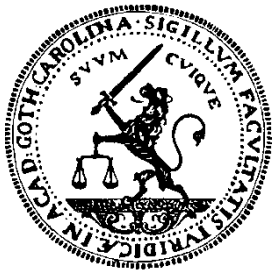

FACULTY OF LAW

LUND UNIVERSITY

Sara Axelsson

AI Training in a Legal Patchwork

Examining Generative AI Training on Images under EU Copyright, Data Protection and Personality Rights

JURM02 Graduate thesis

Graduate thesis, Master of Laws program

30 higher education credits

Supervisor: Ana Nordberg

Semester: Fall 2025

Table of contents

Summary

This thesis examines the scope and practical functioning of the EU legal framework governing the use of image datasets for the training of generative artificial intelligence (GenAI) models, with particular focus on copyright law, and its interaction with privacy, and personality rights. The rapid development of GenAI systems capable of generating high-quality images has intensified legal uncertainty surrounding large-scale data-driven training practices, which frequently rely on reproductions of copyrighted works and the processing of personal data.

The analysis demonstrates that the training of GenAI models on images fall within the text and data mining (TDM) exceptions set out in Articles 3 and 4 of the Directive on Copyright in the Digital Single Market (CDSM Directive). This conclusion is supported by the broad definition of TDM in the directive and reinforced by the Artificial Intelligence Act's explicit reference to the opt-out mechanism in Article 4. However, the thesis shows that significant interpretative and practical uncertainties remain. In particular, unresolved questions concerning lawful access, data retention, and the technical feasibility of effective opt-outs undermine the predictability and workability of the framework, especially in the context of large-scale AI training.

The thesis further assesses the compatibility of the TDM exceptions with the international three-step test. While Articles 3 and 4 CDSM are, in abstract terms, capable of satisfying the test's requirements, their application to GenAI training raises concerns where training practices risk interfering with the normal exploitation of works or disproportionately affecting rightsholders' legitimate interests. The lack of judicial guidance on how the purpose of GenAI models should factor into this assessment contributes to ongoing legal uncertainty.

Beyond copyright, the thesis analyses the role of personality and privacy rights in regulating AI training. It concludes that personality rights offer limited and uneven protection at the training stage, largely due to their fragmented national regulation and frequent reliance on publication or commercial use thresholds. As a result, the protection of individuals whose images are included in training datasets is primarily mediated through EU data protection law. The General Data Protection Regulation (GDPR) imposes an additional, cumulative layer of constraints, requiring lawful processing, necessity, and proportionality, most often assessed through the legitimate interest ground. While GenAI training may, in many cases, rely on legitimate interests, the technical characteristics of AI systems complicate compliance with data subject rights such as erasure, objection, and access.

Taken together, the thesis shows that the EU currently regulates GenAI training through the parallel application of legal frameworks developed for

different purposes and without GenAI in mind. Although both copyright law and data protection law contain internal balancing mechanisms, their cumulative operation produces fragmentation, legal uncertainty, and uneven protection at the training stage. The thesis concludes that further judicial clarification or targeted regulatory development is necessary to ensure a coherent and workable balance between innovation, copyright protection, and fundamental rights in the context of generative AI training.

Sammanfattning

Denna uppsats analyserar tillämpningsområdet och de praktiska utmaningarna i EU-rättens reglering av användningen av bilder vid träning av generativ artificiell intelligens (AI). Fokus ligger på upphovsrätten, och dess samspel med dataskydd enligt GDPR och skyddet för personliga rättigheter. Den snabba utvecklingen av generativa AI-system som kan generera bilder av hög kvalitet har aktualiserat rättsliga frågor kring storskaliga, datadrivna träningsprocesser, vilka ofta innefattar kopiering av upphovsrättsligt skyddade verk och behandling av personuppgifter.

Analysen visar att träning av generativa AI-modeller på bilder omfattas av undantagen för text- och datautvinning (TDM) i artiklarna 3 och 4 i direktivet om upphovsrätt på den digitala inre marknaden (CDSM-direktivet). Denna bedömning grundas på den breda definitionen av text- och datautvinning i direktivet och beaktandet av AI-förordningens hänvisning till opt-out-mekanismen i artikel 4. Samtidigt kvarstår tolknings- och tillämpningssvårigheter. Frågor rörande laglig tillgång, lagringstid och den tekniska möjligheten att genomföra effektiva opt-out-lösningar påverkar regelverkets förutsebarhet, särskilt vid storskalig AI-träning.

Uppsatsen analyserar även TDM-undantagens förhållande till det internationella trestegstestet. Även om artiklarna 3 och 4 CDSM i princip kan anses förenliga med trestegstestets krav, aktualiseras tolkningsfrågor när träningsanvändningar riskerar att påverka det normala utnyttjandet av verk eller rättighetshavarnas legitima intressen. Avsaknaden av rättspraxis om hur generativa AI-modellers syfte ska beaktas i denna bedömning bidrar till rättslig osäkerhet.

Vidare behandlas personlighetsskyddets och integritetsskyddets betydelse för regleringen av AI-träning. Analysen visar att personlighetsskyddet har begränsad räckvidd i träningsfasen, bland annat till följd av fragmenterad nationell reglering och krav på publicering eller kommersiell användning. Skyddet för personer vars bilder ingår i träningsdataset tillgodoses därför i huvudsak genom dataskyddsförordningen (GDPR). GDPR ställer kumulativa krav på laglig behandling, nödvändighet och proportionalitet, vilka i regel bedöms inom ramen för grunden berättigat intresse. Samtidigt medför de generativa AI-systemens tekniska utformning praktiska svårigheter i förhållande till registrerades rättigheter, såsom rätten till radering, invändning och tillgång.

Sammantaget visar uppsatsen att träning av generativ AI inom EU regleras genom parallell tillämpning av rättsliga ramverk som utvecklats för olika ändamål och utan generativ AI i åtanke. Även om upphovsrätten och dataskyddsrätten båda innehåller mekanismer för intresseavvägningar, leder deras samlade tillämpning till fragmentering, begränsad förutsebarhet och ett

ojämnt skydd i träningsfasen. Uppsatsen pekar därmed på behovet av ytterligare rättslig vägledning eller en mer riktad reglering för att säkerställa en fungerande balans mellan innovation, upphovsrättsskydd och grundläggande rättigheter vid träning av generativ AI.

Preface

Snart sex år i Lund flugit förbi, och det är med både glädje och vemod som jag nu avslutar den här fantastiska tiden i livet. För den finns det många jag vill tacka.

Thank you to my exceptional supervisor Ana, for all the great advice and discussions throughout the writing of this thesis.

Det största av tack till alla mina underbara vänner i Lund, de som var med från början och de som har kommit längs vägen. Tack för alla delade stunder och för att ni har gett mig en studietid jag aldrig kommer glömma. Jag ser fram emot att ge mig ut i livet med er.

Tack till Juridiska Föreningen, som har förgyllt alla mina år i Lund och gett mig vänner för livet.

Thank you to my childhood friends in Florida, who always have and always will encourage and inspire me every day.

Tack till min underbara familj, mamma, pappa och Sofie. Tack för all kärlek, för allt stöd och all pepp. För all korrläsning och alla kloka råd. Tack Sofie för att du för evigt är min bästa vän och stöttepelare.

Tack för allt, Lund.

Sara Axelsson

Lund, januari 2026

Abbreviations

AI	Artificial Intelligence
AI Act	Regulation (EU) 2024/1689 Artificial Intelligence Act
CDSM	Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market
CFR	Charter of Fundamental Rights of the European Union
CJEU	Court of Justice of the European Union
ECHR	Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights)
ECtHR	European Court of Human Rights
EDPB	European Data Protection Board
EU	European Union
EPRS	European Parliamentary Research Service
GDPR	Regulation (EU) 2016/679 General Data Protection Regulation
GenAI	Generative Artificial Intelligence
InfoSoc	Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society
NACE	Statistical Classification of Economic Activities in the European Community
TDM	Text and data mining
TEU	The Treaty on European Union
TFEU	The Treaty on the Functioning of the European Union

TRIPS	Agreement on Trade-Related Aspects of Intellectual Property Rights
WCT	WIPO Copyright Treaty (1996)
WIPO	World Intellectual Property Organisation

1 Introduction

1.1 Background

The rapid development of generative artificial intelligence (GenAI) has fundamentally transformed how creative content is produced, used, and consumed within the European Union (EU). Models such as DALL·E 3, Midjourney, and Stable Diffusion can now create state-of-the-art images, opening up new possibilities for artists and creatives.[1] While AI offers many possibilities, it also raises concerns for both creators, data subjects, and developers, particularly regarding copyright compliance and privacy rights. This thesis examines these concerns by analysing the extent to which generative AI developers can rely on the text and data mining (TDM) exceptions under EU copyright law, and whether these exceptions provide a workable balance between the interests of developers and rightsholders. It further explores how privacy and personality rights interact with this framework in the context of image-based AI training.

Unlike earlier forms of machine learning, which relied on relatively narrow and structured datasets, contemporary generative AI models require the largescale ingestion of images scraped from the internet in order to identify complex patterns and generate new visual output.[2] These datasets often include images protected by copyright and personal data relating to identifiable individuals. This technical reality raises complex legal questions, as GenAI training necessarily involves the reproduction, retention, and analysis of protected material. The scale of these processes has therefore renewed tensions between innovation, copyright protection, and fundamental rights.

At the same time, the EU has stressed the strategic importance of fostering AI development, also emphasising the freedom of information and the freedom to conduct a business.[3,4] An example of this is the introduction of the TDM exceptions in Articles 3 and 4 of the Directive on Copyright in the Digital Single Market[5] (CDSM Directive). These provisions were intended to

[1] See Guillaume Girard, 'DALL-E, Stable Diffusion, Adobe Firefly, Midjourney: Which AI Image Generator to Choose?' (16 August 2025) <https://edana.ch/en/2025/08/16/dall-e-stable-diffusion-adobe-firefly-midjourney-which-ai-image-generator-to-choose/> accessed 3 December 2025.

[2] Katherine Quezada-Tavarez, Lidia Dutkiewicz and Noémie Krack, 'Voicing challenges: GDPR and AI research' (2022) 2 *Open Research Europe* 126, 5.

[3] See Freedom to conduct a business in Article 16 Charter of Fundamental Rights of the European Union, 2012/C 326/02, 14 December 2007 (CFR) and Freedom of expression and information in Article 11 CFR.

[4] See European Commission, *AI Continent Action Plan* COM(2025) 165 final (9 April 2025).

[5] Directive (EU) 2019/790 of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

facilitate automated analysis of digital content and reduce barriers to research and innovation. However, these copyright exceptions, which GenAI now relies on, were drafted before the rise of modern GenAI systems. As a result, their actual application to GenAI training as well as interpretive difficulties remains debated among scholars and lacks judicial clarification.

Similarly, the General Data Protection Regulation (GDPR)[6] was adopted before the emergence of contemporary generative AI systems and was not designed with large-scale model training in mind. Where image datasets contain identifiable individuals, data protection law and, to a more limited extent, personality rights become relevant alongside copyright rules. Rather than forming an integrated regulatory framework with copyright law, the GDPR applies as a separate and independent regime. Imposing its own requirements, such as lawful processing, necessity, proportionality, and purpose limitation, that must be satisfied independently of copyright compliance. Personality rights, for their part, seek to protect the integrity and self-representation of the individual, but remain governed by fragmented national rules that play a limited role at the training stage. The interaction of these regimes therefore does not produce a single, coherent standard for AI training, but instead adds layers of context and constraint that may cumulatively restrict or permit specific practices. This layered interaction contributes to regulatory uncertainty for developers, rightsholders, and data subjects, and highlights the importance of analysing how these frameworks function together rather than in isolation.

Taken together, these overlapping frameworks show that the EU currently regulates AI training through rules that were developed independently and for different purposes. Copyright law aims to secure control over creative works, and data protection and personality rights law focuses on safeguarding personal information and personal integrity. When applied simultaneously to large-scale AI training, these regimes often operate in parallel rather than in coordination, leading to uncertainty about how they interact and where their boundaries lie. This fragmentation affects not only rightsholders and data subjects, but also developers who face inconsistent or impractical compliance expectations. As the use of generative AI continues to expand, the question of how these different areas of law can function together in a predictable way, while successfully balancing rights and interests, becomes increasingly important. The following chapters outline the legal framework relevant to this issue and identify the key points of tension that arise when images are used as training data.

[6] Regulation (EU) 2016/679 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.

1.2 Research Aim and Questions

The aim of this thesis is to examine and clarify the scope and practical functioning of the EU legal framework governing the use of image datasets for the training of generative AI models, with particular focus on copyright law, and its interaction with privacy and personality rights.

By analysing the application of the text and data mining exceptions in the CDSM Directive[7] alongside data protection and personality rights, the thesis seeks to assess whether the current framework offers a coherent, balanced, and workable regulatory structure for large-scale AI training.

Namely, the thesis will answer the following research questions:

1. Can the training of generative AI models on images fall within the text and data mining exceptions in Articles 3 and 4 of the CDSM Directive, and if so under what conditions?

2. How does generative AI developers' reliance on the TDM exceptions when training generative AI on image datasets interact with personality and privacy rights, including lawful processing requirements under the GDPR?

3. To what extent do the existing safeguards and balancing mechanisms in EU copyright, data protection and personality right law function in a coherent and workable manner in practice, when applied to the training of generative AI on image datasets?

1.3 Methodology

The thesis employs a doctrinal legal research method, also referred to as the legal dogmatic method. The aim of the doctrinal research method is to give a systematic exposition of the principles, rules and concepts governing a particular legal field, to then be able to analyse these relationships in order to solve the uncertainties and gaps in the existing law.[8] This aligns with the objective of the thesis, which is to examine the text and data mining exceptions on copyrighted materials in EU law, as well as personality and privacy regulations, to understand how these regulations apply to the use of images when training generative AI.

One of the main goals of the doctrinal legal research method is to describe present law by identifying, analysing and synthesizing the content. This is

[7] CDSM Directive (n5).

[8] Jan M. Smits, 'What Is Legal Doctrine? On The Aims and Methods of Legal-Dogmatic Research' in Rob van Gestel, Hans-W. Micklitz and Edward L. Rubin (eds), *Rethinking Legal Scholarship: A Transatlantic Dialogue* (Cambridge University Press 2017) 5.

done through an internal perspective, where the legal system is the subject, as well as it also provides the normative framework for analysis.[9] To achieve this goal, the doctrinal method relies on an important characteristic. Namely, that arguments are derived from authoritative sources, such as existing rules, principles, precedents, and scholarly publications.[10] Since the argument is based on internal, authoritative sources, the hierarchy of norms plays an important role.[11] In EU law, primary sources consist of the EU's constituent treaties,[12] the Charter of Fundamental Rights, and the general principles established by the Court of Justice of the European Union (CJEU). Next in the hierarchy are international agreements, which are not part of either primary or secondary law, but form their own unique category. Secondary law includes all legislative and non-legislative acts adopted by the EU institutions. This includes regulations, directives and decisions adopted by an ordinary or special legislative procedure.[13] Regulations are binding in their entirety and are directly applicable in member states, while directives are binding as to the result to be achieved. Decisions by the CJEU are also binding in their entirety upon those to whom they are addressed, while recommendations and opinions have no binding effect.[14]

Below secondary law we find preparatory works and scholarly writings. Preparatory works are not binding, but play an important role in understanding the legislators' meaning behind each provision, and writings contribute to a broader understanding and discussion of the subject.

Another essential feature of doctrinal research is to view the law as a coherent system. In order to describe the relevant regulations and the issues at hand, the applicable legal elements must be organised and fitted into a functional whole. By systematising the applicable law in this manner, the doctrinal approach is able to accommodate new technological developments by integrating new law into the existing framework.[15]

A systematic interpretation of the relevant legal framework is therefore employed, with the aim of applying the applicable rules to the training of generative AI while avoiding internal conflicts between different areas of law.[16] In accordance with established principles of EU law, this interpretation

[9] Terry Hutchinson, 'Doctrinal research' in Dawn Watkins (ed), *Research Methods in Law* (Routledge 2017) 13 and J M. Smits (2017) (n8).

[10] T Hutchingson (2017) (n9) 14.

[11] Lina Kestemont, 'Handbook on Legal Methodology: From Objective to Method' (Intersentia 2018) 24f.

[12] The Treaty on European Union (TEU) and the Treaty on the Functioning of the European Union (TFEU)) and its protocols.

[13] Publications Office of the European Union, 'EU hierarchy of norms' <https://eur-lex.europa.eu/EN/legal-content/glossary/eu-hierarchy-of-norms.html> accessed 29 September 2025.

[14] Article 288 TFEU (n12).

[15] J M. Smits (2017) (n8) 7.

[16] L Kestemont (2018) (n11) 24-26.

is conducted from an autonomous EU perspective, meaning that key legal concepts are interpreted independently relative to the legal systems of Member States and the international legal order, in order to ensure uniform application across Member States.[17]

While the interpretation of EU law can be compared to the interpretation of international treaties, there is a significant difference, as their interpretation is governed by this autonomous perspective. General principles of treaty interpretation, as reflected in the Vienna Convention on the Law of Treaties,[18] may serve as a point of reference, but they cannot override the autonomous interpretative methods developed by the CJEU, meaning that interpretations grounded in EU law take precedence over alternative interpretations derived from general international law.[19]

In addition, a teleological interpretation is applied, where the provisions are interpreted in light of their objectives and underlying purposes, specifically the EU legislator's aim to balance competing fundamental rights and interests and to ensure the effectiveness of EU copyright and data protection law in a technologically evolving context.[20]

Doctrinal legal research does not only enable a description of the existing legal framework however, but also provides a basis for identifying areas of uncertainty, assessing how specific rules function within the broader legal system, and reflecting on potential future developments. [21] This makes the method particularly suitable for the thesis, as it examines many of the practical difficulties of applying the current legal system to the training of GenAI.

1.4 Materials and Prior Research

This thesis relies primarily on the interpretation of EU primary and secondary law, with particular emphasis on directives such as the CDSM Directive[22] and the Information Society Directive (2001/29/EC) [23] (InfoSoc Directive), as

[17] See Niamh Nic Shuibhne, 'What is the autonomy of EU law, and why does that matter?' (2019) Nordic Journal of International Law, vol. 88, no. 1, 11-12. <https://doi.org/10.1163/15718107-08810002> accessed 20 December 2025.

[18] See Articles 31-33 United Nations. (1969). Vienna Convention on the Law of Treaties. Treaty Series, 1155, 331.

[19] See Pieter Jan Kuijper, 'The European Courts and the Law of Treaties: The Continuing Story', in Enzo Cannizzaro (ed.), *The Law of Treaties Beyond the Vienna Convention* (Oxford, 2011; online edn, Oxford Law Pro, 22 Sept. 2011), 265-270 <https://doi-org.ludwig.lub.lu.se/10.1093/acprof:oso/9780199588916.003.0016> accessed 20 Decmber 2026.

[20] L Kestemont (2018) (n16) 28.

[21] T Hutchinson (2017) (n9) 15.

[22] CDSM Directive (n5).

[23] Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

well as the GDPR[24]. It also draws, to a more limited extent, on international instruments such as the Berne Convention for the Protection of Literary and Artistic Works[25] and the TRIPS Agreement, [26] insofar as these form part of the broader copyright framework. Legislative preparatory works and scholarly writings are used to support interpretation, while case law from the CJEU, as well as selected national case law, is examined where relevant to the analysis.

The focus of this thesis lies in the EU legal framework governing these issues. Examples from Swedish law will however be used to illustrate national implementation and interpretation, as directives may be implemented differently between Member States. Swedish examples will be particularly used in relation to moral and personality rights, which remain outside the scope of EU harmonisation.

The core analysis of the thesis, namely the application of the EU copyright and personality and privacy rights framework to the training of GenAI systems, concerns a new and rapidly evolving area of law. As a result, the field is characterised by a limited body of case law and the absence of interpretative guidance from the CJEU. The analysis therefore relies to a significant extent on academic literature, reflecting the ongoing scholarly debate surrounding the legal treatment of generative AI training within existing EU legal frameworks.

Given the scarcity of case law in this area, the thesis examines a recent decision of the German lower courts, *Kneschke v LAION*,[27] in which the court interprets the TDM exceptions in the context of a non-profit entity engaged in text and data mining. The case is used throughout the thesis to illustrate how the regulatory framework may operate in practice, rather than as part of a comparative analysis. Since the judgment is examined outside its domestic legal context, the conclusions drawn should not be assumed to be fully transferable to all Member States. Moreover, as the decision was originally issued in German, the analysis is based on an unofficial English translation, and any potential interpretative nuances should therefore be approached with caution.

[24] GDPR (n6).

[25] Berne Convention for the Protection of Literary and Artistic Works, Sept. 9, 1886, as revised by the Paris Act of July 24, 1971, 1161 U.N.T.S. 3

[26] Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr. 15, 1994, Marrakesh Agreement Establishing the World Trade Organisation, Annex 1C, 1869 U.N.T.S. 299, 33 I.L.M. 1197 (1994).

[27] *Kneschke v LAION* Unofficial translation 2–3 https://chatgptiseatingtheworld.com/2024/09/28/unofficial-english-translation-of-german-courts-decision-kneschke-v-laion-under-tdm-exception/ accessed 22 October 2025.

1.5 Delimitations

The focus of this thesis is the protection of images, including photographs and artworks. This specific category of works has been chosen to explore the legal dilemmas that arise when multiple rights and interests intersect, particularly the intellectual property rights of the author, alongside the personality and privacy rights that may be implicated when works depict identifiable individuals or contain personal data.

Legal questions also emerge at various stages of the AI model lifecycle. Broadly speaking, these can be divided into the *training stage* and the *output stage*. The latter raises issues of potential copyright infringement where AI-generated outputs resemble existing protected works, particularly in relation to the author's rights of reproduction, communication to the public, or distribution. However, to maintain a clear focus, this thesis will limit its scope to the *input side* of the AI process, examining the legality of copying and storing large volumes of data for the purpose of training and fine-tuning generative AI models.

The thesis is also focused on questions regarding generative AI models and not simpler predictive models or machine learning, as generative AI is currently at the forefront of AI innovation and discussion, and brings further legal questions with it, compared to discriminative models.[28]

The exception in Article 3 CDSM Directive[29] that is discussed in the thesis, contains an exception for research organisations and cultural heritage institutions. Here the paper is focused mainly on research organisations and their definition and demarcations.

1.6 Outline

Chapter 2 provides an explanation of the technical aspects of TDM and training of GenAI that are relevant in relation to the EU copyright framework, as well as a background of the economic and moral rights granted authors under the copyright acquis. It explains the requirements for works to receive copyright protection and which rights are relevant and therefore require an exception for the training of GenAI. Chapter 3 presents the TDM exceptions in Article 3 and 4 in the CDSM Directive.[30] It explains and analyses certain interpretive challenges regarding the application of the exceptions to GenAI training, as well as the application of the three-step test. The chapter ends with an analysis of GenAI's possibility to rely on the TDM exceptions. Chapter 4

[28] Discriminative AI models are trained to recognise patterns in datasets and use those patterns to make predictions or classifications about new samples, whereas generative models have a different generative purpose and may therefore bring new legal challenges.

[29] CDSM Directive (n5).

[30] CDSM Directive (n5).

discusses personality and privacy rights and their application to GenAI training. Chapter 5 explores the interplay of these frameworks, and the functioning of the balancing mechanisms provided in them. Finally, a conclusion of the findings is presented in chapter 6.

2 GenAI Training and the EU Copyright Framework

The European copyright framework is primarily shaped by EU directives, interpreted by the Court of Justice of the European Union (CJEU), and informed by international copyright obligations. These instruments establish minimum standards that Member States must implement, while leaving certain matters, most notably moral rights, to national law. [31] An outline of the basic structure of EU copyright law is necessary in order to assess whether, and on what basis, an exception is required to permit the training of generative AI models, and to understand the limits within which such an exception must operate.

At the international level, all Member States are parties to the Berne Convention,[32] which sets core principles such as automatic protection, national treatment, and minimum standards. Although the EU is not itself a contracting party, EU copyright legislation is designed to comply with these obligations.[33] The Berne Convention,[34] together with the TRIPS Agreement[35] and the WIPO Copyright Treaty (WCT),[36] continues to play an important role in shaping both EU and national copyright law, including through mechanisms such as the three-step test, which remains relevant for assessing the scope and interpretation of copyright exceptions.

Within the EU, copyright harmonisation is chiefly achieved through directives. The InfoSoc Directive[37] harmonises the core economic rights, including reproduction, communication to the public, and distribution, together with a closed list of exceptions and limitations. Additionally, the Database Directive[38] may provide copyright protections to certain databases. More recently, the CDSM Directive[39] introduced specific exceptions for text and data mining, updating the framework to address automated uses of protected content in the digital environment, while preserving the protection of technological protection measures.

---

[31] Morten Rosenmeier, Kacper Szkalej and Sanna Wolk, *EU Copyright Law: Subsistence, Exploitation and Protection of Rights* (Wolters Kluwer 2019) 18.

[32] Berne Convention (n25).

[33] See M Rosenmeier and K Szkalej and S Wolk (2019) (n31) 3, Case C-300/98, *Dior v Evora BV* EU:C:1997:517 and Case C-5/08 *Infopaq Internationas v. Danske Dagblades Forening* EU:C:2009:465

[34] Berne Convention (n25).

[35] TRIPS Agreement (n26).

[36] WIPO Copyright Treaty, Dec. 20, 1996 S. Treaty Doc. No. 105-17 (1997); 2186 U.N.T.S. 121; 36 I.L.M. 65 (1997).

[37] InfoSoc Directive (n23).

[38] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

[39] CDSM Directive (n5).

Copyright protection under EU law is contingent upon originality. A work must constitute the author's own intellectual creation, a standard defined for photographs in Article 6 of the Term Directive[40] and extended by the CJEU to other categories of works.[41] In *Painer*, [42] the Court clarified that originality requires the author to have been able to make free and creative choices, which may be reflected in elements such as pose, lighting, atmosphere, or technique. The court also states that the extent of protection of works is the same for all types and does not depend on possible differences in the degree of creative freedom in the production.[43] Personal photographs, including those shared on social media, may therefore qualify for copyright protection, with the consequence that their reproduction for text and data mining or generative AI training engages the author's exclusive rights.

2.1 Technical Background: Reproductions and AI Training as Text and Data Mining

To determine when exclusive copyright rights may be engaged, and thus when reliance on the TDM exceptions becomes necessary, it is essential to outline the technical relationship between TDM and the training of generative AI. Advances in computational capacity and the availability of large-scale digital datasets have shaped both TDM techniques and contemporary AI training practices, with generative models relying on automated analysis of data to identify patterns and produce new outputs.[44] The following section therefore provides a brief overview of the TDM process and its relevance to GenAI training, focusing on whether, and at which stages, reproductions of protected works occur.

In Article 2(2) of the CDSM Directive,[45] text and data mining is defined as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations." It is further defined in Recital 8 of the CDSM Directive[46] as "the automated computational analysis of information in digital form, such as text, sounds, images or data", with the purpose of "gaining new knowledge and discovering new trends".[47] The act of TDM may

[40] Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights.

[41] See Case C-5/08 *Infopaq International v Danske Dagblades Forening* EU:C:2009:465.

[42] Case C-145/10 *Painer v VerlagsGmbH et al.* EU:C:2011:798.

[43] Case C-145/10 *Painer v VerlagsGmbH et al.* EU:C:2011:798. para 89–98.

[44] Eleonora Rosati, 'The exception for text and data mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Technical aspects' (EPRS Briefing, PE 604.942, European Parliamentary Research Service, February 2018) 2. https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/604942/IPOL_BRI(2018)604942_EN.pdf.

[45] CDSM Directive (n5).

[46] CDSM Directive (n5).

[47] CDSM Directive (n5).

conflict with intellectual property rights mainly during two stages, when accessing the data and when reproducing it in order to carry out the mining process.[48]

Some scholars, however, claim that TDM in its pure form does not copy at all, or that it does copy, but is carried out in relation to mere facts or data that are not protected by copyright, and therefore don't require any authorisation under copyright law.[49] A view that has led to the saying, "the right to read is the right to mine". TDM analyses or "mines" non-creative elements or works that cannot be copyrighted, these are in fact not always copied, and when they are, do not fall under any copyright protections and therefore do not need an exception to be reproduced. Nevertheless, there are often copies and reproductions that accompany or precede TDM methods. This "non-protectable" information, is normally embedded in copyrightable elements, which cannot be freely reproduced, causing the need for a copyright exception.[50]

TDM works by first identifying input materials to be analysed, which could entail either individually collected data or pre-existing databases. A necessary means for this first step is the access to content. Content can be distinguished into two different types, freely accessible, public domain content and content which requires some type of permission, such as a license.[51] All content used for TDM under the current exceptions require legal access to content, which can hinder certain mining and cause legal questions for developers. These issues will be explored further in section 3.1 below.

The second step of text and data mining is usually the extraction and/or copying of content. This includes pre-processing materials by turning them into a machine-readable format compatible with the technology of that specific TDM. Depending on the TDM technique, it can also entail uploading the pre-processed materials on a platform. This second reproduction step is normally where TDM encroaches upon rightsholders' copyright protections,

---

[48] Giorgos Vrakas. *A literature review of "lawful" text and data mining. Open research Europe*, 4, 153. (2024) 6-7. <https://doi.org/10.12688/openreseurope.18013.2> accessed 26 November 2025.

[49] See Recital 9 DSM Directive (n5); Kristina Christensen, 'A European Solution for Text and Data Mining in the Development of Creative Artificial Intelligence: With a Specific Focus on Articles 3 and 4 of the Digital Single Market Directive' (2021) 4(2) *Stockholm Intellectual Property Law Review* 18, 21.

[50] Tim W. Dornis, 'The training of generative AI is not text and data mining', *European Intellectual Property Review* (2025), 7, <https://ssrn.com/abstract=4993782> accessed 26 November 2026.

[51] E Rosati (2018) (n44) 4.

and where application of the exception becomes vital for legal TDM research.[52]

Finally, the third step of TDM is the actual mining activity, which entails extracting the data and recombining it to identify patterns and reach its goal of knowledge discovery.[53]

The development of modern artificial intelligence systems, particularly large-scale language models, depends fundamentally on the ability to extract and analyse vast quantities of data, and then self-train in order to create new content similar to existing content. This vast amount of data is often made available through TDM techniques.[54] The differences and similarities between the technical aspects behind TDM and GenAI training have been debated among scholars. Some scholars, such as Dornis and Stober, argue that the techniques cannot be regarded as the same, because the use of the underlying data differs: TDM is typically conceived as analysing semantic content (for example extracting patterns, correlations or knowledge from text/data) without necessarily reproducing or generating new full-form creative output, whereas GenAI training makes no distinction between semantic and syntactic information[55] and engages in ingesting *all* elements of the work, including style, expression, creativity and syntax, thereby enabling models to generate novel content.[56]

Most scholars, however, see TDM as a kind of first step in the process of training AI, where AI is simply a more sophisticated form of data analysis, and therefore relies on the same technical process.[57] TDM aggregates, transforms and extracts large-scale data, which is then fed into machine-

[52] Christopher Geiger, Giancarlo Frosio & Oleksandr Bulayenko, 'The exception for text and data mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects: In-Depth Analysis' (Policy Department for Citizens' Rights and Constitutional Affairs, Directorate-General for Internal Policies, European Parliament, PE 604.941, February 2018) 5.

[53] E Rosati (2018) (n44) 6–7.

[54] Elena Varese & Carolina Battistella, *Can generative AI rely on the text and data mining (TDM) exception for its training?*, DLA Piper, (2023), <https://www.dlapiper.com/en-us/insights/publications/law-in-tech/l-ai-generativa-puo-fare-affidamento-sulla-eccezione-text-and-data-mining-per-il-suo-addestramento> accessed 29 Oktober 2025.

[55] Semantics refers to information at the level of meaning, while syntax refers to information at the level of signs and symbols. The copyrightable syntax can be seen as the vessel for the works' semantic information. See T W. Dornis (2025) (n50) 4-6.

[56] T W. Dornis (2025) (n50) 21.

[57] See for example – E Rosati (2018) (n44); Martin Senftleben, 'Generative AI and Author Remuneration' (2023) IIC – *International Review of Intellectual Property and Competition Law* 1535, 54; Philipp Hacker, 'A Legal Framework for AI Training Data—from First Principles to the Artificial Intelligence Act' (2021) *Law, Innovation and Technology* 257; Eleonora Rosati, 'Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and Its Role in the Development of AI Creativity' (2019) SSRN Electronic Journal <https://ssrn.com/abstract=3452376> accessed 29 Oktober 2025; Matthias Leistner and Lucie Antoine, 'TDM and AI Training in the European Union – From "LAION" to Possible Ways Ahead?' (2025) *GRUR International,* volume 74, issue 11, 5.

learning models for training. From this perspective, TDM and GenAI training are overlapping and successive phases of a data-driven workflow. Alternatively, as mentioned above, some scholars argue that in a proper copyright framework there is no need for an exception at all, as TDM and AI training extracts data and factual information, which are not protected parts of works.[58] While there is ongoing debate regarding the technical similarities and differences between text and data mining and the training of generative AI systems, this paper aligns with the prevailing view that both processes involve reproductions of protected works. As a result, the author's exclusive right to reproduction is actualized and a copyright exception is required to enable such uses lawfully. The author's exclusive copyrights that may be actualized during the AI training process are presented below.

2.2 Economic Rights

At EU level, economic rights regarding digital photographs and other artistic or literary works are found mainly in the InfoSoc Directive.[59] These economic rights can be divided into two main exclusive rights: reproduction and communication to the public.

2.2.1 Reproduction

The right of reproduction is one of the most vital rights for rightsholders and the largest obstacle for GenAI developers when it comes to AI training and most TDM activities.[60] It is found in Article 2 of the InfoSoc Directive[61], where it states:

> Member States shall provide for the exclusive right to authorize or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part:
>
> (a) for authors, of their works;

In this case, "authors" also entails photographers and artists, and their works are protected as long as they meet the originality requirement.[62] The author's exclusive right to reproduction means that the permission of the author is needed in order to make copies of the work. This right is to be understood in a broad sense. The right doesn't only entail full copies, but also partial

[58] See Thomas Margoni and Martin Kretschmer, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2022) GRUR International, 685–701, 71 <https://doi.org/10.1093/grurint/ikac054> accessed 29 Oktober 2025.

[59] InfoSoc Directive (n23).

[60] See chapter 3 for further discussion.

[61] InfoSoc Directive (n23).

[62] Case C-145/10 *Painer v VerlagsGmbH et al.* EU:C:2011:798. para 95.

reproductions of works.[63] These copies can be digital or physical. Saving a protected work to a digital storage device is therefore also a form of copying.[64]

The broad interpretation of the reproduction right has had significant implications in the digital environment. The CJEU has confirmed that this interpretation also extends to technology-based copies, meaning that almost any technical reproduction falls within the scope of the right. This has brought challenges regarding reproductions in the digital environment. To solve this, legislators have had to focus instead on defining exceptions to this right.[65] One such exception is the exception for temporary reproductions. This exception was relied upon by TDM users prior to the CDSM Directive,[66] and can still be relevant for certain TDM methods.

2.2.1.1 Temporary Reproductions Exception

The exception for temporary reproductions is found in Article 5(1) of the InfoSoc Directive[67] and states that temporary acts of reproduction that are transient or incidental, and an integral and essential part of a technological process, whose sole purpose is to enable a transmission in a network, or a lawful use of a work are allowed. This is however only applicable if the acts of reproduction have no separate economic value. Specifically mentioned acts that should be included in the exception are acts that enable browsing and caching, including those which enable transmission systems to function efficiently. A use is considered lawful if it is authorized by the rightsholder or not restricted by law.[68]

The CJEU has clarified the scope of this exception in cases such as *Infopaq*[69] where it states the importance of the reproduction not exceeding what is necessary for the proper completion of the technological process. According to the court, it is also particularly important that the storage and deletion of the reproduction is not dependent on discretionary human intervention, as it cannot be guaranteed that the person concerned will actually delete the reproduction. This interpretation is supported by the specific acts mentioned in Recital 33 of the InfoSoc Directive,[70] since the acts are by definition, created and deleted automatically and without human intervention.[71]

---

[63] Tito Rendas, *Exceptions in EU copyright law: In Search of a Balance Between Flexibility and Legal Ceratinty,* (Wolters Kluwer 2021) 44.

[64] M Rosenmeier and K Szkalej and S Wolk (2019) (n31) 50.

[65] T Rendas (2021) (n63) 45.

[66] CDSM Directive (n5).

[67] InfoSoc Directive (n23).

[68] Recital 33 InfoSoc Directive (n23).

[69] *Infopaq International A/S v Danske Dagblades Forening* (C-5/08) EU:C:2009:465.

[70] InfoSoc Directive (n23).

[71] *Infopaq International A/S v Danske Dagblades Forening* (C-5/08) EU:C:2009:465 para 61-63.

Before the introduction of the specific TDM exceptions in the CDSM Directive[72], some scholars and stakeholders suggested that the temporary copies exception could, at least theoretically, cover certain automated data analysis processes. The argument was that the copies made during data mining were transient, incidental, and technically necessary for the process to function. However, this interpretation was always controversial.[73] The CJEU's narrow approach to the conditions of Article 5(1), especially the requirement that the copies have "no independent economic significance", meant that large-scale reproductions for analytical purposes were unlikely to fall under the exception.[74] As such, the legal uncertainty surrounding data mining prompted the EU legislator to introduce specific TDM exceptions in Articles 3 and 4 of the CDSM Directive,[75] addressing both research and commercial contexts.

An example of the interpretation regarding TDM and temporary reproductions can be found in a recent case regarding TDM from the lower German court, *Kneschke v. LAION*,[76] where the court quickly concluded that the TDM in question did not fall under the temporary reproduction exception. The case concerned a photographer who alleged that his copyright was infringed through the unauthorized reproduction of his image during the defendant's data analysis process. The defendant, a non-profit association, publicly provided image–text pair datasets free of charge. These datasets were used by both researchers and commercial actors to inter alia train generative AI models. The disputed image was originally published on the website of an image agency and was subsequently captured, downloaded, analysed, and incorporated into the dataset by the defendant.[77]

The defendant in the case claimed that the files were automatically deleted as part of the analysis process. The court, however, found that the deletion was due to deliberate programming by the user, and therefore, not carried out independently of the user. Consequently, the reproduction was not fleeting in the sense required by the exception. Furthermore, the reproduction was not incidental. The files in the case were intentionally downloaded in order to analyse them, and therefore not a reproduction with an independent purpose in relation to the technical process of which it is part.[78]

---

[72] CDSM Directive (n5).

[73] See for example Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko 'Opinion of the CEIPI on the European Commission's Proposal to Reform Copyright Limitations and Exceptions in the European Union' (2017) *SSRN Electronic Journal* 12 and Max Planck Institute for Innovation and Competition, *Position Statement … on the Proposed Modernisation of European Copyright Rules (Part B: Exceptions and Limitations (Art 3 – Text and Data Mining))* (14 January 2017) 1–2.

[74] C Geiger, G Frosio & O Bulayenko (2018) (n52) 7.

[75] CDSM Directive (n5).

[76] *Kneschke v LAION* (n27)

[77] *Kneschke v LAION* (n27)

[78] *Kneschke v LAION* (n27) 8.

Although the exception for temporary reproductions is not wide enough to cover all text and data mining, certain instances of TDM that do involve such temporary reproductions as mentioned in Article 5(1) InfoSoc Directive[79] shall, however, still fall under that exception.[80]

In the context of AI model training, the temporary reproductions exception now has limited relevance. The training of machine learning systems typically involves the creation of large-scale and lasting reproductions of protected works, which are stored and processed repeatedly over time. These activities exceed the transient and incidental nature envisaged by Article 5(1).[81] Consequently, the lawful use of copyrighted material for AI training must now be assessed under the CDSM Directive's[82] dedicated TDM provisions rather than under the temporary reproductions' exception.

The reproduction right is the only right included in the TDM exceptions, meaning that the other economic right, communication to the public, still has to be respected by AI developers. This right will be briefly presented in the subsections below for an understanding of the rights actualized throughout the AI process.

2.2.2 Communication to the Public

The exclusive right of communication to the public is found under Article 3 of the InfoSoc Directive,[83] which covers any act of making works available to the public by wire or wireless means in a manner that allows access from a place and at a time individually chosen by users. The decisive factor is whether members of the public are given the possibility to access the work, irrespective of whether such access is actually exercised.[84]

According to the CJEU, "the public" refers to an indeterminate number of potential recipients, excluding purely private or insignificant groups.[85] Furthermore, if works are made available online through linking or similar means, the assessment depends on whether the communication targets a *new public* not contemplated by the rightholder at the time of the initial communication.[86]

In the context of text and data mining and AI training, the communication right may be engaged where datasets compiled through automated processes are subsequently made available to third parties or shared openly online.

[79] InfoSoc Directive (n23).
[80] Recital 9 CDSM Directive (n5).
[81] InfoSoc Directive (n23).
[82] CDSM Directive (n5).
[83] InfoSoc Directive (n23).
[84] T Rendas (2021) (n63) 50; M Rosenmeier and K Szkalej and S Wolk (2019) (n31) 54.
[85] See Case C-117/15 *Reha Training v GEMA* EU:C:2016:379 para 41-43.
[86] See Case C-466/12 *Svensson et al. v Retriever* EU:C:2014:76 para 24.

When collecting and sharing data previously openly available online, the deciding question regarding copyright infringement will many times be whether the communication is done to a new public. This right falls outside the scope of the TDM exceptions and continues to apply independently of whether reproductions for mining purposes are permitted.

2.3 Protection of Databases

Databases play a central role in text and data mining and the training of generative AI systems by enabling structured access to large volumes of data. At the same time, their use may be subject to intellectual property protection under the Database Directive,[87] which developers must take into account. A database is defined as a collection of independent works, data, or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means. [88]

Databases may be protected by copyright where the selection or arrangement of their contents constitutes the author's own intellectual creation, reflecting free and creative choices.[89] Where this originality threshold is not met, protection may nevertheless arise through the *sui generis* database right, which grants the maker the exclusive right to prevent the extraction or re-utilisation of the whole or a substantial part of the database, provided that a qualitatively or quantitatively substantial investment has been made in obtaining, verifying, or presenting its contents.[90] The investment can be financial, material, or human.

The right to prevent extraction and re-utilisation should be interpreted broadly. *Extraction* can be compared to the right of reproduction, as it involves both the removal and copying of content, whereas *re-utilisation* refers to any act of making the contents available to the public. While insubstantial parts of a database may be extracted or re-utilised, such acts may not be carried out repeatedly or systematically.[91]

Importantly, these forms of protection may coexist. A database used for AI training may be protected either by copyright or by the sui generis right, while the individual works contained within it, such as images, may simultaneously enjoy independent copyright protection. As a result, AI developers relying on

[87] Database Directive (n38).

[88] See Article 1(2) Database Directive (n38).

[89] See Article 3(1) Database Directive; Irini Stamatoudi and Paul Torremans (eds), *EU Copyright Law* (Edward Elgar Publishing 2021); M Rosenmeier and K Szkalej and S Wolk (2019) (n31); Case C-604/10 *Football Dataco et al. v Yahoo! UK Ltd et al.* EU:C:2012:115 para. 38.

[90] Article 7(1) Database Directive (n38).

[91] Article 7(2) Database Directive (n38), see also M Rosenmeier and K Szkalej and S Wolk (2019) (n31) 66-68 and I Stamatoudi and P Torremans (eds) (n89) 235-237.

datasets for training purposes must consider not only the rights of the database maker but also the rights of the individual authors whose works are included.

2.4 Moral Rights

Moral rights of a work are required by Article 6*bis* of the Berne Convention,[92] which all Member States are signatories to. Moral rights are independent of economic rights and remain even after the transfer of such rights. The Convention[93] grants the author the right to claim ownership of the work (paternity rights) as well as the right to object to any distortion of other derogatory action in relation to the work which would be prejudicial to his honour.

Unlike economic rights, moral rights are not harmonised within the EU framework and are instead regulated solely by national provisions.[94] In Sweden, moral rights are regulated in the Swedish Copyright Act (1960:729) (*upphovsrättslagen*) and have been since it came into place in 1960.[95] The rights are rooted in the idea that a work is an expression of the author's personality. The provision gives the rightsholder a right of attribution when the work is reproduced or distributed, to the extent and in the manner required by good practice. A work shall also not be modified in such a way that it violates the author's literary or artistic reputation or compromise the integrity of the work, nor may it be made available to the public in any form or context that would be damaging to the author. The author may only waive these rights for uses of the work that are limited in nature and scope, such as a specific film adaptation.[96]

The rights cannot be transferred, but nevertheless do have an economic value, and any infringement can give rise to a right to compensation. Additionally, the exceptions and limitations to copyright set out in chapter 2 of the Swedish Copyright Act generally do not apply to moral rights.[97]

The right of attribution is actualized when a work is reproduced or made available to the public, regardless of if this is done with permission or not, as the author is seen to have the same interest of being accredited for the work if it is spread legally or illegally, or at a large or small scale. The provision therefore has a general scope, that can be asserted against anyone who exploits the work. Attribution does not have to take place if it is not seen to

[92] Berne Convention (n25).
[93] Berne Convention (n25).
[94] M Rosenmeier and K Szkalej and S Wolk (2019) (n31) 68.
[95] See Chapter 1 Section 3 Swedish copyright Act (1960:729) and Prop. 1960:17 med förslag till lag om upphovsrätt till litterära och konstnärliga verk.
[96] Prop. 1960:17 (n95), 35.
[97] SOU 1956:25 Upphovsmannarätt till litterära och konstnärliga verk, 113 and Prop. 1960:17 (n95), 64.

be in line with best practice, for example when it is difficult to give the author attribution due to technical aspects.[98]

While moral rights are traditionally associated with acts of publication or adaptation, they still hold relevance in the context of generative AI's use of training data. Although training processes do not normally involve attribution or public modification of individual works, large-scale automated uses raise questions as to whether and how moral interests are affected when protected works are ingested, transformed, or processed for model development. The absence of explicit moral-rights safeguards in the text and data mining exceptions further highlights the uncertainty surrounding their application in this context.

Moral rights also illustrate the close conceptual relationship between copyright and personality interests. While moral rights protect the author's personal and reputational bond to the work, personality rights safeguard the personal identity and integrity of individuals depicted in works, particularly in images. These latter rights are not governed by copyright law but arise from separate privacy and image-rights frameworks, which remain fragmented and only partially harmonised at EU level.[99]

When images depicting identifiable individuals are used as training data for generative AI, both sets of interests may be implicated simultaneously: the author's moral rights in the work and the depicted person's personality and privacy rights. This overlap creates potential tensions that are not directly addressed by the copyright framework and that may constrain reliance on the TDM exceptions. These conflicts will be explored further in the following chapters.

[98] SOU 1956:25 (n97), 116.

[99] See for example The Act on the Use of Names and Pictures in Advertising (Lag om namn och bild I reklam (1978:800)), the GDPR and Chapter 4, Section 6a of the Swedish Penal Code (Brottsbalken (1962:700)).

3 GenAI Training Under the CDSM Exceptions and the Three-Step Test

While the temporary reproductions' exception presented above may still be relevant in specific and limited TDM contexts, the TDM exceptions introduced by the CDSM Directive[100] have improved legal certainty for text and data mining activities. At the same time, questions remain as to whether the different purposes of traditional TDM, aimed at analysis and extraction of information, and generative AI training, aimed at developing models capable of producing new outputs, result in different legal treatment. Further uncertainty arises as to whether generative AI training can comply with the requirements of the TDM exceptions under the CDSM Directive,[101] particularly with regard to data retention and the practical effectiveness of opt-out mechanisms.

The following chapter examines the scope and interpretation of the TDM exceptions, identifies key interpretative challenges, and assesses their compliance with the three-step test, specifically in the context of their application to the training of GenAI systems.

3.1 Article 3 CDSM – Exception for Scientific Research

As mentioned previously, the use of TDM within the EU has been uncertain due to the lack of a stable legal framework. Therefore, clearer exceptions and limitations for text and data mining were introduced in the CDSM Directive[102] to limit the legal uncertainty within the Union, harmonizing the rules across borders and broadening the exception to facilitate research and innovation within the EU.[103] The two exceptions are found in Article 3 and Article 4 of the CDSM Directive.[104]

Exceptions for text and data mining for the purpose of scientific research are provided in Article 3 CDSM Directive.[105] The exception allows reproductions and extractions to be made during text and data mining by research organisations and cultural heritage institutions, and the persons attached thereto, for the purpose of scientific research, if they have lawful access of the works.

[100] CDSM Directive (n5).
[101] CDSM Directive (n5).
[102] CDSM Directive (n5).
[103] Recital 8 and 18 CDSM Directive (n5); C Geiger, G Frosio & O Bulayenko (2018) (n52) 24.
[104] CDSM Directive (n5).
[105] CDSM Directive (n5).

As the exception has a very limited nature and scope, the Directive[106] states that any potential harm to rightsholders will be limited, and Member States should therefore not provide compensation for the use of their works under the exception.[107]

Against this background, the following subchapters examine the key conditions and limitations of Article 3 CDSM Directive[108] in more detail, focusing on the requirements of lawful access, the definition of eligible beneficiaries and scientific research, and the boundary between research and commercial use.

3.1.1 Lawful access

The concept of *lawful access* refers to situations where researchers or institutions are legitimately entitled to use content through lawful means. This includes access obtained through open access initiatives, subscription agreements, or other contractual arrangements between rightsholders and research organisations or cultural heritage institutions. When, for example, a research organisation or a cultural heritage institution holds a subscription, the individuals affiliated with it and covered by that subscription are considered to have lawful access. The term also extends to materials that are freely accessible on the internet, meaning that open online availability qualifies as lawful access as well.[109]

Despite the clarification in the recitals, the exact delineation of lawful access is not clear. This has been mentioned as a weakness of the directive among scholars, who see a need for clarification.[110] Lawful access entails different aspects, such as the legitimacy of the source and how the user gains access. The directive does not clarify which of these aspects is decisive, and in practice, both may matter.

The most difficult aspect for users of TDM methods to comply with is the legitimacy of the source itself. Content that is publicly visible online, might have been uploaded without the rightsholder's permission, for example pirated films on YouTube or fan scans on Pinterest. In these cases, the user does not have lawful access to the material, even if they can publicly view the material. Therefore, to avoid the risk of copyright infringement, miners and

[106] CDSM Directive (n5).
[107] Recital 17 CDSM Directive (n5).
[108] CDSM Directive (n5).
[109] Recital 14 and 18 CDSM Directive (n5).
[110] See for example Thomas Margoni, 'TDM and Generative AI: Lawful Access and Opt-Outs' (draft v0.2, August 2024, forthcoming in Auteurs & Media 2024/2) <https://ssrn.com/abstract=5036164> accessed 20 November 2025; Kacper Szkalej, *The* 'Paradox of Lawful Text and Data Mining? Some Experiences from the Research Sector and Where We (Should) Go from Here', *GRUR International,* Volume 74, Issue 4 (2025) 312-313, <https://doi.org/10.1093/grurint/ikaf029> accessed 25 November 2025.

AI developers technically necessarily have to assess the lawfulness of every individual data point before use.[111]

To broaden the scope of the exception, and make its practical use more feasible for researchers, Margoni suggests an alternate interpretation of lawful use for research organisations under Article 3. Arguing that the requirement of lawful access should only cover the actions of the beneficiary, and not the status of the accessed source, differing between "lawful access" and "lawful source."[112] Should this differentiation not be made, it is likely that lawful access will be interpreted in light of cases from the CJEU, such as *ACI Adam*[113] and *GS Media*,[114] according to Margoni. Where the former entails a total ban on unlawful sources, and the latter a presumption of knowledge of unlawfulness of the source when operating for financial gain purposes.[115]

The second option would allow research organisations who do not pursue financial gain to mine any data as long as there is no direct knowledge of the unlawfulness of the source, allowing for a more lenient interpretation. This presumed lack of knowledge regarding the unlawfulness of sources would however cease once they become aware of such unlawfulness, which would result in an obligation to remove the unlawful data. While an image or a link can easily be removed from a website, removing that data point from a trained model may be practically impossible, or at least extremely costly, which raises issues with this seemingly more lenient interpretation as well.[116]

This problem of "de-learning" data presents a major practical challenge that affects the requirements of lawful access, the feasibility of an ex-post opt-out, and the right to be forgotten under the GDPR. The latter issues will be further discussed in the following chapters.

The lawfulness of the beneficiary's access is more apparent, where actions such as circumventing a paywall, a technological protection measure, or a valid contractual limitation are examples of unlawful access. For a more predictable framework, the deciding aspects and the interpretation of lawful access need further clarification.

---

[111] T Margoni (2025) (n110) 12–13; Jonathan Griffiths, Tatiana Synodinou and Raquel Xalabarder, 'Comment of the European Copyright Society Addressing Selected Aspects of the Implementation of Articles 3 to 7 of Directive (EU) 2019/790 on Copyright in the Digital Single Market' (2023) 72 *GRUR International* 26.

[112] T Margoni (2025) (n110) 12–13.

[113] Case C-435/12 *ACI Adam BV et al. v Stichting de Thuiskopie* EU:C:2014:254.

[114] Case C-160/15 *GS Media BV v Sanoma Media Netherlands BV et al.* EU:C:2016:644.

[115] T Margoni (2025) (n110) 13.

[116] T Margoni (2025) (n110) 13, see also Péter Mezei 'A saviour or a dead end? Reservation rights in the age of generative AI' (2024) *European Intellectual Property Review,* 2024, 46(7), 9 <https://ssrn.com/abstract=4695119> accessed 21 November 2025.

3.1.2 Beneficiaries and Requirements for "the purpose of scientific research"

The Article 3[117] exception is limited in scope as it restricts beneficiaries to research organisations and cultural institutions. Research organisations, according to the directive, should, in addition to universities or other higher education institutions, also include research institutes and hospitals that carry out research. These organisations can have different legal forms and structures, but will generally, within the Union, act either on a not-for-profit basis or in the context of a public-interest mission recognised by the respective Member State. Such a public-interest function may be demonstrated through public funding, statutory provisions, or obligations set out in public contracts.[118] The same definition of research organisation has been implemented in the Swedish law when transposing the Directive.[119] Swedish preparatory works maintain the same explanations of research organisations as the recitals of the CDSM Directive,[120] with some further clarifications. Namely that a delineated activity within a legal person can qualify as its own research organisation, and that a for-profit organisation that reinvests all profits in scientific research may also qualify.[121]

The purpose of the text and data mining exception is also limited to a targeted objective, namely, scientific research or educational activities that involve scientific research by these specific beneficiaries.[122] The term scientific research according to Recital 12 of the CDSM Directive[123] should "be understood to cover both the natural sciences and the human sciences."

Research organisations still benefit from the exception when research is carried out in the framework of public-private partnerships. This interpretation is in line with existing Union research policy, which encourages universities and research organisations to collaborate with the private sector. Research organisations can therefore continue to be the beneficiaries of the exceptions, while still relying on their private partners for carrying out text and data mining, including taking advantage of private technical tools.[124]

---

[117] CDSM Directive (n5).
[118] Recital 12 CDSM Directive (n5).
[119] See Chapter 2 Section 15 c Swedish Copyright Act (1960:729).
[120] CDSM Directive (n5).
[121] Ds 2021:30. Upphovsrätten på den digitala inre marknaden p. 60.
[122] See Kalpana Tyagi, 'Copyright, Text & Data Mining and the Innovation Dimension of Generative AI' (2024) 19 *Journal of Intellectual Property Law & Practice* 564 <https://doi.org/10.1093/jiplp/jpae028> accessed 21 November; João Pedro Quintais, 'What Is a "Research Organisation" and Why It Matters: From Text and Data Mining to AI Research' (2025) 74 *GRUR International* 397.
[123] CDSM Directive (n5).
[124] Recital 11 CDSM Directive (n5).

Such a partnership must however be conducted in a manner where the results of the scientific research cannot be accessed on preferential basis by the private party, and where that commercial entity does not hold a controlling position which allows them to influence operations or which leads to such preferential access.[125]

3.1.3 Research vs Commercial Boundary

The collaboration between research and commercial entities invites closer examination of where the boundary between these parties lies. This in turn also raises questions regarding the understanding of the concept of *commercial* within the EU. There is no definition of commercial as it is used in the CDSM Directive[126] or a general definition in EU primary law. Although there seems to be no one harmonised understanding of the commercial concept, the concept of what is seen as commercial within EU law is mentioned in different forms across separate EU frameworks.

In the GDPR,[127] a *commercial activity* is mentioned and defined as everything that is not a personal or household activity, such as personal online activity or social networking. In EU consumer law, we find a definition of *commercial practice.* Article 2(d) of the Directive concerning unfair business-to-consumer commercial practices[128] defines it as "any act, omission, course of conduct or representation, commercial communication including advertising and marketing, by a trader, directly connected with the promotion, sale or supply of a product to consumers."

Within other areas of EU law, a somewhat similar phrase, namely *economic activity* is used. In relation to the Statistical Classification of Economic Activities in the European Community (NACE), which is the system used within the EU to classify economic areas, such as agriculture or manufacturing, we find an explicit definition. According to this, an economic activity takes place when resources are combined to produce specific goods or services. It is characterised by an input of resources, a production process and an output.[129] By this definition, the scope is left very broad and it is not clear whether a profit-making purpose is needed.

---

[125] See Recital 11 CDSM Directive (n5); K Christensen (2021) (n49) 28.

[126] CDSM Directive (n5).

[127] GDPR Directive (n6).

[128] Directive 2005/29/EC of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council ('Unfair Commercial Practices Directive').

[129] European Commission Eurostat, 'Glossary: Statistical classification of economic activities in the European Community (NACE)' (Statistics Explained, n.d.)

Similarly, in the VAT Directive[130], an *economic activity* is defined as "any activity of producers, traders or persons supplying services," where the exploitation of tangible or intangible property for the purposes of obtaining income on a continuing basis is to be *particularly*[131] regarded as an economic activity.

From these examples it is clear that there is no harmonised understanding of what a commercial or economic activity entails. What the different legal frameworks have in common is that it is left as a broad definition where the core lies in resources being used to provide or produce goods or services. The definitions presented above are all from separate frameworks with different purposes, and it is hard to draw conclusions from these when interpreting the research-commercial boundary actualized in the TDM exception.

An interpretation aligned with the NACE definition, requiring merely an input of resources, processing, and an output, is too broad to determine what constitutes a commercial entity under the CDSM Directive,[132] as it would classify most research organisations engaging in TDM as commercial. Additional requirements, such as the presence of a "trader" or a "purpose of obtaining income," better reflect the aim of the exception, but cannot be applied mechanically.

Therefore, when determining the boundary between public-private partnerships and research-commercial, the ground for interpretation needs to lie mainly in the requirements for the beneficiary mentioned above, that are set forth in the CDSM Directive.[133] Where the assessment of the organisation as a commercial entity or not, as well as such a commercial entity's possibilities for influence and control must be determined on a case-to-case basis. An assessment of the requirements for the beneficiary of the research exception that exemplifies this was done in the recent *LAION* case.[134]

LAION, a non-profit who created datasets through TDM processes, then provided these openly online. These were mainly used by research organisations, but also by commercial corporations. Additionally, one of the co-founders of LAION as well as one "Research Scientist" were employed by commercial actors using the open data sets. The plaintiff therefore claimed that LAION could not rely on the exception in Article 3 CDSM Directive[135]

<https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Statistical_classification_of_economic_activities_in_the_European_Community_(NACE)> accessed 14 November 2025.

[130] Council Directive 2006/112/EC of 28 November 2006 on the common system of value added tax.

[131] Author's own emphasis.

[132] CDSM Directive (n5).

[133] CDSM Directive (n5).

[134] *Kneschke v LAION* (n27). For a short summary of the case see Section 2.2.1.1.

[135] CDSM Directive (n5).

due to commercial purpose and influence. The court, however, found that LAION could indeed invoke the exception.

Simply making the datasets freely available to the public made it according to the court "evident", that there was a non-commercial purpose with the research. The fact that other commercial actors used these datasets to develop their AI systems was deemed irrelevant for determining the nature of the defendant's activity, and the activity of the members of the defendant's organisations involved in these commercial actors, was not sufficient to attribute these companies' activities to the defendant as its own.[136] Furthermore, the commercial actors were not deemed to have the kind of decisive influence or control over the defendant for LAION to not be considered a research organisation according to the CDSM Directive.[137] The burden of proof to show this decisive influence was put on the plaintiff, who could not show that this was the case. Messages showing that the co-founder of LAION allegedly expressed his willingness to grant a commercial actor early access to a smaller dataset in exchange for funding, did not meet the requirements for decisive influence according to Article 3.[138]

This might suggest that part of the commercial aspect does lie within some sort of profit-making motive. The court is very clear that the commercial involvement of a few decisive persons within LAION and the use of the datasets by commercial actors is not enough to not grant LAION research status. Instead, a decisive factor seems to be that the datasets are published and available to anyone freely.

A model that begins as a research instrument, and later becomes commercial, might face further legal uncertainties. During the initial training and development of a model used solely for scientific purposes, the exception under Article 3 would be applicable, providing for a broader exception where the provider does not have to take rightsholders' reservations into account. What happens if this model is later put on the market, however, is not completely clear. AI systems and models that are specifically developed and put into service solely for scientific research receive further legal exceptions in the EU Artificial Intelligence Act (AI Act),[139] as they are exempt from the requirements of the act completely.[140] This exemption does not apply, however, for the fine-tuning modifications of a scientific model that is

[136] *Kneschke v LAION* (n27) 15.
[137] CDSM Directive (n5).
[138] *Kneschke v LAION* (n27) 16.
[139] Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence.
[140] See Article 2(6) AI Act (n139).

subsequently commercialized, or to models initially developed for scientific purposes that are later made available on the market.[141][142]

As one of the purposes of the CDSM Directive[143] was to harmonize the legal framework for cross-border uses of works in the digital environment across Member States[144], it suggests that the definition of a model developed for the purpose of scientific research should be understood in the same way according to both the AI Act[145] and the CDSM Directive.[146] With that assumption, an AI model that is initially developed for the purpose of scientific research and therefore relies on the exception under Article 3 CDSM Directive[147], can no longer rely on this exemption once the model is commercialized or made available on the market. This means that all fine-tuning processes or further development of the model must instead be in accordance with the general exception for TDM provided in Article 4 CDSM Directive[148] which also requires developers to comply with rightsholders' reservations and different retention limitations.

Article 3(2) CDSM Directive[149] allows for copies of works to be retained as long as they are stored with an appropriate level of security and only retained for the purpose of scientific research. Consequently, AI models initially developed for scientific research but later repurposed for commercial use may no longer rely on Article 3 to store reproductions, they must instead fall under the general exception in Article 4 which only allows retention of data for TDM purposes.

3.2 Article 4 CDSM – General Exception

During the negotiations of the CDSM Directive,[150] different proposals regarding the scope of the TDM exceptions were put forth. One of the first options suggested an exception only covering uses pursuing non-commercial scientific research purposes. Another option allowed uses for commercial purposes as well, but with limitations. The final option went further and suggested an exception for commercial purposes without restrictions. Ultimately, the second option allowing for commercial uses with exceptions

[141] "Making available on the market" is according to the AI Act defined as the supply of an AI system or GPAI model for distribution or use in the Union market in the course of a commercial activity, whether for payment or free of charge (Article 3(10) AI Act) (n139).

[142] See Recital 25 and 109 AI Act (n139); João Pedro Quintais, 'The Concept of "Research Organisation" and Its Implications for Text and Data Mining and AI Research' (2025) <https://doi.org/10.1093/grurint/ikaf030> accessed 12 November 2025.

[143] CDSM Directive (n5).

[144] See Recital 1-3 CDSM Directive (n5).

[145] AI Act (n139).

[146] CDSM Directive (n5).

[147] CDSM Directive (n5).

[148] CDSM Directive (n5).

[149] CDSM Directive (n5).

[150] CDSM Directive (n5).

was deemed the most proportionate one, and was added as an additional exception, expanding from the original scope of Article 3, and aiming to further balance the interest of the public, developers, and rightsholders.[151]

Article 4 CDSM Directive[152] provides an exception or limitation for reproductions and extractions of lawfully accessible works for the purpose of text and data mining. These may be retained for as long as necessary for the purpose of TDM and only apply on the condition that the use of the works has not been expressly reserved by the rightsholders in an appropriate manner, a reservation that has been known as the opt-out.

The different aspects of lawful access regarding research organisations under Article 3 as explained above are relevant for the requirement under Article 4 as well.[153] Concerning commercial actors however, another aspect, namely the rightsholder's intention is also relevant in the interpretation. Rightsholders may impose contractual or technical restrictions to prevent TDM or rely on the reservation of use according to Article 4, which may result in access for the purpose of TDM to no longer be lawful. These signals, such as metadata, robots.txt, or terms of service, are oftentimes inconsistent or missing, and the technical implementation varies significantly across platforms. As a result, the presence, absence, or malfunctioning of such signals can lead to highly context-dependent assessments and legal disputes.[154]

When training AI models, the question of legal retention of data is also actualized. Training GenAI normally requires a longer retention of data compared to traditional TDM methods. As mentioned above, according to the general exception for TDM in Article 4 CDSM Directive,[155] reproductions may only be retained for as long as necessary for text and data mining, which, in turn, raises the question of whether fine-tuning of AI models constitutes text and data mining, or a subsequent activity to it.[156]

3.2.1 Challenges of the Opt-Out Mechanism

The nature of the rightsholders' opt-out mechanism has faced a lot of criticism among scholars due to its interpretive and technical challenges, which results in a failure to secure the rights of rightsholders while also highly limiting the

[151] Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market 2016/0280 (COD) 8.
[152] CDSM Directive (n5).
[153] See Section 3.1.1.
[154] E Rosati (2019) (n57).
[155] CDSM Directive (n5).
[156] E Varese & C Battistella (2023) (n54).

scope of the exception and hampering AI innovation in the EU.[157] The mechanism raises several unresolved questions, including what constitutes sufficient machine-readability, how and when a reservation must be communicated, the appropriate place of the reservation, and the scope of the opt-out for rightsholders.

Some scholars are of the view that machine-readability is an example of how the reservation can be made, and not a strict requirement. For example, Italy has not implemented the requirement of machine-readability in their national legislation. Contrarily, however, countries such as Germany and Austria require machine-readability for the opt-out to be seen as effective.[158]

According to Recital 18 of the CDSM Directive,[159] a reservation regarding content that has been publicly made available online should "only be considered appropriate" when reserved by the use of machine-readable means, including metadata and terms and conditions of a website or service. This suggest that machine-readability should in fact be considered as a mandatory legal requirement to form a legally effective and "appropriate" reservation.[160] In other cases, when content is not publicly available online, it can be appropriate to reserve the rights by other means, such as contractual agreements or a unilateral declaration.[161]

With machine-readability established as a legal requirement for content publicly available online under the CDSM Directive,[162] the question remains how this notion should be interpreted, given that the Directive[163] provides no definition beyond the brief clarification in the recital. A broad meaning of the term can be based on the premise of TDM's automated character. Machine-readability should therefore entail an opt-out expressed in a form that is understandable to technical tools carrying out an automated analysis.[164] This does not exclude human readability of the reservation. Instead, the importance lies in if the reservation is "sufficiently binary,"[165] and if advanced technologies can reliably decode its content.[166] Despite this general

[157] See for example Stepanka Havlikova 'Technical Challenges of Rightsholders' Opt-out From Gen AI Training after Robert Kneschke v. LAION' (2025) JIPITEC 16(1), 92; P. Bernt Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)' (2019) *Kluwer Copyright Blog* <https://legalblogs.wolterskluwer.com/copyright-blog/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/> accessed 24 November 2025.

[158] S Havlikova (2025) (n157) 92.

[159] CDSM Directive (n5).

[160] This interpretation is shared by Havlikova, see S Havlikova (2025) (n157) 92.

[161] Recital 18 CDSM Directive (n5).

[162] CDSM Directive (n5).

[163] CDSM Directive (n5).

[164] See M Leistner and L Antoine (2025) (n57) 10; S Havlikova (2025) (n157) 92 and P Mezei (2024) (n116) 8.

[165] S Havlikova (2025) (n157) 95.

[166] P Mezei (2024) (n116) 8; S Havlikova (2025) (n157) 95.

understanding, it should be noted that there are still many debates regarding the conservative or liberal interpretation of machine-readability.[167]

The meaning of machine-readability was also considered in *obiter dictum* in *Kneschke v LAION.*[168] The German court states that the term must be interpreted in light of the legislative intent, namely, to enable automated queries by web crawlers. Therefore, "machine-readability" should be understood as "machine understandability". This means that reservations expressed in "natural language",[169] can also be considered "machine-understandable" if technically possible due to the technical developments prevailing at the relevant time of use of the work. The court then refers to Article 53(1) lit. c of the AI Act,[170] which states that providers of AI models must identify reservations according to Article 4(3) of the CDSM Directive[171] "even through the use of state-of-the-art technologies". Undoubtedly, according to the court, this would include applications capable of comprehending text written in natural language, setting a higher standard for the technology used by developers.[172]

The author believes this interpretation can be understood in line with Recital 18 of the CDSM Directive,[173] which exemplifies terms and conditions of a website as an appropriate place for a reservation. Meaning that a natural language reservation can meet the requirements of machine-readability, as long as it is sufficiently binary, as mentioned above. Providing for a proportionate balance between TDM users who have to take all necessary steps to identify reservations, and rightsholders who have to apply expressive and effective mechanisms to make use of their reservation. Issues still remain however, regarding the reliability of these opt-outs, both for rightsholders and developers.

Technical examples of systems that can meet the requirements for machine-readable reservations include Robots.txt files, dedicated TDM fields in the Hypertext Transfer Protocol (HTTP), metadata embedded in HTML content,

---

[167] See for example the US Motion Picture Association, who lobbied that "any reservation that a human could read is equally as machine-readable", Unofficial translation of MPA Legislative Amicus Brief of 31 Jan 2020, p. 2 <https://web.archive.org/web/20240314110407/https://www.bmj.de/SharedDocs/Downloads/DE/Gesetzgebung/Stellungnahmen/2020/013120_StellungnahmeAnpassungdes_Urheberrechts_MPA_DiskE.pdf?__blob=publicationFile&v=3> accessed 24 November 2025; see Hanjo Hamann, 'Artificial Intelligence and the Law of Machine-Readability: A Review of Human-to-Machine Communication Protocols and their (In)Compatibility with Article 4(3) of the Copyright DSM Directive' (2024) 15 *JIPITEC* 102, 110.

[168] *Kneschke v LAION* (n27).

[169] By *natural language* means human language, such as English or Mandarin. See Yue Zhang and Zhiyang Teng (2021) *Natural Language Processing: A Machine Learning Perspective* Cambridge University Press 3.

[170] AI Act (n139).

[171] CDSM Directive (n5).

[172] *Kneschke v LAION* (n27) 13.

[173] CDSM Directive (n5).

and various forms of watermarking in media files. A range of technologies can be used to implement opt-outs, but Robots.txt remains one of the most common. First introduced in 1994 as an anti-scraping tool, it has since been adapted and continues to be widely relied upon today.[174]

Although widely used, Robots.txt still poses several practical limitations when relied upon to signal reservations for TDM or GenAI development. Firstly, rightsholders must either block all automated access, thereby preventing search engines from crawling their webpages and potentially harming their visibility in search results, or maintain a "blacklist" or "whitelist" of specific crawlers. The latter approach requires rightsholders to identify individual scrapers, which is difficult in practice as scrapers choose their own user-agent names. Some major developers, such as OpenAI and Google, publish guidance on which user-agent identifiers can be used to block their crawlers. This transparency, however, creates an uneven playing field: transparent companies are more easily excluded from accessing data, whereas opaque actors that do not disclose their identifiers may go unnoticed and therefore bypass such reservations.[175] These difficulties exemplify the need for a more reliable and standardized opt-out mechanism. Second, the Robots.txt restriction is voluntary, as it does not technically block the automated access, but simply expresses rules for access.[176] Creating questions regarding how to make sure developers are compliant with reservations, especially due to the "black box" nature of training data.

The opt-out mechanism faces further questions in regard to the timing of the opt-out. The CDSM Directive[177] is not clear whether a reservation requires an ex-ante or ex-post, before or after mining, reservation. Although an ex-post reservation cannot be completely ruled out as compliant with the Directive,[178] it brings huge technical challenges, as any removal of mined data from the datasets of GenAI is seen as almost technologically impossible. Therefore, an ex-post reservation has a hard time complying with the ideas behind the reservation rights.[179] For a reservation to be practically usable, and most likely for it to legally comply with the directive, it therefore has to be done before mining begins. In practice, this puts a huge responsibility on rightsholders, who constantly have to monitor for new crawlers, and reserve their rights against each user-agent, before further training of a GenAI model starts in order to efficiently use their reservation rights.[180] To facilitate

[174] Paul Keller, *Considerations for Opt-Out Compliance Policies by AI Model Developers* (Open Future Policy Brief #6, 16 May 2024); H Hamann (2024) (n167) 113; S Havlikova (2025) (n157) 95; see also M Leistner and L Antoine (2025) (n57) 12.

[175] S Havlikova (2025) (n157) 100.

[176] S Havlikova (2025) (n157) 97.

[177] CDSM Directive (n5).

[178] CDSM Directive (n5).

[179] P Mezei (2024) (n116) 9; T Margoni (2024) (n110) 12-13; S Havlikova (2025) (n157) 95-101.

[180] See P Keller (2024) (n174) 7.

rightsholders' possibility of exercising their rights in this manner, the European Data Protection Board (EDPB) suggests that developers observe a reasonable period of time between the collection of a training dataset and its use. The additional safeguard may enable data subjects to exercise their rights during this period, although the practical challenges of having to check each model's individual dataset remains.[181]

Additionally, questions regarding the place of the opt-out can be discussed. Research by Open Future has differentiated between "location-based" and "unit-based" reservations. Location-based reservations are broader and placed on, for example, a website, and applies to all works hosted on that domain. A unit-based reservation, on the other hand, applies only to the individual work. Location-based reservations are suited mainly for rightsholders who administer their own site, while unit-based reservations may be used by rightsholders who have works on files spread across the internet.

When unit-based reservations on different sites are applied, this raises questions about what happens if contradictory statements are found. Do TDM users and GenAI developers have an obligation to remove data that was scraped legally, if it is found with a reservation on another site? This places a large burden on developers, with a fragmented framework that would most likely go beyond reasonable efforts and might unproportionately limit the freedom to conduct a business and harness innovation efforts.[182]

With the adoption of the AI Act,[183] the opt-out mechanism has been reinforced through its explicit reference in Article 53. Several major AI developers have already implemented, or intend to implement, their own opt-out solutions for rightsholders. However, this reintroduces a core problem: rightsholders must express their reservation in the specific technical format required by each individual AI system. This results in a highly fragmented and burdensome framework for rightsholders, rather than a coherent and uniform mechanism.[184]

On the other hand, many scholars warn that the opt-out in its current form, and especially in a more effective form, will have grave consequences for AI

[181] European Data Protection Board, *Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models* (17 December 2024) 29.

[182] S Havlikova (2025) (n157) 95; T Margoni (2024) (n110) 11 and 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) *Journal of Intellectual Property Law & Practice*, Volume 19, Issue 5, 456.

[183] AI Act (n139).

[184] Paul Keller, *'*Generative AI and Copyright: Convergence of Opt-Outs?*'* (2023) *Kluwer Copyright Blog* <https://copyrightblog.kluweriplaw.com /2023/11/23/generative-ai-and-copyright-convergence-of-opt-outs/> accessed 4 November 2025; G Vrakas (2024) (n48) ; P Mezei (2024) (n116) 11.

developers, giving rightsholders excessive power.[185] Massive opt-outs would deprive developers within the EU of training data, and may result in a lack of AI innovation within the Union, as well as a lack of data, causing models to become less effective or biased.[186]

To summarize, the opt-out mechanism is intended to balance the rights of rightsholders and developers relying on the TDM exceptions. In practice, however, it raises numerous practical and interpretive questions, and is today a time-consuming, expensive and fragmented framework for both rightsholders and developers. Greater technical development and clearer legal standardisation are needed to guarantee the opt-out mechanism as a true balancing tool.

3.3 The TDM Exceptions and Training of GenAI

As discussed above, the technical aspects of TDM and training of GenAI differ, but are on a broad level both types of data analysis that rely on the same technical process. Most scholars therefore agree, based on the broad definition of TDM in the CDSM Directive[187] as "any automated analytical technique aimed at analysing text and data in digital form…"[188], that from a technical standpoint, training of GenAI can be included as such an automated analytical technique and therefore deploy the exceptions. From the author's technical understanding of TDM and GenAI, this opinion is shared. However, questions regarding the legislators' original intent of the exceptions can still be raised. TDM shall, according to the Directive,[189] be carried out with the purpose to "generate information which includes but is not limited to patterns, trends and correlations"[190], or to gain new knowledge and discover trends.[191]

Scholars such as Dornis and Stober have questioned whether the process of GenAI, that can be explained as the probability distribution of training data, leading to the production of data similar to that training data, can fall under this purpose. While traditional TDM extract insights from data through, for example, pattern discovery, GenAI is focused on generation of content.[192] Although the generation of content does not seem to fit into Article 2's

[185] See for example P Hugenholtz (2019) (n157) and Maryna Manteghi, 'Overcoming Barriers to Text and Data Mining in the Era of ChatGPT: The Proposed Data Act as a Game-Changer' (2024) 73 *GRUR International* 34, 40.

[186] S Havlikova (2025) (n157) 106 and P Mezei (2024) (n116) 13.

[187] CDSM Directive (n5).

[188] Article 2(2) CDSM Directive (n5).

[189] CDSM Directive (n5).

[190] Article 2(2) CDSM Directive (n5).

[191] See Recital 8 CDSM Directive (n5).

[192] Tim W Dornis and Sebastian Stober, 'Generative AI Training and Copyright Law' (2025) *Transactions of the International Society for Music Information Retrieval* V(N) 4.

examples of patterns, trends and correlations, a wider interpretation of *new knowledge* may still allow for training of GenAI under the exception.

However, as copyright exceptions should be interpreted strictly,[193] it does not seem likely that the legislator had todays GenAI in mind when drafting the Directive,[194] thereby meaning for the technology to be included through a more extensive interpretation. Even though artificial intelligence has been a known phenomenon since as early as 1955, the creative generative capabilities that we know today were not known to a non-specialist public until 2022.[195] Making it unlikely that the lawmaker predicted this technology when the drafting started in 2016 or when the CDSM Directive[196] was decided in 2019.

Even though it might not have been the original intent of the lawmaker to include advanced GenAI in the scope of the exceptions, the implementation of the AI Act[197] seems to have clarified the standpoint of *today's* lawmaker. Article 53(1)(c) AI Act[198] requires providers of general-purpose AI models to comply with a reservation of rights expressed pursuant to Article 4(3) of the CDSM Directive.[199] This explicit reference and requirement to comply with the obligations of the Directive,[200] may imply that developers of AI can rely on the exception as well. An interpretation that is shared amongst most scholars, where many see Article 53(1)(c) of the AI Act[201] as a definite confirmation that GenAI is included in the scope of the TDM exceptions.[202] Although this interpretation is not shared by all, and Alonso and Luchi, for example, argue that this interpretation would invert the normative hierarchy, stating that "regulatory compliance obligations in the AI Act cannot reconfigure the substantive scope of exceptions established under copyright law."[203]

---

[193] Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* EU:C:2009:465, para 56.

[194] CDSM Directive (n5).

[195] George Siemens, 'Not everything we call AI is actually 'artificial intelligence' — here's what you need to know' *The Conversation* (26 December 2022) <https://theconversation.com/not-everything-we-call-ai-is-actually-artificial-intelligence-heres-what-you-need-to-know-196732> accessed 26 November 2025.

[196] CDSM Directive (n5).

[197] AI Act (n139).

[198] AI Act (n139).

[199] CDSM Directive (n5).

[200] CDSM Directive (n5).

[201] AI Act (n139).

[202] See for example: K Szkalej (2025) (n110) 311; T Margoni (2024) (n110); Michelle Rademeyer and Niloufer Selvadurai, 'Out from the Shadows: Developing Effective Copyright Laws for AI Training Datasets and Shadow Libraries' (2025) *Journal of Intellectual Property Law & Practice* 10.

[203] Eduardo Alonso and Nicola Lucchi, 'AI and copyright "hallucinations": does the text and data mining exception really support generative AI training?' (2025) 47 *European Intellectual Property Review* 515, 9.

Given the EU's strong incentive to foster the development of generative AI, the political and economic pressure to remain competitive in the global "AI race", and a broader trend towards technology-sensitive copyright legislation and interpretation, the author believes that it is likely that forthcoming judicial decisions and clarifying guidance will adopt a more extensive interpretation of the relevant exceptions, potentially confirming that the training of GenAI falls within the scope of the CDSM Directive.[204][205]

With GenAI as a part of the scope of the Directive,[206] questions still remain whether its current format is effective enough for the technical realities of GenAI. Issues such as the different retention times of data, the need for strict ex-ante opt-outs, and unclear requirements for lawful access of data, show a framework that is clearly not adapted for GenAI. While the training of GenAI models can, in principle, be interpreted as falling within the exceptions, the mismatch between technical practice and legal design means that the framework often operates ineffectively for both developers and rightsholders.

In sum, there are many differing opinions of GenAI's place or displacement within the TDM exceptions, and it is clear that clarification is needed to provide a stable and predictable framework for both rightsholders and developers reliant on copyright exceptions.

Beyond the question of whether the training of generative AI falls within the scope of the TDM exceptions, its inclusion may also affect the compatibility of those exceptions with the three-step test. It is therefore necessary to examine the TDM exceptions through the lens of the three-step test and assess whether all uses enabled by the exceptions satisfy its proportionality requirements. The following section presents an explanation of the three-step test and then turns to this assessment.

3.4 The Three-Step Test

The three-step test is a general requirement that states that permitted uses of copyright works and protected subject-matter are limited to certain special cases which do not conflict with a normal exploitation thereof and do not unreasonably prejudice the legitimate interest of the concerned rightsholders.[207] The two conditions apply cumulatively and in sequence, and

[204] CDSM Directive (n5).
[205] See M Leistner and L Antoine (2025) (n57) 10.
[206] CDSM Directive (n5).
[207] Eleonora Rosati, 'No Step-Free Copyright Exceptions: The Role of the Three-Step in Defining Permitted Uses of Protected Content (Including TDM for AI-Training Purposes)' (2023) SSRN <https://ssrn.com/abstract=4629528> accessed 5 November 2025.

must both be applicable in order for a certain exception or limitation to copyright or related rights to be valid.[208]

The test was originally established in the Berne Convention[209], but has since then been confirmed in various international agreements, such as the TRIPS Agreement[210] and the WIPO Copyright Treaty[211], as well as in the InfoSoc Directive at EU level.[212] A key question concerning the application of the three-step test has been its addressees, whether it is solely national legislatures that must adhere to it when incorporating relevant exceptions and limitations into domestic law, or whether it also binds EU and national courts.[213]

Although the text of the InfoSoc Directive[214] does not give an explicit answer, national and CJEU case law suggest that the test applies to the courts' interpretation of exceptions and limitations as well. This applies regardless of if the member state has directly transposed the language of the three-step test into law or not. For example, Sweden has not incorporated the text of three-step test directly into law, but the Supreme Court has applied the test directly to interpret the scope of certain exceptions and limitations in Swedish copyright law.[215]

In order for a certain exception or limitation to be compliant with the test, its three conditions need to be fulfilled. First, the exception needs to be limited to certain special cases. Meaning that the exception has to be clearly defined and limited in its scope and reach. The second condition requires the exception to not conflict with the normal exploitation of the work. For instance, a large circulation of counterfeits would inevitably reduce the volume of sales of the original work, and therefore conflict with the rightsholders' normal exploitation.[216] Finally, an exception cannot unreasonably prejudice the legitimate interest of the concerned rightsholders. The third step offers several filters that transform it into a refined proportionality test. The prejudice must be *unreasonable*, as well as the interests need to be *legitimate*, allowing the legislator and the courts to

[208] Claude Masouyé, Guide to the Berne Convention for the Protection of Literary and Artistic Works §9.6 (Paris Act, 1971) (WIPO:1978) ("1978 Guide").

[209] Article 9(2) Berne Convention.

[210] Article 13 TRIPS Agreement.

[211] Article 10 WIPO Copyright Treaty.

[212] Article 5(5) InfoSoc Directive (n23).

[213] E Rosati (2023) (n207) 6f; Eleonora Rosati, 'The Construction of Exceptions and Limitations in the InfoSoc Directive', (2019) *Copyright and the Court of Justice of the European Union* (Oxford, 2019; online edn, Oxford Academic) 148f <https://doi-org.ludwig.lub.lu.se/10.1093/oso/9780198837176.003.0006> accessed 5 November 2025.

[214] InfoSoc Directive (n23).

[215] See Högsta domstolen, *Bildupphovsrätt i Sverige v Wikimedia Sverige*, Ö 849-15, 4 April 2016.

[216] Case C-435/12 *ACI Adam BV et al. v Stichting de Thuiskopie* EU:C:2014:254, para 39.

consider the different interests at stake.[217] In this way, the three-step test is a flexible balancing tool, which serves a purpose restricting some national exceptions and limitations, as well as allowing for copyright protections to be balanced against social, cultural and economic interests.[218]

According to Hugenholtz and Okediji "limitations and exceptions that (1) are not overly broad, (2) do not rob right holders of a real or potential source of income that is substantive, and (3) do not do disproportional harm to the right holders, will pass the test".[219]

The three-step test is explicitly mentioned in Recital 6 of the CDSM Directive[220] and should be used as a tool to achieve a fair balance between the interests of rightsholders and users regarding all exceptions and limitations in the directive, including the TDM exceptions.

3.4.1 TDM Exceptions in Relation to the Three-Step Test

Having established the general function of the three-step test, the following section considers how the exceptions in Articles 3 and 4 of the DSM Directive[221] and their application to the training of GenAI perform under this framework.

The first step of the test, namely the exception's limit to certain special cases seems to be met by both Article 3 and Article 4 CDSM Directive.[222] "Certain special cases", requires exceptions to be clearly defined and narrow in scope.[223] Article 3's limit of beneficiaries and purpose to research organisations mining for the purpose of scientific research creates a narrow scope that should have no problem meeting the requirements of the first step. Article 4 does not restrict the beneficiaries or purpose of the TDM activities, initially leaving a very broad scope. Rightsholders possibility to reserve their rights however, as well as the link to Recital 18, which states that licensing opportunities should not be unduly affected by the exceptions and limitations, limit the scope and reflect the requirements of the three-step test.

---

[217] Christophe Geiger, Daniel J Gervais & Martin Senftleben, 'The Three-Step Test Revisited: How to Use the Test's Flexibility in National Copyright Law' (2014) *American University International Law Review*, 29(3), 605.

[218] Christophe Geiger, Daniel J Gervais & Martin Senftleben 'Chapter 5: Understanding the "three-step test"' (2015) *International Intellectual Property,* 184.

[219] P. Bernt Hugenholtz and Ruth Okediji, 'Conceiving an International Instrument on Limitations and Exceptions to Copyright' (2012) *Amsterdam Law School Legal Studies Research Paper* No. 2012-43, 6 March 2012, 3.

[220] CDSM Directive (n5).

[221] CDSM Directive (n5).

[222] CDSM Directive (n5).

[223] See Eleonora Rosati, 'Copyright reformed: The narrative of flexibility and its pitfalls in policy and legislative initiatives (2011-2021)' (2022) *Asia Pacific Law Review*, Forthcoming, 49-53, <http://dx.doi.org/10.2139/ssrn.4195559> accessed 5 November 2025.

Article 3 seems to be compliant with the second and third step of the test as well. The exception is confined to research purposes, and requires lawful access to works as well as appropriate and secure storage of works. This limited use of copies as well as appropriate safeguards prevent conflict with the normal exploitation of the work and ensures that any prejudice to rightsholders is proportionate.[224]

Under the second step of the three-step test, Article 4 CDSM[225] does not, in principle, conflict with the normal exploitation of works. According to CJEU case law, such a conflict arises where an exception enables uses that place the user in economic competition with licensed uses.[226] Article 4 mitigates this risk by allowing rightsholders to opt out and thereby retain control over commercially exploitable TDM uses. Resulting in the TDM process in accordance with the exceptions in the CDSM Directive[227] to be in line with the three-step test. Nonetheless, practical shortcomings, such as missing or ineffective reservations, may enable unlicensed TDM that could lead to de facto market substitution, particularly in the context of generative AI training.

Under the third step, courts must ensure that the application of Article 4 does not cause unreasonable prejudice to rightsholders' legitimate interests. In line with the CJEU's fair-balance jurisprudence, this requires a case-specific assessment to ensure that the application of Article 4, given the specific circumstances, respects the balance of rights.[228] This case specific assessment may cause issues for certain GenAI developers relying on the TDM exceptions. For example, it would not be possible to copy the whole collection of an artist's works, who might not have opted out simply due to lack of awareness or technical barriers, to then train a model that would create art in their style and essentially compete with or replace them.[229]

This would mean that the purpose of a model could become relevant when assessing its ability to invoke the TDM exceptions. While generative AI models can generally rely on the exceptions for their training processes, models that use rightsholders' works to produce outputs that *directly compete* with those works may fail to satisfy the three-step test, either by interfering with the normal exploitation of the work or by unreasonably prejudicing the legitimate interests of rightsholders. In practice, however, most models do not explicitly market themselves as tools for reproducing specific artistic styles,

[224] See Recitals 15–17 CDSM Directive (n5).

[225] CDSM Directive (n5).

[226] Case C-435/12 *ACI Adam BV et al. v Stichting de Thuiskopie* EU:C:2014:254, para 39.

[227] CDSM Directive (n5).

[228] See for example Case C-392/19 *VG Bild-Kunst v Stiftung Preußischer Kulturbesitz* EU:C:2021:181, para 54 and Opinion 3/15 of the Court (Grand Chamber), EU:C:2017:114, paras 123-125.

[229] E Rosati (2023) (n207) 20.

even though they are technically capable of doing so.[230] This makes it difficult to distinguish which systems create outputs that are "too similar" and thereby enter into direct economic competition. These uncertainties call into question whether the TDM exceptions, in their current form, effectively accommodate all types of generative AI and highlight the need for judicial clarification regarding GenAI's reliance on the exceptions and its compatibility with the three-step test.

The legitimacy of text and data mining under the three-step test was recently addressed in the previously mentioned case, *Kneschke v LAION*.[231] The court examined whether creating a data set that could later be used to train an AI model, that then generates images that can compete with the human made image used to train it, could count as an impairment of the normal exploitation of the work or mean that the legitimate interests of the rightsholder are unduly prejudiced. They found that this was not the case. The fact that the data set *may* be used in future technological developments, that cannot be fully foreseen, was not enough for a legally certain distinction between permissible and impermissible uses. Furthermore, the court found that based on current technical developments, it can never be completely ruled out that findings and data from TDM will be used to train generative AI models, which may then produce works that compete with current rightsholders. To rule out TDM in relation to the three-step test because of this possibility would practically prohibit the practice entirely and would contradict the legislative intent behind the TDM exceptions in the CDSM Directive.[232] The court does not, however, rule on whether the possible following reproductions during the training phase of AI models are in accordance with the TDM exceptions and the three-step test.[233] Meaning that uncertainties regarding the significance of the purpose of a GenAI model still remain.

This assessment under the three-step test does not determine the separate evaluation of individual outputs generated by a GenAI model. Whether a specific output is sufficiently similar to a protected work to infringe the rightsholder's reproduction or communication to the public rights must be assessed independently, on the basis of the output itself.

In sum, the practical challenges in meeting the specific conditions attached to the TDM exceptions, particularly the difficulties in verifying lawful access and complying with opt-outs, together with the lingering uncertainty as to whether large-scale AI training fully satisfies the requirements of the three-step test, indicate that the current legal framework is insufficiently equipped for GenAI development. These uncertainties create operational and legal risks

[230] For example, when asking ChatGPT to create an image in a similar style of Vincent Van Gogh's *The Starry Night*, an almost identical image is created.

[231] *Kneschke v LAION* (n27).

[232] CDSM Directive (n5).

[233] *Kneschke v LAION* (n27) 10–11.

for developers, who must navigate a regime that was not drafted with generative, large-scale, and commercially impactful models in mind. Consequently, further legislative or interpretative clarification, or the adoption of an AI-tailored framework, appears necessary to ensure legal certainty and to align copyright and data-processing rules with contemporary AI training practices.

4 Personality and Privacy Rights in AI Image Training

While recent debates on AI development have largely centred on its use of vast amounts of data and the resulting copyright implications, the processing of such data also raises important questions about personality and privacy rights. While not all data is personal data, these large datasets will often contain personal data that may implicate these rights.

Privacy rights protect individuals from intrusions into their private life and the misuse of personal information, whereas personality rights safeguard the individual's identity and personal attributes in the public sphere. While privacy focuses on non-interference and confidentiality, personality rights protect interests such as reputation, image, name, likeness and personal dignity. The two sets of rights overlap but are conceptually distinct: a violation of personality rights may occur without any invasion of privacy, and vice versa.

These rights, and their relevance for the training of generative AI systems, will be examined in this chapter.

4.1 Personality Rights

In European law, personality rights are not codified in a single instrument but instead emerge from fundamental rights protection of data and private life under Article 7 and 8 in the Charter of Fundamental Rights of the European Union (CFR)[234] and Article 8 European Convention on Human Rights (ECHR),[235] as interpreted by the CJEU and the European Court of Human Rights (ECtHR). Both courts understand "private life" broadly, covering public-facing attributes such as name, image, personal identity, reputation, and honour, elements that collectively function as personality rights at the European level.[236]

Swedish law similarly protects personality interests through a fragmented regulatory framework. Relevant protections include the prohibition on the use of a person's image in commercial advertising without consent,[237] as well as legal safeguards against defamation.[238]

---

[234] CFR (n3).

[235] Convention for the Protection of Human Rights and Fundamental Freedoms.

[236] See for example Case C-131/12 *Google Spain and Google Inc v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* EU:C:2014:317; *Von Hannover v Germany* (2005) 40 EHRR 1.

[237] See The Act (1978:800) on Names and Images in Advertising (Lag (1978:800) om namn och bild i reklam).

[238] See Chapter 5 The Swedish Criminal Code (Brottsbalk (1962:700)).

The protection of personal data or one's image through the personality-rights approach can be seen as a "personhood-oriented" model, differing from data-protection's "privacy-oriented" model. It typically encompasses a commercial aspect aimed at preventing unauthorised exploitation, as well as a moral aspect designed to protect individuals from degrading or humiliating uses.[239]

These characteristics are clearly reflected in Swedish law, which places particular emphasis on commercial exploitation and defamatory harm. In the context of generative AI, personality rights are therefore most likely to be engaged at the output stage. This may occur where AI systems generate images that are identical or highly similar to images of real persons used in training and those outputs are subsequently exploited for commercial purposes without consent. Personality rights may also be implicated where models memorise or hallucinate false or defamatory information about individuals that is presented in generated outputs.[240]

On the training side of GenAI, developers need to use large high-quality datasets without too many repeating data points to avoid these issues, as repeating data points can lead to higher risks of memorization.[241] A larger amount of data, however, can make it more difficult for developers to ensure that they also act in accordance with privacy rights.

Under Swedish law, strong protections are afforded to freedom of expression and information, as there are clear requirements of commercial or defamatory publication. This means that individuals generally cannot rely on personality rights to oppose the use of their personal images for training purposes solely on the basis of perceived exploitation of their personhood. Instead, protection against such uses is primarily provided through the privacy-oriented framework of the GDPR.[242]

However, the situation may differ in some Member States with stronger image rights, such as France or Spain. In these jurisdictions, the general right

---

[239] Bobur Koryogdiev, 'Comparative analysis: image rights under data protection vs personality rights' (2025) 4(9) *Current Approaches and New Research in Modern Sciences* 23 <https://doi.org/10.5281/zenodo.15623959> accessed 3 December 2025.

[240] See for example noyb, 'AI hallucinations: ChatGPT created a fake child murderer' (noyb, 20 March 2025) <https://noyb.eu/en/ai-hallucinations-chatgpt-created-fake-child-murderer> accessed 3 December 2025.

[241] See for example: Péter Mezei, 'Memorization and Generative AI – A Persistent Issue with Copyright Consequences?' in Enrico Bonadio, Péter Mezei and Eduardo Alonso (eds), *The Cambridge Handbook of Generative AI and IP in Europe* (CUP forthcoming 2026) <https://ssrn.com/abstract=5404367> accessed 3 December 2025; Bahadir Akdemir, 'How much do language models actually remember? A deep dive into AI memory and capacity' (Medium, 2024) <https://medium.com/@akdemir_bahadir/how-much-do-language-models-actually-remember-a-deep-dive-into-ai-memory-and-capacity-f7a71e55a93e> accessed 3 December 2025.

[242] GDPR (n6).

to one's image may, under certain circumstances, restrict the taking or use of an individual's image without consent. As a result, personality rights might technically provide a legal basis for objecting to the use of personal images for AI training, although the applicability of these rights to non-public, internal training processes remains contested. [243]

This divergence highlights the fragmented nature of personality and image rights protection within the EU. While such rights may overlap functionally, they remain legally fragmented, resulting in uneven protection for rightsholders and complex, jurisdiction-dependent compliance obligations for AI developers. Given the global circulation of data, these differences create legal uncertainty.[244] Enforcement is also challenging, as personality rights are typically asserted through costly civil proceedings, placing a significant burden on individuals.

In response, scholars and policymakers have called for greater harmonisation of image rights at the EU level, as well as for alternative regulatory approaches, including proposals to integrate the right to one's image into the copyright framework.[245] Image rights as copyright are often seen as a more effective and comprehensive protection, being able to leverage a familiar enforcement infrastructure. Framing image rights as copyright is often viewed as offering more effective and comprehensive protection, due to the availability of established enforcement mechanisms.[246]

Such an approach could protect both the economic value inherent in a person's identity through exclusive economic rights, such as reproduction and communication to the public, and personal integrity through moral rights. At the same time, personal data would remain protected under the GDPR,[247] resulting in a dual layer of protection for rightsholders. However, this model would also impose significant compliance burdens on AI developers, who would be required to clear rights not only from the creator of a work but also

[243] See Spain's "right to one's image" in Article 18 Spanish Constitution of 1978; and France's right of respect to private life in Article 9 France's Civil Code (*Code Civil (Dalloz, 2020)*; see also Tatiana Synodinou, 'Image Right and Copyright Law in Europe: Divergences and Convergences' (2014) 3 *Laws* 183-185 <https://ssrn.com/abstract=4637011> accessed 20 December 2025.

[244] Vincenzo Melilli, 'EU image rights: when the product is us' (World Trademark Review, 12 December 2020) <https://www.worldtrademarkreview.com/article/eu-image-rights-when-the-product-us> accessed 20 December 2025; Fatma Zehra and Eesha Kaul, 'Artificial Intelligence and the Law: Safeguarding Personality Rights in the Digital Age' (2024) 4(5) *Indian Journal of Integrated Research in Law* 989.

[245] See for example Denmark's amendment to its Copyright Act which grants individuals copyright over their likeness and their push for the EU to adopt similar measures, Kulturministeriet, 'Kulturministeren vil udbrede deepfake-lov til resten af Europa' (Kulturministeriet, 2024) <https://kum.dk/aktuelt/nyheder/kulturministeren-vil-udbrede-deepfake-lov-til-resten-af-europa> accessed 20 December 2025; see also V Melilli (2020) (n244).

[246] T Synodinou (2014) (n243).

[247] GDPR (n6).

from the person depicted, in addition to complying with data protection law, for a single image.

However, in the passing of the latest copyright harmonization efforts in the EU through the CDSM Directive,[248] image rights are not mentioned when discussing digital matters. Seemingly making a statement to connect image rights to human rights, and not copyright.[249] Nevertheless, the rapid development of generative AI and the increasing prevalence of deepfakes have reignited calls for EU-level intervention, making future legislative action possible, though uncertain.

Although the current personality rights framework does not clearly protect individuals against the use of personal images for AI training, particularly where national laws require some form of publication, it cannot be entirely excluded that personality rights may be engaged during the replication of images at the training stage. In *Reklos and Davourlis v Greece*,[250] the ECtHR held that the mere taking and retention of photographs of a newborn child without parental consent constituted a violation of Article 8 ECHR.[251] By analogy, it could be argued that the large-scale reproduction of personal images for AI training without consent interferes with individuals' private life and personality interests.

However, it must be noted that the photographer in this case had gained access to a sterile room supposedly only for medical professionals, where the patient should be able to expect a different level of privacy. On the other hand, images used to train GenAI are scraped from publicly available sources on the internet, where the data subject often has published the image willingly. On these fora the same level of privacy can't be expected, suggesting that a replication of such photos may not, on its own violate private life when weighed against rights to information and business.

Despite this, the author believes it is important to keep in mind that personal data is no longer just a question of personal integrity but an economic exploitation issue. In today's digital age and monetization of data, every person's image, and not only those of celebrities with a personal brand, carries economic value. Even when personal images are not published for commercial purposes, they are frequently collected and processed without consent in order to support commercial AI development. This evolving economic reality may strengthen future arguments for extending personality rights protection to the use of personal data in AI training contexts.

[248] CDMS Directive (n5).
[249] V Melilli (2020) (n244).
[250] *Reklos and Davourlis v Greece* App no 1234/05 (ECtHR, 15 January 2009).
[251] ECHR (n235).

In sum, personality and image rights in the EU provide an important but structurally limited layer of protection against the use of personal images in the context of generative AI. While these rights may effectively address harms arising at the output stage, such as unauthorised commercial exploitation or defamatory representations, their ability to constrain the training of AI models remains uncertain and highly dependent on national law. In many Member States, including Sweden, the requirement of publication or commercial use means that personality rights offer little opportunity to oppose the large-scale collection and processing of personal images for training purposes. Even in jurisdictions with stronger image rights, their applicability to non-public, internal training processes remains contested.

As a result, the protection of individuals whose personal images are used in the development of generative AI systems increasingly shifts away from personality-based protections towards data protection law. Unlike personality rights, the GDPR[252] applies directly to the collection, processing, and retention of personal data irrespective of publication or expressive use, making it the primary legal framework governing training-stage practices. The following chapter, therefore examines the role of privacy and data protection rights in regulating the training of generative AI models.

4.2 Privacy Rights

As demonstrated in the preceding section, personality and image rights offer only limited and uneven constraints on the use of personal images in the training of generative AI models. As a result, the assessment of training-stage practices largely falls within the scope of EU data protection law. Under EU law, the processing of personal data is in principle unlawful unless it can be justified under one of the grounds set out in the GDPR.[253] Consequently, the use of personal data for the purpose of training generative AI systems must be assessed through the lens of data protection law.

An image constitutes personal data when there is information concerning an identified or identifiable natural person, either directly or indirectly.[254] Photos of an identifiable person are always seen as personal data. Such photos can also include sensitive data according to Article 9(1) GDPR in the form of, for example, a person wearing glasses (revealing health data) or habits or food

[252] GDPR (n6).
[253] GDPR (n6).
[254] See Article 4(1) GDPR and Recital 26 GDPR (n6).

that may indicate religious views.[255] For sensitive data, special exceptions are required for lawful processing.

Processing of ordinary personal data must be based on one of the six legal bases set out in Article 6 GDPR,[256] where consent is generally seen as the most prominent legal basis. For training of GenAI however, consent is normally not a reliable legal basis, as it involves stringent requirements for valid consent that are generally not attainable due to the large amounts of data collected.[257] For example, relying on the derived consent from a data subject accepting the terms and conditions of a social media platform does not meet the threshold for specific and informed consent.[258]

The key legal basis for processing personal data in order to train GenAI instead falls under legitimate interest according to Article 6(1)(f) GDPR.[259]

The CJEU has identified three criteria for lawful processing under Article 6(1)(f) GDPR[260]: legitimate interest, necessity, and balancing of interests and rights.[261] The legitimate interest of the controller can entail both legal, economic or another intangible interests of the controller or a third party. Although data subjects enjoy strong protections of privacy, even commercial and economic interests will most likely be seen as legitimate, as they are protected by the freedom to conduct business.[262] Additionally, the development of GenAI provides IT tools and services which benefit the general public. The interest in these tools is also supported by the freedom of information.[263]

---

[255] Lee A Bygrave and Luca Tosoni, 'Article 4(1). Personal Data' in Christopher Kuner *et al* (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020), 109-110; Marvin Bartels, 'A Balancing Act: Data Protection Compliance of Artificial Intelligence' *GRUR International*, Volume 73, Issue 6, June 2024, 533, <https://doi.org/10.1093/grurint/ikae060> accessed 2 December 2025.

[256] GDPR (n6).

[257] See Claudio Novelli *et al*, 'Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity' (2024) *Computer Law & Security Review* 55, 106066, 5 <https://www.sciencedirect.com/science/article/pii/S0267364924001328> accessed 2 December 2025; European Parliament, *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence* (EPRS Study, STOA Scientific Foresight Unit, June 2020); Pablo Trigo Kramcsák, 'Can legitimate interest be an appropriate lawful basis for processing Artificial Intelligence training datasets?' (2023) 48 *Computer Law & Security Review* 105765, 5.

[258] See Article 4(11) GDPR (n6); K Quezada-Tavarez, L Dutkiewicz and N Krack (2022) (n2) 7.

[259] GDPR (n6).

[260] GDPR (n6).

[261] Case C-252/21 *Meta Platforms Inc et al v Bundeskartellamt* EU:C:2023:537, para 106.

[262] The freedom to conduct business is found in Article 16 CFR (n3) and Chapter 2 Article 17 The Instrument of Government (1974:152) (*Regeringsformen*) in Swedish law.

[263] See Article 11 CFR (n3); M Bartels (2024) (n255) 532; P Kramcsák (2023) (n257) 7–8.

For the processing to be seen as necessary, the interest pursued cannot be achieved as effectively by other means less intrusive for the data subject.[264] Which is in line with the data minimisation principle in Article 5(1)(c) GDPR. Minimizing the amount of data used is, however, a difficult task for developers of GenAI, as less data reduces the quality and accuracy of the model. Data-intensive processing should therefore be allowed as long as it is necessary for the purpose of the model.[265] The use of data that can be seen as irrelevant to the quality however, such as credit card numbers, may not meet the necessity requirements. Developers can also minimize personal data by using less data from social media, for example.[266]

The balancing test must be done on a case-by-case basis, and it is therefore hard to draw a general conclusion. Some argue, however, that there are strong arguments for the developer when it comes to training. The personal data used is often already published on the internet, and certain scholars argue that processing for AI does not bring any tangible risks for the data subject purely during the training phase.[267] This is a debated conclusion however, and models have been shown to reveal personal data through data leakage and model inversion, which does not support the idea of a completely risk free process for data subjects.[268]

When considering the reasonable expectations of the data subjects according to Recital 47 GDPR,[269] Bartels also argues that data subjects are aware that the data they publish on the internet can be read and used by others, and training on this data should therefore be allowed.[270] It can however be questioned if a data subject who posted images on social media ten years ago, reasonably expects these images to be used to train today's GenAI models. For processing to be lawful according to the balancing test, the final outcome must be that the legitimate interest of the controller or any third party outweighs the interests and fundamental rights of the data subject.[271] Although the processing of personal data involves privacy risks, these may prove sufficiently limited to permit the interests of developers and third parties, together with the broader benefits of GenAI, to outweigh them, thereby enabling lawful processing. Nevertheless, further guidance is needed

[264] *Meta Platforms Inc et al v Bundeskartellamt* EU:C:2023:537, para 108.

[265] See Paul Vogel, 'Training of AI Systems: How to Comply With the GDPR?' in *Privacy, Data Protection and Data-driven Technologies* (Routledge 2024) 166.

[266] M Bartels (2024) (n255) 532.

[267] See M Bartels (2024) (n255) 532

[268] Nicholas Carlini, et al. 'Extracting training data from diffusion models.' In: Presented at the 32nd USENIX Security Symposium (USENIX Security 23); 2023. <https://www.usenix.org/conference/usenixsecurity23/presentation/carlini> accessed 2 December 2025; C Novelli *et al.* (2024) (n257) 5; *Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models* (n181) 14.

[269] GDPR (n6).

[270] M Bartels (2024) (n255) 532.

[271] P Kramcsák (2023) (n257) 7.

on how to balance individual data protection rights with the public interest of GenAI development.

In order to process sensitive data according to Article 9(1) GDPR,[272] one of the exceptions in Article 9(2) needs to apply. The most promising of the exceptions for GenAI is Article 9(2)(g), which allows processing if necessary for reasons of substantial public interest. The processing shall be proportionate to the aim, and developers have to provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject. The exceptions may therefore be applicable in areas such as health, research, and education. According to this, it may be as unlikely, that GenAI models which process sensitive data, often by accident during web scraping, in order to create artistic images, are deemed to have such a substantial public interest that it is to be balanced over the protection of data subjects' sensitive data.

The technical difficulties concerning the deletion of personal data create a problem for GenAI developers regarding compliance with both the right to be forgotten in Article 17 GDPR[273] and the right to object in Article 21 GDPR.[274] According to Article 17, data subjects have a right to have their personal data erased, for example, if the processing no longer is necessary (Art. 17(1)(a)) or if it has been unlawfully processed (Art. 17(1)(d)). Article 21(1) further allows data subjects to reject processing of their data, when it has been lawfully based on legitimate interest, "on grounds relating to his or her particular situation" unless "the controller demonstrates compelling legitimate grounds for the processing". The controller's legitimate interest needs to "overwhelmingly" override the interests, rights and freedoms of the data subject.[275] The EDPB has clarified that this right needs to be ensured during the development of AI.[276]

If a data subject is deemed to have a right of erasure or objection, it will entail similar technical questions as an ex-ante opt-out for copyright holders. It is unclear how AI developers can comply with the right to erasure, as data is often not presented in a structured format and may be contained in multiple forms, making it difficult to isolate specific data points. "Deletion" of data often requires retraining of the model, which may substantially alter the dataset's overall quality and is not an economically sustainable solution for

[272] GDPR (n6).
[273] GDPR (n6).
[274] GDPR (n6).
[275] Gabriela Zanfir-Fortuna, 'Article 21 Right to Object' in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (OUP 2020) 517.
[276] See *Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models* (n181) 21.

developers. There is therefore a need for further guidance on what "deletion" of personal data in the training data of a GenAI model entails.[277]

The same technical difficulties of accessing specific datapoints also make it unclear if AI developers can comply with the data subjects' right of access to the personal data being processed according to Article 15 GDPR,[278] as well as the right to rectification of incorrect data according to Article 16 GDPR[279] and the principle of data accuracy according to Article 5(1)(d) GDPR.[280] However, although these principles are cornerstones of data protection laws, they are not absolute. Recital 39 calls for "reasonable steps" to ensure accurate and rectified data, opening up for a balance between other rights and freedoms. This may entail that the freedom of information[281] for third parties and the freedom to conduct business,[282] accompanied by the unreasonable steps required for developers to delete and access specific data points, open up for compliance despite inaccuracies, as long as they don't significantly impact individuals.[283]

Models developed for the purpose of scientific or historical research may according to Article 89 GDPR,[284] receive exceptions from the rights of access, rectification and objection, as long as appropriate safeguards such as pseudonymization are in place and core principles such as data minimisation are followed. The article does not provide a legal base for processing by itself, and developers still need to rely on a legal base according to Article 6 or Article 9 GDPR,[285] such as legitimate interest.[286] The right to erasure still needs to be respected, unless it is likely to render impossible or seriously impair the achievement of the objectives of that processing.[287] These exceptions ease some requirements for developers with a research purpose, but do not solve all issues, as pseudonymization and data minimisation will be difficult to comply with when using large datasets.

To conclude, processing personal data to train GenAI can most likely rely on legitimate interest as a ground for lawful processing. The balance test on which this ground rests on requires the different interests at stake to be

[277] See C Novelli *et al.* (2024) (n257) 7; M Bartels (2024) (n255) 529.
[278] GDPR (n6).
[279] GDPR (n6).
[280] GDPR (n6).
[281] Article 11 CFR (n3).
[282] Article 16 CFR. (n3).
[283] See C Novelli *et al.* (2024) (n257) 7.
[284] GDPR (n6).
[285] GDPR (n6).
[286] Article 6(1)(f) GDPR (n6).; Christian Wiese Svanberg, 'Article 89 Safeguards and Derogations Relating to Processing for Archiving Purposes in the Public Interest, Scientific or Historical Research Purposes or Statistical Purposes' in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (OUP 2020) 1245.
[287] Article 17(3)(d) GDPR (n6).

carefully weighed in each case, and special attention should be given to safeguards protecting personal data. The requirement for "compelling legitimate grounds" for processing when the data subject rejects processing[288] is however not as clear, and in general, further judicial guidance is needed regarding GenAI's technical capabilities to comply with all privacy requirements under the GDPR.

[288] Article 21(1) GDPR (n6).

5 Interplay and Systemic Tensions

The training of generative AI models on large collections of images engages multiple areas of EU law simultaneously. Copyright law, particularly through the TDM exceptions in Articles 3 and 4 of the CDSM Directive,[289] determines when reproductions for training are permissible. At the same time, the inclusion of identifiable individuals within image datasets implicates personality rights and activates the GDPR's[290] requirements for lawful processing of personal data. These regimes regulate different objects, namely artistic works, identity, and data. They were developed in response to different policy concerns and at a time when the technical realities of GenAI training were not yet foreseeable. As a result, neither set of rules fully accommodates how these systems operate, leaving several provisions difficult to apply in practice. Developers therefore face a fragmented and often misaligned legal landscape in which compliance with one regime does not ensure compliance with another, and in which personality-based interests are unevenly protected at different stages of the AI lifecycle.

This fragmented framework and different levels of protection can be shown through the personality rights and copyright framework, which both protect interest with an economic and a moral dimension. Copyright is grounded in the author's expended creative effort and in the integrity of the work, whereas personality rights protect the integrity and self-representation of the individual represented. Despite these parallels, copyright law offers clearer points of control at the training stage, as reproductions of protected works are directly regulated and subject to harmonised exceptions. Personality rights, by contrast, are not clearly applicable to internal training processes, particularly in legal systems that require some form of publication or commercial use, which results in weaker and less predictable protection at this stage.

Another illustrative example of compliance in a fragmented framework is the requirement of lawful access in the CDSM Directive[291] and lawful processing under the GDPR.[292] Two distinct requirements in separate frameworks, which may result in one framework restricting training, while the other allows it. A developer may have lawful access to training material through either a paid license or a lawfully uploaded (or maybe only lawfully accessed[293]) public source on the internet. However, if that material contains personal data, the developer must also identify a valid lawful basis for processing under Article

[289] CDSM Directive (n5).
[290] GDPR (n6).
[291] CDSM Directive (n5).
[292] GDPR (n6).
[293] See discussion about this question in Section 3.1.1.

6 GDPR.[294] For the training of GenAI systems, this will typically rely on legitimate interests.

The outcome of the legitimate interest balancing test may render the processing unlawful, if the interests or rights of data subjects outweigh those of the developer or third parties. This may occur even if, from a copyright perspective, the same use would be considered proportionate under the three-step test. In such cases, data protection law restricts training that would otherwise be permissible under the CDSM Directive,[295] demonstrating how the two frameworks operate cumulatively and how privacy rights can limit uses that copyright law would allow.

On the other hand, a developer may have a legitimate interest to process personal data under the GDPR, meaning that the processing would be permissible from a data protection perspective. However, if the data was obtained by scraping a website in breach of its terms of service or by circumventing a paywall, the access is unlawful. In such a situation, the TDM exceptions in the CDSM Directive[296] cannot be invoked. As a result, copyright law restricts training that would otherwise be allowed under data protection law.

Furthermore, if a balancing of interests between individual data subjects and developers was done simply relying on personality rights, the current framework would struggle to achieve a meaningful balance between personal integrity and the interest of developers, as individuals have limited opportunities to control the use of their images at the training stage. When read together with the GDPR,[297] however, a more complete balancing framework emerges. Although data protection law is formally concerned with personal data rather than personal identity, the safeguards it imposes in practice: lawfulness, necessity, proportionality, and purpose limitation, may indirectly protect identity-based interests as well. This functional overlap, however, also shifts the burden of protection onto data protection enforcement mechanisms, requiring proper enforcement of the safety mechanisms in the GDPR,[298] which, as shown above, is not fully technically feasible when training GenAI.

Developers must further ensure that they have a legitimate interest to process data and implement appropriate safeguards to protect both data subjects and copyright holders. In practice, this requires satisfying two distinct balancing exercises: the legitimate-interest test under the GDPR[299] and the

[294] GDPR (n6).
[295] CDSM Directive (n5).
[296] CDSM Directive (n5).
[297] GDPR (n6).
[298] GDPR (n6).
[299] GDPR (n6).

proportionality assessment embedded in the three-step test under copyright law. Although these tests originate from separate legal frameworks and pursue different objectives, they share notable similarities, particularly in how they weigh competing interests and in how the purpose of the AI model may influence the overall assessment.

In both assessments, rights and interests such as the developer's freedom to conduct a business[300] and society's and third parties' freedom of information,[301] are balanced against individuals' data protection rights[302] in the GDPR,[303] and rightsholders' right to their intellectual property[304] in the CDSM Directive[305]. In the three-step test, there is an additional requirement for the use to *unreasonably* prejudice the legitimate interest of the rightsholder. The use in this case needs to disproportionately harm the right of the copyright holder to be unlawful. This would entail that some harm is okay, and the interests at stake need to be weighed against each other.

The assessment in each case will partly rely on the purpose of the GenAI model at hand. The developer's interest in conducting a business will always be actualized. The general benefit of a model, however, will most likely be judged differently if it is a research model analysing health images or a model allowing users to create artistic image outputs. This might result in some types of models being able to rely on the TDM exceptions and fulfilling requirements for legitimate interests under GDPR,[306] while other types may not, despite relying on the same technological process.

The purpose of the model is also relevant from a personality-rights perspective. Although personality rights are not fully enforceable at the training stage in most current national frameworks, certain uses of personal images may nevertheless be perceived as more intrusive to personal identity than others. Models developed with the aim of generating sexualised or explicit content illustrate this point.[307] Even in the absence of defamatory output, the use of personal images to train such systems may interfere with personal integrity and dignity in ways that differ qualitatively from more abstract or analytical applications. The commercial use of one's personal images in this manner could be seen as violation of integrity and therefore personality rights. These emerging uses of generative AI raise questions as to

[300] Article 16 CFR (n3).
[301] Article 11 CFR (n3).
[302] See Respect for private life in Article 7 CFR (n3) and Article 8 (n235) and Protection of personal data in Article 8 CFR (n3).
[303] GDPR (n6).
[304] Article 17 CFR.
[305] CDSM Directive (n5).
[306] GDPR (n6).
[307] See for example Grok Imagine's "spicy mode"; Jibin Joseph, 'Grok App Adds AI Image and Video Generator With NSFW "Spicy" Mode' (*PCMag Australia*, 6 August 2025) <https://au.pcmag.com/ai/112455/grok-app-adds-ai-image-and-video-generator-with-nsfw-spicy-mode> accessed 20 December 2025.

whether existing national legislation sufficiently protects personal integrity in the AI context, and may prompt further harmonisation and legal development of image rights at EU level.

In the assessment of legitimate interest under the GDPR,[308] the different values of AI may give training of GenAI a general legitimate interest. However, this needs to be evaluated on a case-to-case basis, where safety measures such as data minimisation, pseudonymization and the fulfilment of other data subject rights such as right of access and right of rectification also play a role. The CDSM Directive[309] also provides a technical safeguard in the form of the opt-out mechanism. These safeguards should in theory support developers of GenAI in proportionality assessments as they limit exploitation and privacy risks for rightsholders and data subjects.

However, as discussed throughout the thesis, the technical characteristics of GenAI training make it difficult to comply with several of these safeguards. In both legal frameworks, the challenge of removing data once it has been incorporated into a model creates practical barriers. This affects the ability to meet data subjects' rights of access and erasure, and similarly complicates rightsholders' ability to exercise an effective ex-ante opt-out. These technical limitations raise doubts about how effective the balancing tests truly are in the context of GenAI training. They may function coherently in theory, yet prove difficult to apply meaningfully in practice.

Resulting in parallel frameworks where, when applied simultaneously, should be able to efficiently balance rights of developers, data subjects and rightsholders. However, they may fall short in practice due to the technical complexity of GenAI and the fact that neither framework was designed with these realities in mind.

The opt-out mechanism in the CDSM Directive[310] illustrates this tension. The current system creates a fragmented and burdensome process for both rightsholders and developers. [311] At the same time, an opt-out that operates effectively for rightsholders could disproportionately restrict the interests of developers and other users who depend on GenAI systems, as broad opt-outs may lead to insufficient training data and ultimately reduce model quality, also raising questions regarding the opt-out mechanism in itself as an appropriate balancing tool.

Overall, this chapter has shown that the EU legal framework governing the training of generative AI on image datasets relies on the cumulative application of copyright law, data protection, and personality-based interests,

[308] GDPR (n6).
[309] CDSM Directive (n5).
[310] CDSM Directive (n5).
[311] See Section 3.2.1.

resulting in a complex and often strained balance between distinct regulatory frameworks. Although the CDSM Directive[312] and the GDPR[313] each contain internal balancing mechanisms aimed at weighing developer interests against those of rightsholders and individuals, their parallel application produces fragmentation, legal uncertainty, and uneven protection at the training stage. Copyright law provides clearer points of control but is confined to protected works, while personality rights remain weakly enforceable and are largely channelled through data protection law. This functional reliance on the GDPR[314] to address identity-related concerns, despite its different normative focus, places significant regulatory weight on data protection and exposes structural and practical issues in how competing interests are balanced across the frameworks.

[312] CDSM Directive (n5).
[313] GDPR (n6).
[314] GDPR (n6).

6 Conclusion

This thesis has examined how the training of generative AI models on images is regulated under EU copyright law, data protection law, and personality rights. The analysis shows that the training of generative AI models can, in principle, fall within the scope of the text and data mining exceptions in Articles 3 and 4 of the CDSM Directive.[315] This conclusion is supported by the broad definition of text and data mining and reinforced by the AI Act's[316] explicit reference to Article 4 reservations. However, the effectiveness of these exceptions is constrained by unresolved interpretative issues concerning lawful access, data retention, and the practical operation of opt-out mechanisms. As a result, while the exceptions formally enable GenAI training, they provide limited legal certainty in practice.

With respect to privacy and personality rights, the thesis demonstrates that their relevance to generative AI training is uneven and largely mediated through data protection law. Personality rights play a limited role at the training stage, while the GDPR[317] introduces an additional set of conditions that operate alongside the CDSM framework[318]. Taken together, copyright, data protection, and personality-based rules form a layered regulatory structure in which the interests of developers, rightsholders, and data subjects are addressed through separate but cumulative legal assessments. Although these frameworks are capable of accommodating competing interests in principle, technical aspects of GenAI make full compliance with these balancing aspects technically challenging, and their parallel application remains difficult to coordinate in practice.

It is clear that further judicial clarifications are needed regarding certain interpretations, such as lawful access or what counts as "deletion" of data in terms of GenAI, as well as what legal importance the purpose of the model holds, in order to properly rely on TDM exceptions and grounds for processing personal data when training GenAI. Alternatively, a more AI-specific framework that explicitly balances rights with AI technology in mind could solve the issues at hand, but currently seems far away.

Ongoing technical developments are likely to improve AI systems' ability to comply with legal requirements. In particular, advances that reduce memorisation may also reduce the risk of copyright infringement and unlawful processing of personal data. However, the timing and effectiveness of such developments remain uncertain, and AI systems currently in use or

[315] CDSM Directive (n5).
[316] AI Act (n139).
[317] GDPR (n6).
[318] CDSM Directive (n5).

under development require judicial clarification to ensure compliance with existing copyright and data protection frameworks.

To truly reach a balance of rights, however, the author believes that the basis of the current framework might not be enough, and that future solutions will benefit in thinking outside of the box. Looking towards examples such as a large-scale licensing solution, which might make an effective opt-out mechanism into a working balancing tool,[319] or a general levy system, which would allow developers to pay a statutory fee to compensate copyright holders instead of having to rely on individual licenses.[320]

Overall, the complicated nature of this framework should not be an argument against its importance. Rather, it underscores the need for the EU to develop realistic and workable measures that protect both data subjects and rightsholders, continuing to set an example for the protection of individual rights, while still enabling AI innovation.

[319] See for example Martin Senftleben, 'Copyright Data Improvement for AI Licensing – The Role of Content Moderation and Text and Data Mining Rules' (2024) 7 <https://papers.ssrn.com/abstract=4817796> accessed 3 December 2025.

[320] See for example Grzegorz Tylec et al., 'Is It Possible to License Works Used in the Learning Process of Artificial Intelligence Algorithms?' (2024) 13 <https://papers.ssrn.com/abstract=4729495> accessed 3 December 2025.

Bibliography

Literature

Alonso E and Lucchi N, ‘AI and Copyright “Hallucinations”: Does the Text and Data Mining Exception Really Support Generative AI Training?’ (2025) 47 *European Intellectual Property Review* 515

Bartels M, ‘A Balancing Act: Data Protection Compliance of Artificial Intelligence’ (2024) *GRUR International,* Volume 73, Issue 6 <https://doi.org/10.1093/grurint/ikae060> accessed 2 December 2025

Bygrave LA and Tosoni L, ‘Article 4(1). Personal Data’ in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020)

Christensen K, ‘A European Solution for Text and Data Mining in the Development of Creative Artificial Intelligence: With a Specific Focus on Articles 3 and 4 of the Digital Single Market Directive’ (2021) 4(2) *Stockholm Intellectual Property Law Review* 18

Dornis T and Stober S, ‘Generative AI Training and Copyright Law’ (2025) *Transactions of the International Society for Music Information Retrieval* V(N) 4

Dornis T, ‘The training of generative AI is not text and data mining’, European Intellectual Property Review (2025), 7, <https://ssrn.com/abstract=4993782> accessed 26 November 2026

Geiger C, Frosio G and Bulayenko O, ‘Opinion of the CEIPI on the European Commission's Proposal to Reform Copyright Limitations and Exceptions in the European Union’ (2017) *SSRN Electronic Journal*

Geiger C, Gervais DJ and Senftleben M, ‘Chapter 5: Understanding the “three-step test”’ (2015) *International Intellectual Property*

Geiger C, Gervais DJ and Senftleben M, ‘The Three-Step Test Revisited: How to Use the Test’s Flexibility in National Copyright’ (2014) 29(3) *American University International Law Review*

Griffiths J, Synodinou T and Xalabarder R, ‘Comment of the European Copyright Society Addressing Selected Aspects of the Implementation of Articles 3 to 7 of Directive (EU) 2019/790 on Copyright in the Digital Single Market’ (2023) *GRUR International* 26

Hamann H, ‘Artificial Intelligence and the Law of Machine-Readability’ (2024) 15 *Journal of Intellectual Property, Information Technology and E-Commerce Law* 102

Hacker P, ‘A Legal Framework for AI Training Data—from First Principles to the Artificial Intelligence Act’ (2021) *Law, Innovation and Technology* 257

Havlikova S, ‘Technical Challenges of Rightsholders’ Opt-out From Gen AI Training after Robert Kneschke v. LAION’ (2025) 16 *Journal of Intellectual Property, Information Technology and E-Commerce Law*

Hugenholtz P.B and Okediji R, ‘Conceiving an International Instrument on Limitations and Exceptions to Copyright’ (2012) *Minnesota Legal Studies Research Paper* No 12-43

Hutchinson T, ‘Doctrinal Research’ in Dawn Watkins (ed), *Research Methods in Law* (Routledge 2017)

Keller P, *Considerations for Opt-Out Compliance Policies by AI Model Developers* (Open Future Policy Brief #6, 16 May 2024)

Kestemont L, *Handbook on Legal Methodology: From Objective to Method* (Intersentia 2018)

Koryogdiev B, ‘Comparative Analysis: Image Rights under Data Protection vs Personality Rights’ (2025) 4(9) *Current Approaches and New Research in Modern Sciences* <https://doi.org/10.5281/zenodo.15623959> accessed 3 December 2025

Kuijper PJ, ‘The European Courts and the Law of Treaties: The Continuing Story’ in Enzo Cannizzaro (ed), *The Law of Treaties Beyond the Vienna Convention* (Oxford University Press 2011)

Leistner M and Antoine L, ‘TDM and AI Training in the European Union – From “LAION” to Possible Ways Ahead?’ (2025) 74 *GRUR International* 11

Manteghi M, ‘Overcoming Barriers to Text and Data Mining in the Era of ChatGPT: The Proposed Data Act as a Game-’ (2024) 73 *GRUR International* 34

Margoni T and Kretschmer M, ‘A Deeper Look into the EU Text and Data Mining Exceptions’ (2022) 71 *GRUR International* 685, <https://doi.org/10.1093/grurint/ikac054> accessed 29 Oktober 2025

Margoni T, 'TDM and Generative AI: Lawful Access and Opt-Outs' (draft v0.2, August 2024, forthcoming in Auteurs & Media 2024/2) <https://ssrn.com/abstract=5036164> accessed 20 November 2025

Masouyé C, Guide to the Berne Convention for the Protection of Literary and Artistic Works (Paris Act, 1971) (WIPO:1978) ("1978 Guide")

Max Planck Institute for Innovation and Competition, *Position Statement ... on the Proposed Modernisation of European Copyright Rules (Part B: Exceptions and Limitations (Art 3 – Text and Data Mining))* (14 January 2017)

Mezei P, 'A saviour or a dead end? Reservation rights in the age of generative AI' (2024) European Intellectual Property Review, 2024, 46(7) <https://ssrn.com/abstract=4695119> accessed 21 November 2025

Mezei P, 'Memorization and Generative AI – A Persistent Issue with Copyright Consequences?' in Enrico Bonadio, Péter Mezei and Eduardo Alonso (eds), *The Cambridge Handbook of Generative AI and IP in Europe* (CUP forthcoming 2026) <https://ssrn.com/abstract=5404367> accessed 3 December 2025

MPA Legislative Amicus Brief of 31 Jan 2020 (Unoffical translation) <https://web.archive.org/web/20240314110407/https://www.bmj.de/SharedDocs/Downloads/DE/Gesetzgebung/Stellungnahmen/2020/013120_StellungnahmeAnpassungdes_Urheberrechts_MPA_DiskE.pdf?__blob=publicationFile&v=3> accessed 24 November 2025

Novelli C and others, 'Generative AI in EU Law: Liability, Privacy, Intellectual Property and Cybersecurity' (2024) *Computer Law & Security Review* 55 <https://www.sciencedirect.com/science/article/pii/S0267364924001328> accessed 2 December 2025

Quezada-Tavarez K, Dutkiewicz L and Krack N, 'Voicing Challenges: GDPR and AI Research' (2022) 2 *Open Research Europe* 126

Quintais JP, 'The Concept of "Research Organisation" and Its Implications for Text and Data Mining and AI Research' (2025) <https://doi.org/10.1093/grurint/ikaf030> accessed 12 November 2025

Quintais JP, 'What Is a "Research Organisation" and Why It Matters: From Text and Data Mining to AI Research' (2025) *GRUR International* 397

Rademeyer M and Selvadurai N, 'Out from the Shadows: Developing Effective Copyright Laws for AI Training Datasets' (2025) *Journal of Intellectual Property Law & Practice*

Rendas T, *Exceptions in EU Copyright Law: In Search of a Balance Between Flexibility and Legal Certainty* (Wolters Kluwer 2021)

Rosati E, 'Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and Its Role in the Development of AI Creativity' (2019) *SSRN Electronic Journal* <https://ssrn.com/abstract=3452376> accessed 29 Oktober 2025

Rosati E, 'Copyright reformed: The narrative of flexibility and its pitfalls in policy and legislative initiatives (2011-2021)' (2022) *Asia Pacific Law Review*, Forthcoming <http://dx.doi.org/10.2139/ssrn.4195559> accessed 5 November 2025

Rosati E, 'No Step-Free Copyright Exceptions: The Role of the Three-Step in Defining Permitted Uses of Protected Content (Including TDM for AI-Training Purposes)' (2023) SSRN <https://ssrn.com/abstract=4629528> accessed 5 November 2025

Rosati E, 'The Construction of Exceptions and Limitations in the InfoSoc Directive', (2019) Copyright and the Court of Justice of the European Union (Oxford, 2019; online edn, Oxford Academic) <https://dx.doi.org/10.2139/ssrn.4516462> accessed 5 November 2025

Rosenmeier M, Szkalej K and Wolk S, *EU Copyright Law: Subsistence, Exploitation and Protection of Rights* (Wolters Kluwer 2019)

Martin Senftleben, 'Copyright Data Improvement for AI Licensing – The Role of Content Moderation and Text and Data Mining Rules' (2024) <https://papers.ssrn.com/abstract=4817796> accessed 3 December 2025

Senftleben M, 'Generative AI and Author Remuneration' (2023) 54 *International Review of Intellectual Property and Competition Law* 1535

Shuibhne N, 'What Is the Autonomy of EU Law, and Why Does That Matter?' (2019) 88(1) *Nordic Journal of International Law*

Smits JM, 'What Is Legal Doctrine? On The Aims and Methods of Legal-Dogmatic Research' in Rob van Gestel, Hans-W. Micklitz and Edward L. Rubin (eds), Rethinking Legal Scholarship: A Transatlantic Dialogue (Cambridge University Press 2017)

Stamatoudi I and Torremans P (eds), *EU Copyright Law* (Edward Elgar Publishing 2021)

Tatiana Synodinou, ‘Image Right and Copyright Law in Europe: Divergences and Convergences’ (2014) 3 *Laws* <https://ssrn.com/abstract=4637011> accessed 20 December 2025

Szkalej K, The ‘Paradox of Lawful Text and Data Mining? Some Experiences from the Research Sector and Where We (Should) Go from Here’, *GRUR International,* Volume 74, Issue 4 (2025) <https://doi.org/10.1093/grurint/ikaf029> accessed 25 November 2025

Trigo Kramcsák P, ‘Can Legitimate Interest Be an Appropriate Lawful Basis for Processing Artificial Intelligence Training Datasets?’ (2023) 48 *Computer Law & Security Review* 105765

Tyagi K, ‘Copyright, Text & Data Mining and the Innovation Dimension of Generative AI’ (2024) *Journal of Intellectual Property Law & Practice* 564 <https://doi.org/10.1093/jiplp/jpae028> accessed 21 November 2025

Tylec G and others, ‘Is It Possible to License Works Used in the Learning Process of Artificial Intelligence Algorithms?’ (2024) <https://papers.ssrn.com/abstract=4729495> accessed 3 December 2025

Vogel P, ‘Training of AI Systems: How to Comply With the GDPR?’ in *Privacy, Data Protection and Data-driven Technologies* (Routledge 2024)

Vrakas G, ‘A Literature Review of “Lawful” Text and Data Mining’ (2024) 4 *Open Research Europe* 153 <https://doi.org/10.12688/openreseurope.18013.2> accessed 26 November 2025

Wiese Svanberg C, ‘Article 89 Safeguards and Derogations Relating to Processing for Archiving Purposes in the Public Interest, Scientific or Historical Research Purposes or Statistical Purposes’ in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (OUP 2020)

Zanfir-Fortuna G, ‘Article 21 Right to Object’ in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020)

Zehra F and Kaul E, ‘Artificial Intelligence and the Law: Safeguarding Personality Rights in the Digital Age’ (2024) 4(5) *Indian Journal of Integrated Research in Law*

Zhang Y and Teng Z, *Natural Language Processing: A Machine Learning Perspective* (Cambridge University Press 2021)

Ziaja G, ‘The Text and Data Mining Opt-Out in Article 4(3) CDSMD: Adequate veto right for rightholders or a suffocating blanket for European artificial intelligence innovations?’ (2024) 19 *Journal of Intellectual Property Law & Practice* 19(5)

EU Reports and Publications

European Commission, *AI Continent Action Plan* COM (2025) 165 final (9 April 2025)

European Commission (Eurostat), ‘Glossary: Statistical Classification of Economic Activities in the European Community (NACE)’ <https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Economic_activity> accessed 14 November 2025

European Data Protection Board, *Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models* (17 December 2024)

European Parliament Research Service, Rosati E, ‘The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Technical Aspects’ (EPRS Briefing PE 604.942, 2018)

European Parliament, ‘The impact of the General Data Protection Regulation (GDPR) on artificial intelligence’ (EPRS Study, STOA Scientific Foresight Unit, June 2020)

Policy Department for Citizens’ Rights and Constitutional Affairs, Geiger C, Frosio G & Bulayenko O, ‘The exception for text and data mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects: In-Depth Analysis’ (PE 604.941, 2018)

Publications Office of the European Union, 'EU hierarchy of norms' <https://eur-lex.europa.eu/EN/legal-content/glossary/eu-hierarchy-of-norms.html> accessed 29 September 2025

Websites

Akdemir B, ‘How much do language models actually remember? A deep dive into AI memory and capacity’ (Medium, 2024) <https://medium.com/@akdemir_bahadir/how-much-do-language-models-actually-remember-a-deep-dive-into-ai-memory-and-capacity-f7a71e55a93e> accessed 3 December 2025

Carlini N and others, ‘Extracting Training Data from Diffusion Models’ (USENIX Security Symposium 2023) <https://www.usenix.org/conference/usenixsecurity23/presentation/carlini> accessed 2 December 2025

Girard G, ‘DALL-E, Stable Diffusion, Adobe Firefly, Midjourney: Which AI Image Generator to Choose?’ (Edana, 16 August 2025) https://edana.ch/ accessed 3 December 2025

Hugenholtz P.B, ‘The New Copyright Directive: Text and Data Mining (Articles 3 and 4)’ (2019) *Kluwer Copyright Blog* <https://legalblogs.wolterskluwer.com/copyright-blog/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/> accessed 24 November 2025

Joseph J, ‘Grok App Adds AI Image and Video Generator with NSFW “Spicy” Mode’ (PCMag Australia, 6 August 2025) <https://au.pcmag.com/ai/112455/grok-app-adds-ai-image-and-video-generator-with-nsfw-spicy-mode> accessed 20 December 2025

Keller P, ‘Generative AI and Copyright: Convergence of Opt-Outs?’ (2023) *Kluwer Copyright Blog* <https://legalblogs.wolterskluwer.com/copyright-blog/generative-ai-and-copyright-convergence-of-opt-outs/> accessed 4 November 2025

Kulturministeren vil udbrede deepfake-lov til resten af Europa’ (Kulturministeriet, 2024) <https://kum.dk/aktuelt/nyheder/kulturministeren-vil-udbrede-deepfake-lov-til-resten-af-europa> accessed 20 December 2025

Melilli V, ‘EU Image Rights: When the Product Is Us’ (World Trademark Review, 12 December 2020) <https://www.worldtrademarkreview.com/article/eu-image-rights-when-the-product-us> accessed 20 December 2025

noyb, ‘AI Hallucinations: ChatGPT Created a Fake Child Murderer’ (20 March 2025) <https://noyb.eu/en/ai-hallucinations-chatgpt-created-fake-child-murderer> accessed 3 December 2025

Siemens G, ‘Not Everything We Call AI Is Actually “Artificial Intelligence”’ (The Conversation, 26 December 2022) <https://theconversation.com/not-everything-we-call-ai-is-actually-artificial-intelligence-heres-what-you-need-to-know-196732> accessed 26 November 2025

Varese E and Battistella C, ‘Can Generative AI Rely on the Text and Data Mining Exception for Training?’ (DLA Piper, 2023) <https://www.dlapiper.com/en-us/insights/publications/law-in-tech/l-ai-

generativa-puo-fare-affidamento-sulla-eccezione-text-and-data-mining-per-il-suo-addestramento> accessed 29 Oktober 2025

Table of Legislation

EU Legislation

Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society

Directive 2005/29/EC concerning unfair business-to-consumer commercial practices in the internal market

Directive 2006/112/EC of 28 November 2006 on the common system of value added tax

Directive 2006/115/EC on rental right and lending right and on certain rights related to copyright

Directive 2006/116/EC on the term of protection of copyright and certain related rights

Directive 96/9/EC on the legal protection of databases

Directive (EU) 2019/790 of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC

Regulation (EU) 2016/679 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

Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence. (Artificial Intelligence Act)

EU Preparatory Works

Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market 2016/0280 (COD)

Treaties

Agreement on Trade-Related Aspects of Intellectual Property Rights (1994)

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

Consolidated version of the Treaty on European Union (2012)

Consolidated version of the Treaty on the Functioning of the European Union (2012)

European Union Charter of Fundamental Rights (2012)

European Convention of Human Rights (1994)

The Vienna Convention on the Law of Treaties (1969)

World Intellectual Property Organisation Copyright Treaty (1996)

Swedish Legislation

Brottsbalken (1962:700)

Lag (1978:800) om namn och bild i reklam

Regeringsformen (1974:152)

Upphovsrättslag (1960:729)

Swedish Preparatory Works

Ds 2021:30. Upphovsrätten på den digitala inre marknaden

Prop. 1960:17 med förslag till lag om upphovsrätt till litterära och konstnärliga verk

SOU 1956:25. Upphovsmannarätt till litterära och konstnärliga verk

Other Legislation

Code Civil (Dalloz, 2020) (French Civil Code)

Spanish Constitution of 1978

Table of Cases

Court of Justice of the European Union

ACI Adam BV and Others v Stichting de Thuiskopie (C-435/12) EU:C:2014:254

Dior v Evora BV (C-300/98) EU:C:1997:517

Football Dataco Ltd and Others v Yahoo! UK Ltd and Others (C-604/10) EU:C:2012:115

Google Spain SL and Google Inc v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González (C-131/12) EU:C:2014:317

GS Media BV v Sanoma Media Netherlands BV and Others (C-160/15) EU:C:2016:644

Infopaq International A/S v Danske Dagblades Forening (C-5/08) EU:C:2009:465

Meta Platforms Inc and Others v Bundeskartellamt (C-252/21) EU:C:2023:537

Painer v Standard VerlagsGmbH and Others (C-145/10) EU:C:2011:798

Reha Training Gesellschaft für Sport- und Unfallrehabilitation mbH v GEMA (C-117/15) EU:C:2016:379

Svensson and Others v Retriever Sverige AB (C-466/12) EU:C:2014:76

VG Bild-Kunst v Stiftung Preußischer Kulturbesitz (C-392/19) EU:C:2021:181

Opinions of the court

Opinion 3/15 of the Court (Grand Chamber), EU:C:2017:114

European Court of Human Rights

Von Hannover v Germany (2005) 40 EHRR 1

Reklos and Davourlis v Greece App no 1234/05 (ECtHR, 15 January 2009)

National Courts

Germany

Kneschke v LAION eV (Landgericht Hamburg, 2024) (Unofficial translation)

Sweden

Bildupphovsrätt i Sverige v Wikimedia Sverige, Ö 849-15, 4 April 2016