that cannot be retroactively altered by future standards or compensation schemes. This thesis does not offer a solution for this aspect, except to reiterate, as noted in sub-section 4.3.5, that a distinct, retrospective mechanism (such as a one-off fund for historical use) would constitute a separate measure from amending Article 4. The Anthropic Agreement (sub-section 5.4.1) demonstrates that such retrospective authorisation is not unprecedented in other legal systems, even if its implementation within the EU remains unspecified.⁴⁶¹

⁴⁶⁰ See Chapter 5, sub-sections 5.3.1, 5.4.3 and Section 5.5 of this thesis.

⁴⁶¹ See Chapter 4, sub-section 4.3.5 of this thesis; see Chapter 5, sub-section 5.4.1 of this thesis.

Second, the problem of global non-conformity (sub-section 5.3.4) is also not addressed: amendments to EU legislation cannot extend the EU's jurisdiction. Addressing this limitation, if at all possible, would require international cooperation or deployment-based mechanisms as discussed in sub-section 5.3.4, neither of which is further examined in this thesis.

The proposed approach also has implications for the three-step test outlined in Article 5(5) of the InfoSoc Directive, which any exception (whether existing or amended) must satisfy: limitation to specific cases, no conflict with normal exploitation, and no unreasonable prejudice to the legitimate interests of rightholders.⁴⁶² The current, uncompensated Article 4 of the DSM Directive is least compatible with the third criterion, as uncompensated commercial use is a clear example of “unreasonable prejudice”, even though Article 4 has not been repealed on this basis. Introducing compensation through Option B directly addresses this concern, a result that Option A alone cannot achieve. Although the three-step test and the Charter's fair balance doctrine (Section 5.5) are formally distinct, one serving as a condition for exceptions under the InfoSoc Directive, the other as a constitutional balancing requirement, they both indicate that a compensated Article 4 is more consistent with these frameworks than the current uncompensated version.

In terms of political feasibility, the context described in sub-section 6.4.1 can be regarded as favourable to this combined approach. Standardisation, the central element of Option A, is not a novel proposal introduced by this paper; it is currently under active discussion and, as indicated by the March 2026 resolution, is an explicit requirement of the European Parliament.⁴⁶³ Compensation, the core of Option B, continues to encounter resistance from industry stakeholders who benefit from the current uncompensated regime under Article 4 of the DSM Directive. However, two recent developments are reducing this opposition. First, the Anthropoc settlement

⁴⁶² InfoSoc Directive, art 5(5).

⁴⁶³ See Chapter 6, sub-section 6.4.1 of this thesis.

illustrates that the alternative to a legislatively mandated compensation mechanism is not the absence of compensation, but rather compensation determined through litigation, which is less predictable and outside industry control (a \$1.5 billion outcome that AI developers might have preferred to avoid through a statutory levy).⁴⁶⁴ Second, the rapporteur for the March 2026 resolution explicitly considered a licensing model operating alongside or instead of Article 4 of the DSM Directive, using the rationale of Article 17 as a reference point. This suggests that a system of permits contingent upon payment or compliance, a structural element shared by Option B and Option D, and closely related to Article 17, is no longer solely a theoretical proposal of this thesis but is now under discussion in various forms in Strasbourg.⁴⁶⁵ Channelled through the Article 25 review of the DSM Directive being now formally underway, a combined A+B reform is not a departure from the EU's current trajectory but an articulation of where several of its current strands already point.

⁴⁶⁴ See Chapter 5, sub-section 5.4.1 of this thesis.

⁴⁶⁵ See Chapter 6, sub-sections 6.4.1 and 6.4.2 of this thesis.

7. Conclusion

7.1. Summary of Findings by Research Questions

RQ1: Does the training of AI models constitute copyright-relevant reproduction under EU law?

Chapter 3 demonstrates that the training of AI models constitutes a form of reproduction subject to copyright under Article 2 of the InfoSoc Directive at multiple stages of the training process. Applying the originality standard established in the Infopaq International case and refined in subsequent court rulings, such as Painer and Cofemel, this thesis establishes that web scraping, dataset storage, and data pre-processing are acts of reproduction within the meaning of Article 2 of the InfoSoc Directive. The vast majority of content processed in this manner reflects the author's own intellectual creation and is therefore eligible for protection. The exception for temporary reproduction under Article 5 of the InfoSoc Directive, considered as a possible alternative basis for these acts, does not apply. The storage of datasets and the resulting model weights are permanent rather than temporary or incidental; the reproduction serves several simultaneous commercial purposes rather than a single purpose as required by Article 5(1); and it possesses independent economic significance, which constitutes the entire source of value of the resulting model. The conclusion from RQ1 is clear: in the absence of a licence or an applicable statutory exception, reproducing copyright-protected works for AI training infringes the rights of the relevant rightholders under Article 2 of the InfoSoc Directive.

RQ2: To what extent do the TDM exceptions of the DSM Directive permit the use of protected works for AI training purposes?

Chapter 4 examined whether the TDM exceptions introduced by the DSM Directive provide a basis for a licence or exception for AI training. The analysis demonstrates that Article 3 of the DSM Directive, which applies to TDM mining conducted by research organisations and cultural heritage institutions for scientific research, does not extend to AI development by commercial

developers for commercial purposes, and is therefore unavailable to the entities whose activities form the focus of this thesis. In contrast, Article 4 of the DSM Directive is, in principle, capable of covering commercial AI training, as it is not limited by the purpose or status of the beneficiary. However, its application is subject to two cumulative conditions: the reproduction must be carried out on works to which the developer has lawful access, and the rightholder must not have exercised the right of refusal provided in Article 4(3) of the DSM Directive. In cases where training data is obtained from sources known or reasonably believed to be unauthorised, such as datasets collected from pirate book repositories and similar sources, neither Article 3 nor Article 4 of the DSM Directive can legitimise the resulting reproduction, regardless of the opt-out issue. Overall, the answer to RQ2 reveals an incomplete and structurally inadequate system that does not correspond to the scale and nature of modern commercial AI training. Article 4 of the DSM Directive may, under favourable circumstances, provide a legal basis for training using lawfully accessible materials from which no opt-out has been received, yet it operates within a framework not designed for foundation models and leaves significant categories of training activities without any clear legal basis.

RQ3: How effective is the opt-out mechanism under Article 4 of the DSM Directive in balancing innovation interests and rightholders protection?

Chapter 5 addresses the sole mechanism within this system that the legislator appears to have intended as the primary safeguard for rightholders, namely the opt-out mechanism under Article 4(3) of the DSM Directive, and assesses its practical effectiveness. The results are consistently negative. The absence of any mandatory technical standard for expressing machine-readable opt-outs creates a fragmented and uncertain landscape of competing conventions, with no guarantee that an opt-out expressed in one format will be recognised or complied with by any particular AI developer. The burden of proof inherent in the opt-out mechanism places the onus on individual rights holders to identify, track, and enforce their opt-outs across an indeterminate

number of datasets and developers. This stands in stark contrast to the scale and resources available to AI developers, whose conduct the mechanism is designed to restrict. Furthermore, this mechanism is unable to address historical training data. Works already incorporated into existing models prior to the opt-out declaration remain embedded in the model weights, and no retroactive measures are provided for this category of use. The analysis also reveals an almost complete absence of enforcement mechanisms specific to the opt-out itself, leaving rights holders reliant on general intellectual property protection tools that were not designed for the context of AI training and are poorly suited to it. Collectively, these findings support the conclusion that the opt-out mechanism in its current form is structurally inadequate to achieve the intended balance between innovation interests and the protection of rights holders' rights under Article 4(3) of the DSM Directive, and that reform of this mechanism is urgently needed.

7.2. Overall Assessment

Drawing the findings of the three research questions together, the overall assessment indicates that EU copyright law establishes a framework for regulating AI training that is both incomplete in scope and inadequate for addressing the scale and commercial significance of current practices. The framework is incomplete in the sense that it does, in principle, offer certain pathways (most notably Article 4 of the DSM Directive) through which some AI training activities may be reconciled with the exclusive rights set out in the InfoSoc Directive. However, these pathways are narrow, conditional, and poorly aligned with the realities of foundational model development. Specifically, the lawful access requirement excludes datasets sourced without authorisation, the opt-out mechanism is weakened by the deficiencies discussed in Chapter 5, and neither Article 3 nor Article 4 of the DSM Directive was designed to accommodate the scale, commercial interests, or technical characteristics of contemporary LLM training.

For major AI developers operating within or targeting the EU market, this combination of partial legal coverage and structural inadequacy results in significant legal exposure under existing

legislation. When training corpora include copyright-protected works obtained without lawful access, or when valid opt-out made under Article 4(3) of the DSM Directive is disregarded or cannot be practically honoured due to insufficient technical standardisation, the reproduction of these works during training exceeds the scope of any available exception and therefore constitutes, *prima facie*, a violation of Article 2 of the InfoSoc Directive. The magnitude of such infringement, given that training corpora may contain hundreds of billions of documents, is unprecedented in copyright law. The legal proceedings examined in Chapter 5, including a significant settlement involving a major AI developer in the US, demonstrate the potential scale of liability that similar claims could generate within the EU.

It should be noted that the DSM Directive was adopted in 2019, when TDM provisions were primarily intended for regulation of scientific research and traditional data analysis, rather than for establishing legislative control over training of generative foundation models at the scale now observed. While the drafters of the DSM Directive could not have anticipated developments that emerged years after its adoption, the resulting legal framework, regardless of interpretative efforts, was not designed to address the current challenges. The transparency obligations introduced by Article 53 of the EU AI Act constitute a positive and complementary development, as they may, over time, enhance the ability of rightholders to determine whether and how their works have been used in training. Nevertheless, these obligations function alongside, rather than as a substitute for, the copyright system and do not resolve the core issues of legality, licensing, or compensation discussed in this thesis. Transparency, in the absence of a substantive legal framework to which it can be effectively connected, addresses only a portion of the overall problem.

7.3. Reform Recommendations

On the basis of the findings summarised above, this thesis advances the following consolidated reform recommendations.

Firstly, the European Commission should adopt an implementing act to establish a mandatory, harmonised technical standard for the machine-readable expression of opt-outs from TDM, in accordance with Article 4(3) of the DSM Directive. As noted in Chapter 5, the absence of such a standard is a primary factor contributing to the practical ineffectiveness of the opt-out mechanism. Implementing a mandatory standard based on existing technical proposals, such as the Robots Exclusion Protocol and emerging provenance specifications, would address this deficiency directly, without requiring amendments to the DSM Directive's core provisions.

Secondly, Article 4 of the DSM Directive should be amended to introduce a mandatory compensation mechanism for commercial TDM, including AI training, within its scope. This amendment would address the identified asymmetry between the value generated by AI training and the absence of compensation for rightholders whose works contribute to that value. It would also align the approach to commercial TDM with compensation-based models established in other exceptions and limitations from adjacent spheres of EU copyright law.

Thirdly, the EU should establish a collective licensing mechanism tailored for AI training, building upon the extended collective licensing system outlined in Article 12 of the DSM Directive. This mechanism would provide a practical solution to the transaction costs that impede individual licensing at the scale necessary for training foundational models, while ensuring that rightholders who have not opted out continue to receive remuneration through collective management organisations.

Fourthly, an explicit harmonisation provision should be introduced to clarify the relationship between the transparency obligations under Article 53 of the EU AI Act and the provisions of the DSM Directive concerning TDM. Currently, both instruments operate in parallel, without any formal mechanism linking the information disclosed under the AI Act to the exercise or

enforcement of rights under the DSM Directive; an explicit linking provision would allow transparency obligations to play a genuinely functional role within the copyright framework, rather than existing as a standalone disclosure requirement.

Fifthly, the Commission should establish a mechanism for regular review in the form of a regular assessment of the application of provisions on TDM to AI training. This assessment should draw on information obtained under the AI Act's transparency regime, the development of case law, including the expected referral to the CJEU discussed in Chapter 5, and the evolving licensing landscape. Given the rapid development of both technologies and associated commercial practices, a static legislative framework risks becoming obsolete in the short term. A structured review mechanism would provide the institutional capacity to identify and respond to such changes before they result in the gap between law and practice identified in this thesis with regard to the DSM Directive in its current form.

7.4. Broader Significance

The findings of this thesis should not be read as suggesting that copyright law and the development of AI are inherently irreconcilable objectives. Rather, the analysis in the preceding Chapters demonstrates that the current legal instruments were developed for a different technological era. The contradictions identified between AI training and the rights of authors and other rightholders stem from legislative intent and timing, not from fundamental legal principles. A framework that requires AI developers to obtain lawful access to training data, respects rightholders' decisions to preserve their rights, and allocates a portion of the value generated by AI systems to the original creators is both conceptually coherent and commercially feasible. This approach represents an extension of principles already established in EU copyright law to a new and economically significant context.

The EU has the institutional capacity to establish such a framework. The adoption of the InfoSoc Directive, the DSM Directive, and the EU AI Act within a relatively short timeframe demonstrates the Union's responsiveness to successive technological changes. Furthermore, the consultations and resolutions discussed in Chapter 5 indicate that EU institutions are attentive to the issues addressed in this thesis and are actively considering further action. The remaining challenge is not one of legal or technical feasibility, but of political will: whether the Union is prepared to address the identified gaps and contradictions through concrete legislative amendments before the accumulation of litigation, licensing practices, and divergent national case law forecloses the space for a coherent, Union-wide solution.

It is in this regard that the present thesis seeks to make its contribution. By systematically tracking the application of the reproduction right and TDM exceptions to the AI training process, by assessing the opt-out mechanism in light of practical realities, as documented in recent litigation and regulatory developments, and by translating the findings into a concrete set of reform proposals, this thesis aims to lay part of the academic foundation upon which the legislative clarification required in this area can be built. The issues considered here are likely to continue to evolve rapidly in the months and years following the submission of this thesis; what should not change, however, is the underlying recognition that the law's response to these questions is a matter of choice, and that the choice remains open.

8. Bibliography

Books

- Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (Oxford University Press 2020).
- Thomas Dreier and Bernt Hugenholtz (eds), *Concise European Copyright Law* (2nd edn, Kluwer Law International 2016).
- Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning* (MIT Press 2016).
- Sebastian Raschka, *Build a Large Language Model (From Scratch)* (Manning 2024).
- Jörg Reinbothe and Silke von Lewinski, *The WIPO Treaties on Copyright: A Commentary* (2nd edn, Oxford University Press 2015).
- Emanuela Rosati, *Copyright and the Court of Justice of the European Union* (Edward Elgar 2019).
- Emanuela Rosati, *Originality in EU Copyright: Full Harmonization through Case Law* (Edward Elgar 2013).

Journal Articles

- Stella Biderman and others, ‘Emergent and Predictable Memorization in Large Language Models’ (Advances in Neural Information Processing Systems 37, 2023)
- Nicolas Carlini and others, ‘Extracting Training Data from Large Language Models’ (USENIX Security Symposium 2021) 2633.
- R H Coase, ‘The Problem of Social Cost’ (1960) 3 Journal of Law and Economics 1.
- Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, ‘The Exception for Text and Data Mining in the DSM Directive’ (2021) 52 IIC 174.
- Patrick Henderson and others, ‘Foundation Models and Fair Use’ (2023) 24 Journal of Machine Learning Research 1.
- Peter Bernt Hugenholtz, ‘The Database Right after Ryanair’ (2020) 51 IIC 1019.
- Peter Bernt Hugenholtz, ‘The New Copyright Directive: The EU Content Directive (Directive 2019/790)’ (2021) 7 JIPITEC 3.
- Matthias Leistner, ‘Urheberrecht an der Schnittstelle zwischen Online-Plattformen und Nutzern (2020)’ 122 GRUR 1.

- Thomas Margoni and Martin Kretschmer, ‘A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology’ (CREATe Working Paper 2022/1, University of Glasgow 2022).
- Milad Nasr and others, ‘Scalable Extraction of Training Data from (Production) Language Models’ (2023) arXiv:2311.17035.
- Martin Senftleben, ‘Breathing Space for Cloud-Based Business Models’ (2013) 4 JIPI-TEC 87.
- Martin Senftleben, ‘The New EU Text and Data Mining Exceptions’ (2021) 70 GRUR International 1.
- Martin Senftleben and Laurens Buijelaar, ‘Robot Creativity: An Incentive-Based Neighbouring Rights Approach’ (2020) 42 EIPR 797.
- Benjamin Sobel, ‘Artificial Intelligence as a Threat to Copyright and How the Law Should Respond’ (2023) 26 Stanford Technology Law Review 1.
- 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 Journal of Intellectual Property Law & Practice 453.

EU Legislation

- Charter of Fundamental Rights of the European Union [2012] OJ C326/391 (CFREU)
- 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).
- Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights [2004] OJ L157/45 (IPRED).
- Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market (DSM Directive).
- Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L77/20 (Database Directive).
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1 (GDPR).

- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence [2024] OJ L1689/1 (EU AI Act).
- Treaty on the Functioning of the European Union [2012] OJ C326/47 (TFEU).

Other Legislation

- Copyright Act (Japan), art 30-4, as amended by Act No. 30 of 2018.
- 17 USC § 107 (US Copyright Act).

Cases (CJEU)

- Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569.
- Case C-302/10 *Infopaq International A/S v Danske Dagblades Forening (Infopaq II)* [2012] ECLI:EU:C:2012:16.
- Case C-145/10 *Eva-Maria Painer v Standard VerlagsGmbH* [2011] ECR I-12533.
- Case C-360/13 *Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd* [2014] ECLI:EU:C:2014:1195.
- Case C-476/17 *Pelham GmbH v Ralf Hütter* [2019] ECLI:EU:C:2019:624.
- Case C-469/17 *Funke Medien NRW GmbH v Bundesrepublik Deutschland* [2019] ECLI:EU:C:2019:623.
- Case C-683/17 *Cofemel — Sociedade de Vestuário SA v G-Star Raw CV* [2019] ECLI:EU:C:2019:764.
- Joined Cases C-403/08 and C-429/08 *Football Association Premier League Ltd v QC Leisure* [2011] ECR I-9083.
- Case C-607/11 *ITV Broadcasting Ltd v TV Catchup Ltd* [2013] ECLI:EU:C:2013:147.
- Case C-255/02 *Halifax plc v Commissioners of Customs and Excise* [2006] ECR I-1609.
- Case C-275/06 *Productores de Música de España (Promusicae) v Telefónica de España SAU* [2008] ECR I-271.
- Case C-360/10 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85.
- Case C-250/25 *Like Company v Google Ireland Ltd* (reference lodged 3 April 2025, pending).

Cases (National Courts / US)

- *Kadrey v Meta Platforms Inc*, No 3:23-cv-03417 (ND Cal 2023).
- *The New York Times Co v Microsoft Corp*, No 1:23-cv-11195 (SDNY 2023).
- *Getty Images (US) Inc v Stability AI Ltd*, No 1:23-cv-00135 (D Del 2023).
- *Bartz v Anthropic PBC*, No 3:24-cv-05417-WHA (ND Cal, settlement preliminarily approved 25 September 2025).
- *Robert Kneschke v LAION e.V.*, Landgericht Hamburg, 27 September 2024.

EU Institutional Documents

- European Commission, ‘Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market’ COM(2016) 593 final, 14 September 2016.
- European Commission, ‘Impact Assessment on the Modernisation of EU Copyright Rules’ SWD(2016) 301 final, 14 September 2016.
- European Commission, ‘Communication on Artificial Intelligence for Europe’ COM(2018) 237 final, 25 April 2018.
- European Commission, ‘Commission Launches Consultation on Protocols for Reserving Rights from Text and Data Mining under the AI Act and the GPAI Code of Practice’ (Shaping Europe’s Digital Future, 1 December 2025).
- European Parliament, Committee on Legal Affairs, ‘Report on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market’ (A8-0245/2018, 2018).
- European Parliament, Answer to Parliamentary Question E-000528/2025 (2025).
- European Parliament, ‘Report on Copyright and Generative Artificial Intelligence — Opportunities and Challenges’ A10-0019/2026.
- European Parliament, Resolution of 10 March 2026 on Copyright and Generative Artificial Intelligence — Opportunities and Challenges (2025/2058(INI)), P10_TA(2026)0066.

Technical Standards

- Internet Engineering Task Force, *Robots Exclusion Protocol*, RFC 9309 (2022).
- Coalition for Content Provenance and Authenticity, *C2PA Technical Specification* (2023).

Online/News Sources

- Authors Alliance, ‘What Happens When Your Publisher Licenses Your Work for AI Training?’ (2025).
- Citevera, ‘The Complete robots.txt Guide for AI Crawlers in 2026’ (2 May 2026).
- Civil Society Alliances for Digital Empowerment, ‘Commission Opens Consultation on Machine-Readable Copyright Opt-Out Protocols under the AI Act’ (16 December 2025).
- Contently, ‘How Do I Get into Training Datasets Used by ChatGPT or Claude?’ (19 November 2025).
- Coronium, ‘The Closing Web in 2026: How AI Crawler Blocking and Pay-Per-Crawl Changed Web Scraping’ (2026).
- Digital Applied, ‘AI Crawler & Bot Traffic 2026: The Data’ (2026).
- Digital Applied, ‘AI Crawler Access Control: The 2026 Decision Matrix’ (2026).
- eMarketer, ‘OpenAI to Fund Four Axios Newsrooms in Content Licensing Deal’ (2025).
- European Federation of Journalists, ‘EFJ Welcomes European Parliament’s Call for an EU Regulatory Framework that Protects Copyright in the Age of AI’ (10 March 2026).
- Fieldfisher, ‘Anthropic’s \$1.5 Billion Copyright Settlement with Authors Could Set Precedent in AI Industry’ (16 October 2025).
- Inside Privacy, ‘European Commission Launches Consultations on the EU AI Act’s Copyright Provisions and AI Regulatory Sandboxes’ (18 December 2025).
- Rob Kelly, ‘AI Content Licensing Deals: June 2026 Update’ (Media and the Machine, June 2026).
- Morrison Foerster, ‘To Scrape or Not to Scrape? First Court Decision on the EU Copyright Exception for Text and Data Mining in Germany’ (4 October 2024).
- No Hacks, ‘The AI User-Agent Landscape in 2026: A Complete Reference’ (13 April 2026).
- Nomos, ‘European Parliament Resolution: MEPs Want to Strengthen Copyright Protection to Meet the Challenges of Generative Artificial Intelligence’ (31 March 2026).
- Oxford Law Blogs, ‘AI and the Redefinition of Law in Code: The Future of Legal Intention under Article 4 DSM’ (15 December 2025).
- RAILS Blog, ‘Landmark Court Decision in the EU: Copyright Permissibility of Text and Data Mining for the Purpose of AI Training’ (5 January 2025).

- Ropes & Gray LLP, ‘Anthropic’s Landmark Copyright Settlement: Implications for AI Developers and Enterprise Users’ (8 September 2025).
- Soar, ‘AI Bots robots.txt Guide: GPTBot, ClaudeBot, PerplexityBot’ (13 April 2026).
- Society for Computers and Law, ‘European Commission Consults on Protocols for Reserving Rights from Text and Data Mining under the AI Act and the GPAI Code of Practice’ (5 December 2025).
- SoftwareSeni, ‘How to Govern AI Crawler Access to Your Website in 2026’ (20 February 2026).
- Susman Godfrey LLP, ‘Susman Godfrey Secures \$1.5 Billion Settlement in Landmark AI Piracy Case’ (2025).
- Taylor Wessing, ‘Copyright in 2026: Clarification, Review and Reform’ (16 February 2026).
- Will Scott, ‘How AI Licensing Deals Determine Search Visibility in 2025’ (4 October 2025).