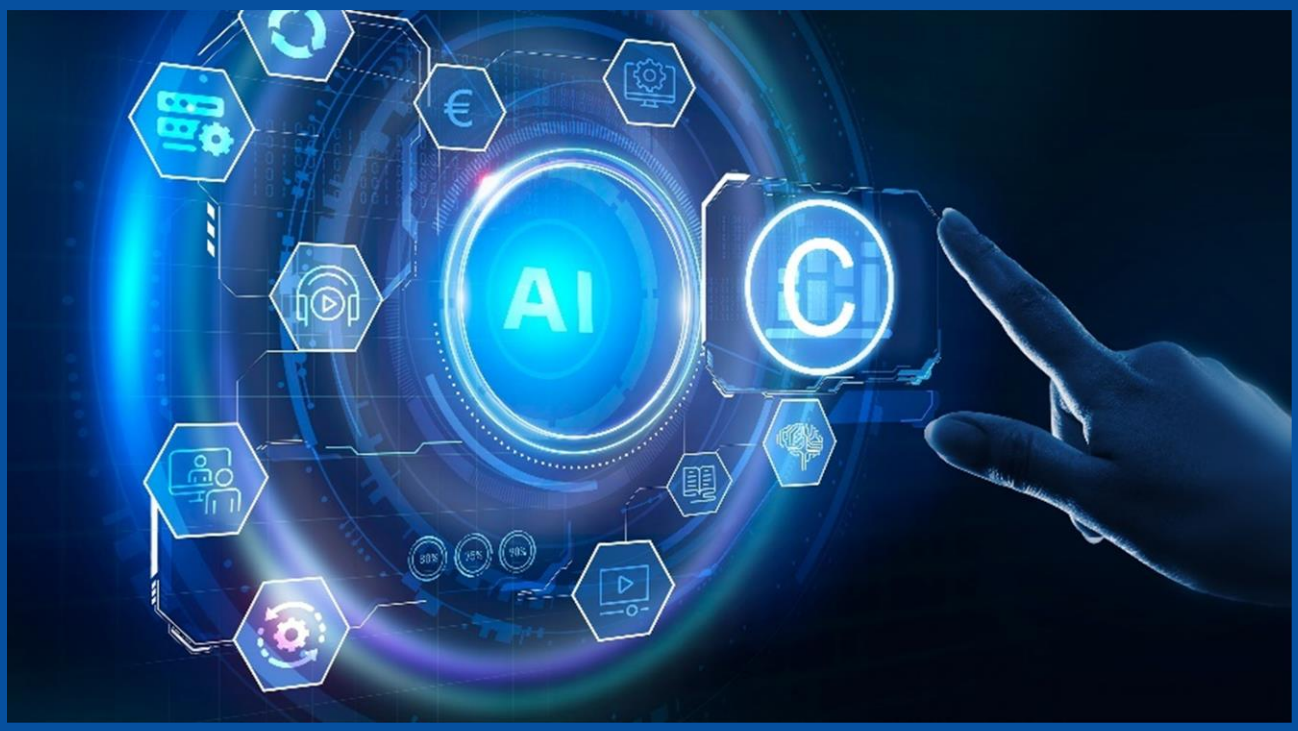
EUIPO
EUROPEAN UNION
INTELLECTUAL PROPERTY OFFICE

www.euipo.europa.eu

THE DEVELOPMENT OF GENERATIVE ARTIFICIAL INTELLIGENCE FROM A COPYRIGHT PERSPECTIVE



May 2025

THE DEVELOPMENT OF GENERATIVE ARTIFICIAL INTELLIGENCE FROM A COPYRIGHT PERSPECTIVE

TB-01-25-001-EN-N ISBN: 978-92-9156-369-2 DOI: 10.2814/3893780

Foreword

In an era marked by rapid technological transformation, copyright remains a cornerstone of Europe's cultural diversity and economic strength. The European Union's creative industries—firmly supported by a robust copyright framework – play a vital role in sustaining employment, fostering innovation, and preserving cultural heritage. Copyright-intensive sectors alone account for more than 17 million jobs and nearly 7% of the EU's GDP, underlining the central role of intellectual property in driving Europe's prosperity and global competitiveness.

Over the past three decades, successive waves of digital innovation have reshaped the way content is created, distributed and accessed. Throughout these transformations, copyright law has adapted to ensure that creators receive recognition and remuneration for their work, thereby sustaining the creative sectors that enrich our societies. However, the emergence of Generative Artificial Intelligence (GenAI) presents unprecedented challenges and opportunities, necessitating a re-evaluation of existing legal frameworks and support mechanisms to address the complexities introduced by this technology.

GenAI is already transforming the way we create, communicate, and innovate. While it offers immense potential as a source of growth and competitiveness in the future, it blurs the existing lines of content creation and introduces a new paradigm where not all content is created by humans. It therefore raises profound questions about how copyright can continue to serve its purpose while supporting innovation. It is essential to find a balance between these two objectives.

GenAI is often described as a black box, with little transparency around its input, functioning and outputs. This makes understanding its impact on copyright even more complex. This evolution prompts critical questions: How does GenAI use copyright-protected content? What is the European Union (EU) legal framework applicable to such use, and how can copyright holders reserve their rights and opt-out content from GenAI training? What are the developing technologies to mark or identify AI-generated content? And finally, what are the opportunities for copyright holders to license the use of their content by GenAI? All questions that need answers if we are to fully understand the development of GenAI from a copyright perspective.

This study is designed to clarify how GenAI systems interact with copyright – technically, legally, and economically. It examines how copyright-protected content is used in training models, what the applicable EU legal framework is, how creators can reserve their rights through opt-out mechanisms, and what technologies exist to mark or identify AI-generated outputs. It also explores licensing opportunities and the potential emergence of a functioning market for AI training data. Although the study is intended for experts in the field, it lays the groundwork for developing clear and accessible informational resources for a broader audience.

Furthermore, this report will provide insights for policymakers to maximise the innovative potential of the EU in light of these new technologies. As the Draghi report on the future of EU competitiveness recently underlined, and as highlighted in the European Commission AI Continent Action Plan, Europe must lead in the digital and AI transformation, not only by investing in infrastructure and skills, but also by shaping the regulatory frameworks that govern emerging technologies. Copyright is a key component of such a framework. It is central to maintaining Europe's capacity to innovate on its own terms—grounded in values of fairness, transparency, and respect for intellectual property.

The EUIPO Strategic Plan 2030 reinforces this vision. It calls on the office to support the strengthening of the IP ecosystem in line with technological developments, such as the rise of GenAI, demonstrating the need for action and new solutions to support both innovation and copyright protection. This study represents an early and important step in meeting that strategic commitment. But it is also a starting point. Much more is needed to guide and support rights holders, AI developers, and policymakers through this fast-changing environment, if we are to realise the full potential of EU digital markets for creators and businesses.

To that end, the EUIPO will launch the Copyright Knowledge Centre by the end of 2025. With regard to GenAI, this new Centre will equip copyright holders with clear, practical information on how their works may be used in the development of GenAI – and how they can effectively manage and protect their intellectual assets. It will also provide a platform for stakeholders, enabling creators, developers, and institutions to share needs, identify gaps, and explore opportunities for collaboration. Drawing on the insights of this study, the Centre will provide a foundation for discussions among experts on how copyright can effectively support content creation and innovation in the GenAI landscape.

It is essential to make copyright rules work in a way that keep human creators in control and ensure their proper remuneration, while allowing AI developers of all sizes to have competitive access to high-quality data. Balancing both interests can be facilitated by simple and effective mechanisms for copyright holders to reserve their rights and the use of their content, as well as licensing and mediation mechanisms to facilitate the conclusion of license agreements with AI developers. As GenAI applications and markets mature, further reflections might also be needed on whether content generated by AI deserves protection through existing or new intellectual property rights.

At the EUIPO, we stand ready to play our part. By working in close cooperation with European and international institutions to contribute our expertise on IP protection and awareness, and in the development of technical solutions and mediation services to help ensure that, as with earlier digital innovation cycles, copyright keep supporting creators and technological progress.

Acknowledgements

This study has been prepared by a research team of the University of Turin Law School and the Nexa Center for Internet & Society of the Polytechnic of Turin for the European Union Intellectual Property Office (EUIPO).

A list of researchers and collaborators who contributed to this project is included as *Annex I*.

Table of Contents

Executive Summary

Over the past several years Artificial Intelligence (AI) technologies have experienced major advances, with the release of Large Language Models (LLMs) and Generative AI (GenAI) systems. GenAI services to generate text, code, image, video, and audio content are now widely available. This has led policymakers and regulators to examine how existing legal frameworks should evolve to address the implications of large-scale AI adoption, and to balance innovation with intellectual property (IP) protection.

In this context, this study explores the developments of GenAI from the perspective of EU copyright law. It is structured around three main components, (1) **a technical, legal and economic analysis** to further understand the functionality of GenAI and the implications of its development, as well as a detailed examination of copyright-related issues regarding the (2) **use of content in GenAI services development** and the (3) **generation of content**.

Technical, Legal, and Economic Background

In the EU, two legal instruments are particularly relevant for framing the implications of GenAI developments from a copyright perspective:

The **Copyright in the Single Market Directive** (CDSM Directive) creates a legal framework for **'text and data mining'** (TDM). TDM is a central part of GenAI development, as it is the main process through which content is collected, analysed and used as an **input** to develop an AI model's parameters and weights. This process often requires the reproduction of training content, which may involve the exclusive rights of copyright and database owners. The CDSM provisions on TDM provide for specific limitations to these exclusive rights. Article 3 of the CDSM allows for TDM by scientific research organisations while Article 4 allows TDM by any user, including commercial AI developers. Importantly, the exception under Article 4 is subject to rights holders ability to **reserve their exclusive reproduction rights**, commonly referred to as **'opting-out'** of the TDM exception. To be valid, such an opt-out reservation must be **made expressly**, **by the right holder**, and in an **appropriate manner**, including **'machine-readable means'** for content made publicly available online. To use content for training where

an opt-out reservation has been placed, AI developers need an authorisation by the right holder, for example through licences.

The **EU Artificial Intelligence Act (AI Act)** sets out a regulatory framework for AI technologies in the EU, with specific obligations on the providers of general–purpose AI (GPAI) models. Regarding copyright, these obligations refer to the **compliance with Article 4 of the CDSM Directive**, on the TDM opt-outs expressed by copyright holders. The AI Act addresses a broad range of concerns such as risk management, transparency, data governance, ethical considerations and compliance with fundamental rights across all AI systems. GPAI system providers are also required to **publish sufficiently detailed summaries of the training data** they utilise, to facilitate the ability of copyright holders to enforce their rights where relevant. The AI Act also places obligations on the deployers of GenAI systems to **ensure that generative output is detectable** in a machine-readable format.

The global GenAI landscape involves a rising number of legal disputes between rights holders and GenAI system providers, with a substantial number occurring in the United States of America (USA). To date, there have been four court cases identified in the EU that relate to copyright and AI training, the September 2024 case ***Kneschke vs. LAION*** being a noteworthy first. While the German court deemed that LAION (a major provider of text-image datasets used for GenAI training) benefited from the Article 3 CDSM exception for scientific research TDM, it made several *obiter dicta* references that provide insights into how future courts might interpret the legal requirements for valid TDM rights reservations under Article 4 CDSM.

In parallel, several high-value agreements on the use of copyright protected content for AI training have been reached, between rights holders and GenAI developers. **Direct licensing** by copyright holders who effectively opt-out their content from being used under Article 4 CDSM, has the potential to bring new revenues streams. The study identifies several factors driving such agreements, including (i) the perception of impending **data shortages** for machine learning, (ii) the role of **data quality** and the importance of metadata and data annotation, (iii) the **attitude towards risk** of GenAI developers and relative negotiating power, (iv) the role of **synthetic data as a substitute** to training input, and (v) the emergence of **content aggregation services** which serve as commercial intermediaries for smaller rights holders who seek to access the emerging training data market.

While the specific dynamics of direct licensing markets differ between content sectors, the publishing sector (and in particular the press and scientific publishing) is uniquely positioned to take advantage of **licensing opportunities associated with Retrieval Augmented Generation** (RAG, see also part on GenAI Output) applications that are central to the development of some GenAI services.

Several key considerations that may affect **licensing terms** are also identified, including (i) the **development of benchmark market rates**, (ii) **the metrics used for remuneration** (iii) innovation in the types of licensing being offered, (iv) the potential to **link input-based and output-based licensing permissions**, and (v) **reciprocal exchange of commercial assets**. The evolution of these aspects should be followed to understand the dynamic of direct licensing markets, as **standard contractual practices and norms** eventually emerge.

An emerging issue is the potential for **'data laundering'** to arise from the interplay between scientific-research TDM activities covered by Article 3 CDSM Directive, and commercial TDM activities for AI training covered by Article 4 CDSM Directive. The relationships between scientific researchers building datasets pursuant to Article 3 CDSM Directive, and commercial AI developers using these datasets for their own purposes, has raised concerns of scientific research privileges being exploited for commercial purposes.

Generative Artificial Intelligence Input

Data collection process is the first stage in GenAI training, and it must comply with copyright obligations. Depending on the context, copyright obligations may include respecting TDM optouts, or where necessary, entering into direct licensing agreements with rights holders. Collected data must then be cleaned, annotated, and processed before it is used in the AI training, which consist of multiple stages from **model pre-training** to **model fine-tuning,** and possible reinforcement learning.

While several large datasets are publicly available for AI training, they may include **pirated content**, as well as unspecified, incorrect, or standard **licences not tailored to the actual use of the dataset**. These issues may result in copyright liability passing down the AI value chain from the AI dataset creator to the GenAI developer and GenAI service deployer, all of whom must comply with their obligations under EU copyright law and the AI Act.

Content publicly available online is a central source of data used in AI training processes. While **web crawling** has traditionally been used for search engine indexing, **web scraping** is now widely used to collect massive quantities of data for the development of AI training datasets. As a result, many of the measures used by copyright holders to control access to their works, focus on addressing this practice. The **Robots Exclusion Protocol** (REP) currently serves as a *de facto* standard for managing web crawling and scraping activities and has largely been deployed as a primary strategy for TDM rights reservations. However, there is a prevailing consensus amongst stakeholders that REP is not optimal as a TDM opt-out mechanism and serves more as a temporary solution. This is mainly due to REP's **inherent limited granularity and use-specificity**, its need for intermediation by website managers, unenforceability, and the voluntary disclosure of web-crawler identities. In that respect, REP is also sometimes complemented by traffic management strategies for restricting web-crawlers access to online content in the first place.

Given the complexity of the AI ecosystem, and the specific needs and business models of different content sectors, **no single opt-out mechanism** has emerged as the sole standard used by rights holders. Instead, **legally-driven measures** and **technical measures** are used by rights holders to express their TDM rights reservations. The legally-driven measures for rights reservations reviewed in the study include unilateral declarations, licensing constraints, and website terms and conditions. Meanwhile, the technical measures for rights reservations include REP, TDM Reservation Protocol (TDMRep), Robots Meta Tags, the C2PA Content Authenticity Initiative, the JPEG Trust standard, as well as services developed by SpawningAI, the Liccium Trust Engine Infrastructure (linked to the ISO ISCCcode identifiers), and Valuenode's Open Rights Data Exchange platform.

The study is comparing such measures in relation to seventeen key criteria: (i) typology, (ii) user-specificity, (iii) use-differentiation, (iv) granularity, (v) versatility, (vi) robustness, (vii) timestamping, (viii) authentication, (ix) intermediation, (x) openness, (xi) ease of implementation, (xii) flexibility, (xiii) retroactivity, (xiv) external effects, (xv) generative application, (xvi) offline application, and (xvii) market maturity. This analysis supports the understanding on the **respective advantages and limitations** of the different measures to support the expression and implementation of TDM reservations by right holders, their readability by TDM users, as well as their effectiveness to support licensing for different use cases.

In general, none of the reservation measures analysed support enforcement of an expressed reservation. TDM users are generally responsible for properly configuring their data collection policies, scraping tools, and data cleaning procedures, to comply with expressed TDM reservations. Legally-driven measures are typically applied to specific copyright-protected works, but also entire repertoires of works. Technically-driven measures are categorised as either **'location-based'** (i.e., associated to the location of a piece of content online) or **'asset-based'** (i.e., associated with the actual content irrespective of where it is made available online). Both approaches have their distinct advantages and limitations.

The diversity of measures is reflected in the indications from stakeholders interviews that their content management and rights reservation strategies often use a combination of various legal and technical measures.

The study identifies a trend towards **open standards** and open-source licensing in technical reservation solutions to support wide adoption and interoperability. Stakeholders on both the right holder and GenAI development sides of the TDM process generally seem to support increased efforts for **standardisation of rights reservation measures**, as well as the **flexibility to incorporate multiple measures** to adapt to different use cases. As the GenAI ecosystem keeps evolving, a number of standard practices are expected to emerge to address conceptual and practical challenges in adapting reservation measures to the specific needs of different content sectors and use cases throughout the AI value chain.

The current situation regarding rights reservation measures suggests a r**ole for public authorities**, such as national IP offices or similar national or supranational institutions. Institutional support may take the form of **technical support** in implementing and administering federated databases of TDM reservations expressed by right holders. **Nontechnical support** may consist of increasing public awareness of the copyright issues surrounding the deployment and use of GenAI technologies, providing information on various rights reservation measures (including comprehensive lists of web scraper identifiers), and analysing industry trends in terms of technical developments and commercial licensing terms.

Generative Artificial Intelligence Output

The technical process of content generation depends on the type of GenAI model, as typical model architectures differ between the types of content they generate. Given the high costs of

training AI models and the inherent limitations of constantly (re)training models on new content, there is a trend of **increased deployment of RAG technologies** that combine aspects of information retrieval mechanisms with GenAI capabilities. This improves model performance without having to frequently (re)train models on updated training datasets. RAG is gaining prominence in AI-driven search engines, also known as 'answer engines', presenting new challenges and opportunities for copyright holders. RAG comes with its own copyright issues that may depend on whether the application is based on static RAG and locally stored content used for retrieval, or on dynamic RAG which may incorporate forms of web scraping.

Given that the AI Act requires transparency on the content produced by GenAI systems, several measures have been developed to **identify and disclose the nature of synthetic content**. These **generative transparency measures** include **provenance tracking,** (including the C2PA Initiative, the JPEG Trust Initiative, and the block-chain based Trace4EU project), **detection measures** for AI-Generated content (including StyleGan3-detector for images, or Deezer's detection methods for audio), as well as **content processing solutions** (including various protocols for watermarking and digital fingerprinting), and **membership inference attacks**.

This study compares a selection of these **generative transparency measures** in relation to ten key criteria: (i) typology, (ii) versatility, (iii) openness, (iv) market maturity, (v) human readability, (vi) cost implications, (vii) robustness, (viii) interoperability, (ix) scalability, and (x) reliability. This comparison supports the understanding on the relative advantages and limitations of each measure.

Once a model is trained on input data, the patterns and correlations extracted during the machine learning process are embedded in its parameters. The extent to which these representations influence the model's outputs depends on its architecture. While some GenAI models abstract knowledge in a way that makes direct extraction of training data unlikely, others – particularly LLMs and generative vision models – may exhibit **'memorisation'**. This may lead to a situation where certain outputs can closely resemble or even replicate training inputs. Memorisation is thus a technical issue which creates a legal issue, with potential for **plagiaristic output** and **content 'regurgitation'** (explicit reproduction of the trained content).

GenAI system providers have developed various technical solutions to address memorisation. These measures include various tools to **compare generated content** with potential input sources, **filters for preventing duplicative output**, and different approaches to **prompt rewriting or filtering**. An emerging technical research field to address these issues consist of **'model unlearning'** and **'model editing'**. These are methods for erasing, adjusting or updating the information coded into the model's parameters, enabling AI developers to solve issues detected after the model's deployment. In addition to these technical measures, other means are also used to address the challenge of potentially infringing output. Several GenAI system providers offer some form of **legal indemnification** to mitigate the risk for their customers.

The issues surrounding GenAI outputs and copyright also suggests a potential **role for public institutions** active in the field of IP. **On information for GenAI developers and policy makers** they could openly share information on measures available to mitigate potential infringing output and detect synthetic content, and good practices developing in that field. **On information for the general public**, they could provide information on ethical prompts usage and cooperate with other relevant bodies to increase the public's capacity to identify generative output. On the **technical side**, public institutions could serve as forums for information sharing and collaboration supporting the interoperability of output transparency measures across platforms and GenAI systems.

Concluding observations

The study takes a structured approach to clarify, from a technical point of view, the interaction between GenAI and copyright. The study shows, firstly, that **no single solution has emerged as the sole standard** opt-out mechanism for rights holders to express their TDM rights reservations, or transparency measure to identify and disclose the nature of synthetic content. Secondly, although the global GenAI landscape involves a **rising number of legal disputes**, the study also notes that **several high-value agreements have been reached** between rights holders and GenAI developers. Lastly, the current situation suggests a **possible role for public authorities** in providing technical support for implementing and administering databases of TDM reservations and raising awareness on measures and good practices to mitigate potential infringing output.

As a disrupting technology, the development of GenAI has caused shifts in the creative and the IT industries, and significantly altered how rights holders and AI developers operate. While it may take some time before a new balance is established, the study importantly showed the relevance of accessing essential information about works' origin and permissible uses in view of proper respect, benefit and enforcement of copyright.

1 Introduction

1.1 Purpose and Scope

This report provides a technical analysis on the interface between **Generative Artificial Intelligence ('GenAI')** and EU Copyright Law. It is published by the European Union Intellectual Property Office (EUIPO) as part of the work of the *European Observatory on Infringements of Intellectual Property Rights* ('the Observatory'), which is a network of public and private experts and specialist stakeholders.

In recent years, the development of Artificial Intelligence (AI) technologies, especially GenAI, have been at the centre of public attention and debate. **GenAI systems, including Large Language Models (LLMs), draw insights from large quantities of training data to develop algorithmic processes which can generate and output new content with similar characteristics**. Rapid developments in these technologies and their widespread use and deployment have resulted in **rising concerns about copyright-related implications**. While these technologies represent new forms of innovation and have the potential to transform the creative industries, they also create tension with the interests of copyright holders. In any event, such technologies must be developed and managed in a manner consistent with applicable intellectual property laws.

The European AI Strategy, published by the European Commission in 2018, stressed that *"Reflection will be needed on interactions between AI and intellectual property rights, from the perspective of both intellectual property offices and users, with a view to fostering innovation and legal certainty in a balanced way."* ([1]) Pursuant to this strategy, the EU was the first jurisdiction in the world to adopt a comprehensive legislation on the regulation of AI technologies, in the form of the *Regulation (EU) 2024/1689*, commonly referred to as the 'AI Act', adopted in June 2024. These legal developments should be considered alongside existing EU laws on the protection of intellectual property rights, including specific copyright

([1]) Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions - Artificial Intelligence for Europe, European Commission, 25 April 2018 COM(2018) 237.

provisions establishing the exceptions for 'text and data mining' and the notion of rights reservations ('opt-outs' of text and data mining uses) made by rights holders.

The implications of GenAI technologies on the European intellectual property landscape have been discussed in detail within the various Expert Groups of the EUIPO Observatory. In March 2022, the EUIPO published a study on the *'Impact of Artificial Intelligence on the Infringement and Enforcement of Copyright and Designs'* ([2]). In February 2022, the European Commission also published two reports - *'Opportunities and Challenges of Artificial Intelligence Technologies for the Cultural and Creative Sectors'* ([3]), and *'Study on copyright and new technologies: copyright data management and artificial intelligence'* ([4]). These discussions and reports have taken place in parallel with the European Commission's activities regarding the legal framework for the regulation of AI technologies.

Given these various developments, the Observatory commissioned this study to document and foster a deeper understanding of technical developments, issues, and solutions in terms of the interface between copyright law and GenAI systems. This study complements ongoing activities within the European Commission AI Office and Copyright Unit. According to the 2025 Work Programme of the Observatory, it is anticipated that following this study, and in close cooperation with the European Commission, the EUIPO will explore the possibility of developing services facilitating opt-out mechanisms, and respect of the opt-out expressed for the benefit of both copyright holders and AI companies ([5]).

The main objective of the study is to analyse copyright implications at both the input and output stages of GenAI, focusing on the related **technical solutions:**

([2]) Study on the Impact of Artificial Intelligence on the Infringement and Enforcement of Copyright and Designs (Impact of Technology Deep Dive Report I), European Union Intellectual Property Office (EUIPO), March 2022.

([3]) Opportunities and Challenges of Artificial Intelligence Technologies for the Cultural and Creative Sectors, SMART 2019/0024, European Commission, Directorate-General for Communications Networks, Content and Technology (DG-CNCT), February 2022.

([4])Study on copyright and new technologies: copyright data management and artificial intelligence, SMART 2019/0038, European Commission, Directorate-General for Communications Networks, Content and Technology (DG-CNCT), February 2022.

([5]) Work Programme 2025, European Union Intellectual Property Office (EUIPO), European Observatory on Infringements of Intellectual Property Rights, October 2024.

On the input side, the purpose is to analyse technical solutions and practices currently used, or still under development, to reserve, limit or licence the **use of copyright protected works as training material** for the development of GenAI systems.

On the output side, the purpose is to analyse technical solutions and practices to **identify content** generated by, or with the help of GenAI, as well as practices to **prevent the generation of content that might infringe** on existing intellectual property rights. The central premise of this study is thus a 'solution-driven approach'.

The scope of this study includes a **background analysis** of technical, legal and market developments, encompassing an examination of GenAI input and output processes. The focus is explicitly on the copyright implications of these processes, and the measures used by different actors across the AI ecosystem to address copyright management concerns. Where relevant, the study addresses economic and institutional considerations.

To understand this report, readers should consider three important contextual notes.

- **First, no aspect of this report should be understood as implying any specific legal interpretation of any particular legal provisions**. Any discussion on legal provisions is intended solely to highlight the different elements of the legal environment surrounding EU copyright law and GenAI, as well as its influence on GenAI input and output measures and without prejudice how the Commission and the European Court of Justice of the European Union (CJEU) may interpret these concepts. This is critical as many legal concepts in this domain are novel and evolving, and their legitimate interpretation falls within the purview of competent judicial authorities.

- **Second, while this report aims to review key technical measures adopted within the GenAI ecosystem, no aspect of the analysis should be taken as an endorsement of any specific measure**. While this study seeks to be comprehensive in scope, the technical measures analysed are not exhaustive, and the omission of any particular measure, provider or development should not be interpreted as a lack of relevance or significance.

- **Third, at many instances in this report, references are made to copyright law** in general using terms such as 'rights holders' and 'copyright protected works'. **These**

references should be understood in broad terms unless specified otherwise, and Chapter 2 of this report clearly addresses the various rights holders that comprise the ecosystem under analysis. Where these broad terms are used, the analysis should not be interpreted as intentionally excluding other relevant rights holders whose intellectual property rights are not in the strict sense considered to be 'copyright' (including in particular, database owners and holders of related rights). Any exceptions to this general approach will be made clear by context.

1.2 Methodology

The research on which this report is based was completed between September 2024 and March 2025. Its methodological approach consisted of two main streams of research activity conducted in parallel: **desk research** and **stakeholder interviews**. These research activities were organised in relation to three main components, each of which corresponds to a substantive chapter in this report: (i) Technical, Legal and Economic Background, (ii) GenAI Input, and (iii) GenAI Output. Each of these components was then analysed in terms of various sub-components which considered all relevant technological, legal, and economic dimensions of the interface between GenAI and copyright law, specifically in the EU context.

Desk research activities were conducted to review a wide range of publications and sources including academic journal articles, industry white papers, and reports from national and international institutions. Given the dynamic and evolving nature of the AI market, desk research activities also included reviews of 'grey literature' sources such as technology blogs, industry discussion fora, and press articles. In relation to technical content, the research process focused on measures used by various AI industry stakeholders to manage copyright issues both on the input and output sides of GenAI processes. The findings of the desk research process were used to develop and refine questions that guided stakeholder interviews.

Interviews were conducted by identifying potential interviewees within four key stakeholder groups: **AI Companies** (providers of AI models and systems), **Technical Solution Providers** (technology infrastructure and service providers), **Rights holders** (from various content sectors), and **Public Organisations** (civil society organisations and government agencies).

In total, 30 interviews were conducted during the study period. An anonymised list of interviewees is reported at *Annex II*.

Based on the desk research findings, general interview templates were developed for each stakeholder group. These templates are reported as *Annex III*. Prior to each interview, questionnaire templates were modified to apply specifically to each interviewee, based on a review of publicly available information about the stakeholder. These specific questionnaires were distributed to interviewees in advance, and the interviews took a semi-structured form. The purpose of the interviews was to validate findings from the desk research, document perspectives on GenAI and copyright issues that would otherwise be absent from public literature, and to gain detailed insights into the practical issues facing different stakeholder groups. The interviews also contributed to the identification of TDM reservations measures, non-reservation measures, as well as generative transparency solutions, that were subsequently analysed and compared. Subsequent to each interview, a written follow-up questionnaire was conducted with the interviewee in order to gather further technical details and clarify key points.

Further insights were gained from two workshops conducted in December 2024 and March 2025, where preliminary findings of the study were presented to a group of experts (from the Observatory's '*Cooperation with Intermediaries' and 'Impact of Technologies*' Expert Groups). Overall, stakeholders' engagement was extremely high, with interviewees showing great appreciation for being invited to participate in the study, and a strong willingness to share insights. For the purpose of preserving confidentiality of stakeholder's business and proprietary interests, insights from interviews are generalised and anonymised when incorporated into this report.

2 Technical, Legal and Economic Background

2.1 Evolution of Artificial Intelligence

The concept of artificial intelligence (AI) gained scientific traction in the 1950s with the Alan Turing Test (Turing, 1950) and the Dartmouth Conference of 1956, where John McCarthy coined the term 'artificial intelligence' (Russell et al., 2010) ([6]).

The high expectations for this technology went unmet for several decades, leading to multiple periods of stagnation known as 'AI winters', until the 1990s, when the necessary advancements in hardware technologies were achieved. The 1990s saw a revival of **machine learning** and neural network research, driven in part by renewed government funding in the USA. This resurgence thrived alongside the rapid growth of the internet, the widespread adoption of personal computers, and advancements in character and speech recognition.

By the 2000s and 2010s, further advancements in computing, specifically the use of Graphics Processing Units (**GPUs** ([7])), allowed for faster processing of vast datasets, such as ImageNet ([8]), that provided an essential resource for training **deep learning** models. These models enabled automatic extraction of complex patterns from data and started to outperform traditional statistical methods. Combined with improvements in **neural network** training, this led to breakthroughs in fields like image recognition, natural language processing, and autonomous systems.

([6]) In 1961, Joseph Weizenbaum developed ELIZA, a computer program that simulated conversation by responding to human input with natural language and pre-programmed empathetic replies. As one of the earliest chatbots, ELIZA showcased the potential for machines to mimic human-like dialogue. Notably, Weizenbaum noticed that users frequently ascribed human qualities, such as understanding and empathy, to the program, a phenomenon that became known as the 'Eliza effect' (Weizenbaum, 1976).

([7]) A GPU (Graphics Processing Unit) is a specialised electronic circuit designed to accelerate the processing of images, videos, and computations. Initially developed for rendering graphics, GPUs are now widely used in parallel processing tasks, such as machine learning, scientific simulations, and cryptocurrency mining, due to their ability to handle multiple tasks simultaneously.

([8]) ImageNet is a large-scale, labelled image database designed for visual object recognition research. It contains over 14 million images categorised into thousands of classes, organised according to the WordNet hierarchy, and has been widely used to train and benchmark computer vision models.

2.1.1 The Rise of Generative AI

By the mid-2000s, improvements in neural network training methodologies, increased access to extensive datasets, and advancements in computational capabilities yielded significant breakthroughs in applications such as speech recognition and image classification. These innovations laid the foundational groundwork for Generative Artificial Intelligence (GenAI). GenAI is a subfield of artificial intelligence leveraging deep learning techniques to synthesise novel content rather than merely analysing existing datasets.

Central to modern deep learning systems are **neural networks**, computational architectures consisting of multiple interconnected layers of nodes, each characterised by weighted parameters. These layers process input data by identifying fundamental features and progressively abstracting them into more complex structures, thereby enabling neural systems to discern and represent intricate data patterns effectively. Such advancements have precipitated substantial progress across diverse AI domains, including facial and image recognition, natural language processing, and conversational agents (chatbots). An example is Apple's introduction of Siri in 2011, representing the first widely recognised virtual assistant.

These methodological advancements facilitated the emergence of **Generative Adversarial Networks (GANs)**, first introduced by Goodfellow et al. (2014), significantly transforming GenAI. GANs employ an **adversarial two-player framework** consisting of a generator and a discriminator iteratively refining their performance through mutual competition, as shown in *Figure 2.1.1-1* ([9]). This novel adversarial training methodology notably advanced generative model capabilities.

([9]) Within the adversarial framework, the generator network produces synthetic data samples, and the discriminator network assesses their authenticity. The generator iteratively enhances its performance by minimising adversarial loss, progressively generating increasingly realistic outputs.

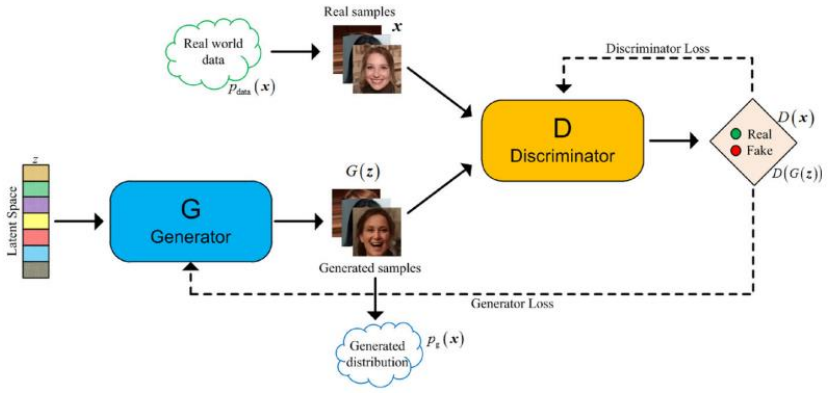


Figure 2.1.1-1: Adversarial training of both the discriminator and generator of a GAN architecture (Ahmad et al., 2024).

While GANs popularised adversarial training as a powerful mechanism for generating high-quality synthetic datasets, they were not the sole generative model paradigm emerging during this time. **Variational Autoencoders (VAEs)** were concurrently developed, presenting an alternative approach predicated on **probabilistic modelling** rather than adversarial dynamics ([10]). The properties of VAEs render them particularly effective for applications necessitating structured and smooth data interpolation. *Figure 2.1.1-2* illustrates the primary components of VAE training, delineating how the encoder maps input data into latent distributions and how the decoder subsequently reconstructs data samples from these latent representations.

([10]) VAEs integrate autoencoder architectures—neural networks designed explicitly to compress data into condensed representations and subsequently reconstruct them—with latent variable models, which characterise underlying data structures through continuous probability distributions. Rather than mapping input data directly onto deterministic latent vectors, VAEs encode inputs into probabilistic latent spaces, facilitating the generation of new samples via sampling from these learned distributions.

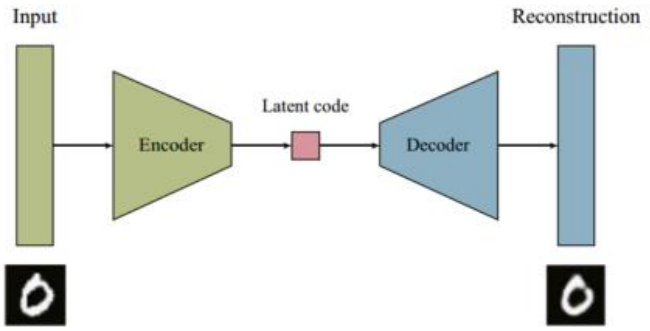


Figure 2.1.1-2: Architecture of a Variational Autoencoder (VAE) (Mehrjardi et al., 2023). The "Input" is the training data, the "latent code" is the model's learnt distribution of the training data (encoded into the model's parameters) and the "Reconstruction" is the output generated.

Additionally, **Diffusion Models** achieved image generation by simulating a process that gradually transforms a simple data distribution into a more complex one, mimicking the target distribution. Diffusion Models **generate images by starting with random noise and gradually transforming it into a clear image**. This process simulates how complex data, like images, can be created from a simple, random starting point by slowly adding structure and details over time. In *Figure 2.1.1-3* below, the iterations of the denoising process performed by Stable Diffusion, a widely recognised AI system based on a Diffusion Model, are illustrated.

Figure 2.1.1-3: Intermediate steps of the denoising phase at the base of Diffusion Models' capability to generate images (World Intellectual Property Organization, 2024)

The landscape of GenAI was further enriched by the emergence of powerful deep learning architectures like **Transformers** ([11]). While **not strictly generative models themselves**, Transformers excel at handling massive datasets and capturing **long-range relationships within sequential text data** ([12]). This achievement paved the way for **LLMs** built upon Transformer architectures (World Intellectual Property Organization, 2024). In particular, the public launch of OpenAI's LLM 'ChatGPT' in late 2022 marked an important shift in deployment, public awareness, and public use of GenAI systems. The table in *Annex IV* lists some examples of GenAI models available by January 2025, with their type and usage.

GenAI systems need **high-performance hardware** like GPUs or TPUs ([13]). ([14]) To improve scalability, this often leads to the adoption of a **cloud or an edge-computing infrastructure** (Wang et al., 2023), typically built through partnerships with cloud service providers such as Microsoft Azure, Amazon Web Services (AWS), or Google Cloud.

Recent advancements in GenAI architectures, however, suggest a shift towards more energy-efficient models with lower computational demands. For instance, DeepSeek, an emerging LLM framework, has been reported to require significantly less computing power compared to traditional models of similar scale (Meng et al., 2025). While this remains an evolving landscape, such optimisations could influence the future infrastructure needs of GenAI systems, potentially reducing dependency on large-scale GPU clusters and cloud-based computing.

([11]) For the detailed definition of 'transformer' see the *Glossary.*

([12]) In sequential data like text, relationships between words or concepts can occur over long distances. Transformers excel at identifying and learning these long-range dependencies because of their self-attention mechanism. This mechanism enables the model to weigh and relate every word in a sequence to every other word, regardless of distance.

([13]) A TPU (Tensor Processing Unit) is a custom chip by Google designed to accelerate machine learning tasks. Compared to GPUs, TPUs offer faster processing, higher efficiency, and lower power consumption for deep learning models, making them more suitable for large-scale AI workloads.

([14]) For example, although the exact computing infrastructure used by the ChatGPT service is not publicly available, one can estimate the cost to run the service each day based on the model architecture of GPT-3. To deploy such a large model, a distributed computing system with at least 2,048 CPUs and 2,048 GPUs is needed.

2.1.2 Relevant Aspects for Generative AI systems

GenAI systems are complex and are based on the interaction of several technologies and processes that can involve human intervention, each designed to fulfil a specific function. Relevant aspects are:

- **Training dataset:** large collection of data ingested by the model during training. In general, a model must be trained on the same data type of its output, e.g., language models are trained using text data, text-to-image models are trained using captioned images, and so on.

- **Training data collection**: training data can be collected using different methods, detailed in *Section 3.1.2*. The strategies adopted influence both the overall quality of the dataset, and consequently the quality of the final model, as well as the compliance with any rights that may exist in the collected data.

- **Data cleaning and tokenisation:** preparing inputs by removing irrelevant data and segmenting text into tokens (see *Section 3.1.3*);

- **Machine learning**: a method for training models that allows systems to identify patterns, eliminating the need for explicit programming, as discussed more extensively next in *Section 2.1.2.1*.

- **Foundation models**: large AI systems **pre-trained** on vast datasets, enabling versatility across tasks like text, images, sounds or video processing. They learn general patterns, which can be fine-tuned for specific applications (as detailed below in *Section 2.1.2.2*).

- **Refined (or fine-tuned) models**: Foundation models are further trained or fine-tuned on **task-specific** data to specialise them (an example, fully detailed in *Section 3.1.4* is ChatGPT, a model derived from GPT);

- **Retrieval-Augmented Generation (RAG)**: technique that combines generative capabilities with an **external knowledge base**, comprising of information stored in documents or databases rather than acquired through machine learning, to improve

accuracy and context relevance (for example, in applications used for customer support, education, healthcare, legal analysis and content creation);

- **User Interface:** provides access to GenAI capabilities, often through a textbox for a prompt, but it can also be a visual image/text editor.

2.1.2.1 Machine Learning (ML)

Machine learning is the technique used to train models that enable systems to learn patterns and make predictions or decisions without being explicitly programmed.

The main types of ML are:

- **Supervised Learning**: a technique where models learn from labelled data (the labels are essentially classes splitting the training data into categories) by trying to predict the labels and minimising prediction errors;
- **Unsupervised Learning**: a method for discovering patterns, structures, or relationships in unlabelled data. Differently from supervised learning, the training happens without indications on expected outputs;
- **Semi-Supervised Learning**: a hybrid approach that combines a small amount of labelled data with a large amount of unlabelled data to improve learning;
- **Reinforcement Learning**: a process where an agent learns to make decisions by interacting with an environment and maximising cumulative rewards (e.g., a model can be trained to play a videogame by trying random inputs and learning which ones deliver the highest score);
- **Reinforcement Learning with Human Feedback (RLHF)**: technique used to manually align the model's behaviour with user preferences; the **human labels the training data** to indicate the expected output, then the model's parameters are fine-tuned to obtain the correct results (supervised fine-tuning) or to maximise a reward function, which is a numerical value assigned to model outputs to guide the learning process by reinforcing desirable behaviour;

- **Deep Learning**: a subset of ML that uses multi-layered neural networks to extract and learn complex data representations. Each layer applies transformations to the layer's input data, and its output forms the next layer's input. For example, when deep learning is performed with images, each layer can make different computations on a subset of the image's pixels to learn the relations between them;
- **Online Learning**: a dynamic approach where models are updated incrementally as new data becomes available, supporting real-time learning;
- **Transfer Learning**: a method of reusing knowledge from one task to enhance learning in a related but different task.

2.1.2.2 Foundation Model (FM)

Foundation models (FMs) are a class of AI models characterised by their ability to generalise across tasks by pre-training on massive, diverse datasets. These models leverage architectures like transformers and can work with different types of data, including text, images, sounds, video and multimodal inputs. In 2021, IBM ([15]) defined FMs as the *"future of the AI: flexible, reusable AI models that can be applied to just about any domain or industry task."*

A considerable number of FMs were released between September 2023 and March 2024, ranging in size, modality, and capability. According to the *Stanford Center for Research on Foundation Models* (CRFM), over 120 FMs were publicly released in this period, bringing the total number of known FMs globally to over 330. Some examples of FMs are: Mistral Large, Anthropic Claude 3, Stability AI Stable Cascade, OpenAI Sora, Google Gemini 1.5 ([16]).

FMs are generally made available on a spectrum from closed (e.g., proprietary, commercial, or internal-use models) to open-source (e.g., models with weights and training instructions available to users) ([17]).

([15]) What are foundation models?, IBM Research, 9 February 2021 (accessed 14 March 2025).

([16]) AI Foundation Models technical update report, CMA, 16 April 2024 (accessed 14 March 2025).

([17]) For a list of open-source models, *see* Introducing the European Open Source AI Index', European Open Source AI Index (accessed 14 March 2025).

2.2 Legal Framework for Generative AI

The previous section outlines how model training, using large datasets, is a critical stage in developing GenAI. In some instances, datasets may contain content protected by Intellectual Property (IP), such as copyright. As a result, rights of copyright holders need to be respected when GenAI developers utilise such data and content. This section outlines the **applicable legal framework for copyright and related rights** within the EU, along with the specific **obligations** imposed by EU law on actors involved in GenAI ([18]).

2.2.1 Rights holders, Exclusive Rights and Exceptions

For this study, the focus is on copyright and related rights, although there may be issues with other intellectual property rights such as trade mark or design rights, that are not addressed.

While IP rights cover a large portion of creative content in circulation today, a considerable amount of material, potentially used for GenAI training, might fall outside this protection and is in the **'public domain'**. Public domain may include materials that were never eligible for copyright or related rights protection in the first place, works whose copyright protection has lapsed due to expired terms, and materials for which rights holders intentionally waived their exclusive rights. While IP law may not restrict the use of such content for training, **other legal constraints or conditions may apply**. Contractual agreements, such as a website's terms and conditions, may impose limitations on its use (see *Section 2.2.3*). Additionally, while personal data about individuals may not be covered by copyright, its usage could be governed by data protection laws, particularly the General Data Protection Regulation (GDPR) ([19]).

([18]) The analysis is grounded in relevant EU legislation and does not extend to the national implementations of EU directives by Member States.

([19]) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) OJ L 119, 4.5.2016.

2.2.1.1 Reproduction Right

The foundations of modern EU copyright law are set out in *Directive 2001/29/EC* commonly referred to as the 'Information Society' (or 'InfoSoc') Directive ([20]). Article 2 of the InfoSoc Directive sets out an exclusive **'reproduction right'** which is enjoyed by authors in relation to their copyright-protected works, as well as specific beneficiaries of related rights ([21]). Reproduction of any work (or the specified subject matter protected by related rights) requires authorisation, irrespective of whether a reproduction is temporary or permanent, the means or form of production, and whether the reproduction is whole or in part.

While authors are by default the initial owners of copyright in their works (as are performers, producers, and broadcasting organisations in relation to their respective subject matter protected by related rights), exclusive rights, including the right of reproduction, may be transferred, assigned, or contractually licenced to another party ([22]). The prevalence of full contractual assignment of copyright differs by content sector. In certain sectors, it is common practice to **contractually assign (or exclusively licence) copyright to an intermediary** which acts as the economic agent mandated to manage the commercialisation of a work. For example, this model is common in the music and literature sectors, where publishers are designated as rights holders by authors through assignment or licence and play key roles in content distributions and licensing.

Copyright and related rights are subject to **'limitations and exceptions'**, which permit certain acts of reproduction without the explicit authorisation of a right holder. In the EU, the InfoSoc Directive (and other relevant copyright Directives) set out mandatory and optional limitations and exceptions to exclusive rights, the latter being at the discretion of Member States to implement into their national laws. This system of a 'closed list' of permissible exceptions is in contrast with the 'open standard' of 'fair use' in the USA copyright law, which is not limited to specific statutory categories of uses but is instead applied flexibly on a case-by-case basis ([23]).

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

([21]) Ibid. Article 2 grants an exclusive right of reproduction to the following rights holders in respect of their respective subject matter: (a) authors, for their works; (b) performers, for fixations of their performances; (c) phonogram producers, for their phonograms; (d) film producers, for the originals and copies of their films; and (e) broadcasting organisations, for fixations of their broadcasts.

([22]) Ibid. Recital 30.

([23]) Title 17 USC §107 (U.S. Code Title 17, Section 107: Fair Use).

2.2.1.2 Database Rights

Separate exclusive rights in databases are set out in *Directive 96/9/EC* ('Database Directive') ([24]). 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"* ([25]). The directive provides for two distinct forms of protection: (1) **copyright protection for original databases** (which *"by reason of the selection or arrangement of their contents, constitute the author's own intellectual creation"*) ([26]), and (2) **a *sui generis* right** (subject matter-specific unique intellectual property right) in databases (where *"there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents"*) ([27]).

For databases protected by copyright, database authors have exclusive rights to authorise reproductions of their databases, irrespective of whether such reproductions are temporary or permanent ([28]). As for databases protected by a *sui generis* database right, the maker of the database has the right to **prevent extraction and/or re-utilisation** (of the whole or of a substantial part, evaluated qualitatively and/or quantitatively) of the contents of that database ([29]).

However, the EU Data Act (*Regulation 2023/2854*) clarified that data obtained from or generated by so-called 'Internet of Things' (IoT) products and services (i.e., connected products that obtain, generate, or collect environmental data and are able to communicate such product data) **is not eligible for protection** under the *sui generis* database right.

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

([25]) Ibid. Article 1.

([26]) Ibid. Article 3.

([27]) Ibid. Article 7.

([28]) Ibid. Article 5.

([29]) Ibid. Article 7(1).

2.2.1.3 Rights in Computer Programs

Computer programs are protected under EU copyright law by Directive 2009/24/EC ('Computer Programs Directive') ([30]) as literary works ([31]). Nevertheless, the Directive sets out various provisions relevant for **computer programs as unique subject matter**, which might not apply to more traditional literary works, and provides for specifics right of reproduction (permanent or temporary) and alteration (including translation, adaptation, and arrangement) ([32]). Additionally, the Directive specifies that protection is not extended to logic, algorithms and programming languages, to the extent that such comprise ideas and principles ([33]).

2.2.1.4 Press Publisher's Rights

Publishers of press publications have specific rights provided for in Directive (EU) 2019/790 ('Copyright in the Digital Single Market Directive' or 'CDSM Directive') ([34]). Under the CDSM Directive, 'press publication' generally (but not strictly) refers to collection of **literary works of a journalistic nature** periodically published under a single title, with some informative purpose, and under some editorial control.

Publishers of press publications are the beneficiaries of specific protection (neighbouring right) against certain **online uses** by information society service providers (such as social media platforms and search engines). Commonly referred to as a 'press publisher's right', it grants publishers the exclusive rights of reproduction and 'making available to the public' as established in the InfoSoc Directive. It does not affect the rights of the authors of individual press articles which are incorporated into a press publication. In practice, authors of literary

([30]) Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version) (Text with EEA relevance). OJ L 111, 5.5.2009, p. 16–22.

([31]) Ibid. Article 1.

([32]) Ibid. Article 4.

([33]) Ibid. Recital 11.

([34]) Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.)

works like newspaper articles may or may not entirely assign their rights to press publishers based on employment or other contractual relationships.

2.2.1.5 Technological Protection Measures and Rights-Management Information

In addition to the specific exclusive rights in copyright and related rights, the InfoSoc Directive also provides for certain protections in relation to 'technological protection measures' (TPM) and 'rights-management information', both of which play an important role in rights management.

TPMs are means (such as access controls or encryption) used to prevent or restrict acts which are not authorised by a right holder in relation to their respective subject matter. The national laws of EU Member States are required to provide adequate legal protection against the **circumvention of effective technological measures** used by rights holders. Importantly, technological measures should not prevent the beneficiaries of specific limitations and exceptions to copyright and related rights to benefit from them.

Rights-management information refers to information provided by a right holder regarding the identification of a work (or respective subject matter), the author, or the terms and conditions of permitted use. Such information is understood to include both human-readable statements (such as text descriptions attached to a work) and machine-readable statements (metadata which is embedded into digital copies of works). Several content industries have specific industry standards for metadata and rights-management information, including systems of unique codes and alphanumeric identifiers. This information is typically important for not only communicating the scope of permitted uses (and may interface with TPM) but is also used to track distribution and use of works for the purpose of copyright licensing. The national laws of Member States are required to provide adequate legal protection against the **removal or alteration of any electronic rights-management information**.

At EU level, these provisions on protection of technological measures and rights-management information do not explicitly extend to computer programs (as a special copyright subject matter). However, in practice, such provisions generally apply given that computer programs are subsumed under the more general rubric of copyright-protected literary works.

Furthermore, the Computer Programs Directive requires remedies against circulating or possessing means for removal or circumvention of technical devices used to protect computer programs. Under the InfoSoc and CDSM Directives respectively, provisions against circumvention of TPM and removal of rights-management information are explicitly applicable to database rights (both copyright and *sui generis*) and the rights of publishers of press publications.

2.2.1.6 Relevant Exceptions and Limitations

Under EU copyright law a broad definition of exclusive rights encompasses a wide range of acts, some of which are exempted from infringement under specific exceptions – the 'rule versus exception' framework. Traditionally, these exceptions are viewed as derogations from general rules and principles that uphold the fundamental right to property protection. The CJEU has recently interpreted exceptions also as expressions of fundamental rights and interests in turn (Borghi, 2021). As clarified by the Court, the 'rule versus exception' framework necessitates a careful **balancing of conflicting rights and interests**, all of which are equally protected under primary EU law. *Section 2.2.1.12* contains an overview of the exceptions and limitations that are directly relevant in the context of GenAI.

2.2.1.7 Text and Data Mining Exceptions

The CDSM Directive defines **'text and data mining'** (TDM) 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"* ([35]). The significance of TDM has grown considerably due to advancements in computing power and the rise of a data-driven economy. The Recitals of the CDSM Directive articulate the significance, widespread usage, and value of TDM practices in the context of research and innovation:

> *New technologies enable the automated computational analysis of information in digital form, such as text, sounds, images or data, generally known as text and data*

([35]) Directive (EU) 2019/790 Article 2(2).

> *mining. Text and data mining makes the processing of large amounts of information with a view to gaining new knowledge and discovering new trends possible. Text and data mining technologies are prevalent across the digital economy; however, there is widespread acknowledgment that text and data mining can, in particular, benefit the research community and, in so doing, support innovation. …* ([36])
>
> *In addition to their significance in the context of scientific research, text and data mining techniques are widely used both by private and public entities to analyse large amounts of data in different areas of life and for various purposes, including for government services, complex business decisions and the development of new applications or technologies…* ([37])

The content that might be analysed as part of TDM practices may include copyright-protected works (text, images, audio, video, or code), or subject matter protected by other related rights (databases, online press publications, fixations of performances, broadcasts, phonograms, etc.), but also material that is not eligible for copyright protection (such as mere facts or data) and content in the public domain. Since **TDM can involve reproducing protected works**, authorisation from rights holders is required unless a specific exception or limitation applies.

In order to increase legal certainty and an enabling framework to improve the '*Union's competitive position as a research area*' and to '*encourage innovation also in the private sector*', the CDSM Directive introduced **two mandatory exceptions** for the purpose of TDM activities in Article 3 (for research organisations and cultural heritage institutions) and Article 4 (for all other users engaged in TDM).

Article 3 (Text and data mining for the purposes of scientific research)

1. Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for reproductions and extractions made by research organisations and

([36]) Ibid. Recital 8.

([37]) Ibid. Recital 18.

cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.

2. Copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results.

3. Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.

4. Member States shall encourage rightholders, research organisations and cultural heritage institutions to define commonly agreed best practices concerning the application of the obligation and of the measures referred to in paragraphs 2 and 3 respectively.

Article 4 (Exception or limitation for text and data mining)

1. Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.

2. Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining.

3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.

4. This Article shall not affect the application of Article 3 of this Directive.

Article 7 (Common provisions)

1. Any contractual provision contrary to the exceptions provided for in Articles 3, 5 and 6 shall be unenforceable.

> 2. Article 5(5) of Directive 2001/29/EC shall apply to the exceptions and limitations provided for under this Title. The first, third and fifth subparagraphs of Article 6(4) of Directive 2001/29/EC shall apply to Articles 3 to 6 of this Directive.

A comparison of Article 3 and Article 4 reveals differences in relation to key features of these provisions. Primarily, the Article 3 exception can be used by only two classes of users: (i) **research organisations**, and (ii) **cultural heritage institutions**, that need to meet specific criteria established in Article 2 of the Directive. The Article 4 exception can be **used by any user**, irrespective of the purpose of TDM or the user's institutional status or commercial orientation.

The Article 3 exception is a much more substantive limitation to the exclusive rights of a right holder for two reasons. **Rights holders do not have the possibility of opposing TDM practices** when undertaken by research organisations or cultural heritage institutions for the purpose of scientific research, once the TDM activity is based on lawful access to the content. **Reproductions for TDM under Article 3 cannot be ruled out or restricted by contractual terms** ([38]), and Member States are not allowed to establish fair compensation mechanisms for rights holders ([39]).

In contrast, Article 4 is limited where the use is **'expressly reserved'** by the right holder 'in an appropriate manner'. Such reservations are commonly referred to as **'TDM opt-out'**, and the reservation of rights is an active decision and assertion made by the right holder to derogate from this general permission.

For both the Article 3 and Article 4 exceptions, **lawful access** to works or other subject matters is a necessary precondition. This means that "pirated" content cannot be used for TDM under either article. The condition of lawful access under Article 3 may be problematic if it allows contractual terms or TPMs to limit TDM practices (Strowel & Ducato, 2021). This is possibly less of an issue under Article 4, as TPMs and contractual terms may overlap with rights-

([38]) Directive (EU) 2019/790 Article 7(1): 'Any contractual provision contrary to the exceptions provided for in Articles 3 [...] shall be unenforceable'.

([39]) Ibid. Recital 17.

reservations systems, and so for commercial TDM the legal access and respecting opt-out requirements may be mutually reinforcing.

A further difference is that Article 3 creates a new relationship of rights and obligations between the TDM user and rights holders which stem from the fact that research organisations may need to – and are permitted to – **retain copies of works for scientific research purposes**, including the verification of research results ([40]). In Article 4, reproductions and extractions **may only be retained 'for as long as is necessary for the purposes of TDM'.** The Article 3 exception can be considered as wider due to the absence of potential right holder reservations and because it extends to the subsequent retention of reproductions. The beneficiary of the exception under Article 3 has an obligation to ensure the security of content reproduction storage, and rights holders have a right to apply measures to ensure the security and integrity of these storage networks.

Lastly, Articles 3 and 4 differ in terms of the actual rights which are subject to the exception. Both articles are exceptions to (i) reproduction rights of copyright owners, related rights holders, copyright-protected database owners, to (ii) extraction and reutilisation rights of *sui generis* database makers, and to (iii) reproduction and making available rights of online press publishers. The Article 4 exception is *de jure* wider than the Article 3 exception in that it also covers the rights of reproduction and alteration of owners of copyright in computer programs.

The differences in the respective scopes of Article 3 and Article 4 are summarised in the tables below.

CHARACTERISTIC OF EXCEPTION	ARTICLE 3	ARTICLE 4
Class of beneficiary	Research Organisations and Cultural Heritage Institutions	Anyone

([40]) Ibid. Recital 15.

CHARACTERISTIC OF EXCEPTION	ARTICLE 3	ARTICLE 4
Object of the exception	Works and subject matter to which the user has lawful access	Lawfully accessible works and subject matter
Scope of the exception	Scientific research	Any use
Contractual override	Not permitted	Permitted
Right holder Reservation	No	Opt-out possible
Obligations on beneficiary	Security of data storage	Respect of rights reservations
Retention of copies beyond TDM	Yes, for purposes of scientific research, including the verification of results	No

EXCEPTION TO EXCLUSIVE RIGHT	ARTICLE 3	ARTICLE 4
Reproduction of copyright-protected works *(Directive 2001/29/EC Art 2)*	X	X
Reproduction of related-rights subject matter *(Directive 2001/29/EC Art 2)*	X	X
Reproduction of copyright-protected databases *(Directive 96/9/EC Art 5(a))*	X	X

EXCEPTION TO EXCLUSIVE RIGHT	ARTICLE 3	ARTICLE 4
Extraction/Reutilisation of sui generis databases *(Directive 96/9/EC Art 7(1))*	X	X
Reproduction/Making available of press publications *(Directive 2019/790 Art 15(1))*	X	X
Reproduction/Alteration of computer programs *(Directive 2009/24/EC Art 4(1)(a)/(b))*		X

2.2.1.8 Requirements for a Valid Reservation of Rights ('Opt-out')

As noted above, TDM under CDSM Article 4 is permitted on the condition that the use of the content *"has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online"*. Further context is given by CDSM Recital 18, which states: *"In the case of content that has been made publicly available online, it should only be considered appropriate to reserve those rights by the use of machine-readable means, including metadata and terms and conditions of a website or a service. Other uses should not be affected by the reservation of rights for the purposes of text and data mining. In other cases, it can be appropriate to reserve the rights by other means, such as contractual agreements or a unilateral declaration."*

Based on the legal provisions, a valid opt-out under Article 4 (3) must meet three requirements: the reservation is (A) ***expressly* made** (B) **by the *right holder***, and (C) **in an *appropriate* manner**.

As highlighted in the analysis below, the interpretation of the exact requirements for a valid opt-out is an issue for which there are some uncertainties and even competing views, and this may eventually evolve through case law.

2.2.1.8.1 The 'Expressly' Requirement

There are different ways how the 'expressly' requirement may be interpreted and has been implemented in practice. A **strict interpretation** may demand an **explicit reference** to the act of 'text or data mining' or the enabling legislation (either CDSM Article 4 or the national transposition thereof). For example, the October 2023 public statement from French music rights Collecting Management Organisation (CMO) SACEM explicitly refers to opt-outs of TDM uses, makes an explicit reference to AI, and cites the relevant French copyright provisions on commercial TDM reservations ([41]).

More general declarations, such as broader contractual prohibitions on web scraping, have also been observed (as discussed in *Section 2.2.2*). Opt-out declarations may be made in various forms, such as metadata protocols, or website terms and conditions. In the LAION Case (discussed in *Section 2.3.1*), the Court of Hamburg (in obiter dictum statements) noted that the website terms and conditions likely met the conditions for a valid opt-out.

Overall discourses on TDM rights reservations often focus on the **Robots Exclusion Protocol** ('REP', also referred to as *'robots.txt'*) which is an instruction provided by websites to various 'robots' indicating general or specific restrictions and permissions to access and scrape its content. While REP is discussed in detail later in this report (see *Section 3.4.2.1*), it was not originally designed to address copyright management issues. As **a broad instruction to web scrapers,** it does not make explicit reference to any copyright protected work, legislative provision, or specific use case. However, current discourse in the industry has pointed towards REP being the benchmark for opt-out provisions.

A Dutch Case involving RSS (Really Simple Syndication) and alleged copyright infringement (HowardsHome Case) touched on the issue of TDM and valid opt-outs. The District Court of Amsterdam found that there was no valid reservation of rights to prevent TDM, because the REP instructions on the plaintiff's website excluded specific AI bots (including GPTBot, ChatGPT-User, CCBOT, and anthropic-ai), but not the bot of the defendant (who - for avoidance of doubt - was not an AI provider). This case does not provide a precedent on what

([41]) 'Sacem, in favour of virtuous, transparent, and fair AI, exercises its right to opt-out', Sacem, 12 October 2023 (accessed 14 March 2025).

might be considered a valid TDM opt-out, but it does suggest that a court may consider REP to be a valid rights reservation if the defendant's specific bot was included in the list of disallowed bots.

The view, based on the limited case law on TDM, as well as discourse in the industry and scholarship, is that the 'expressly' requirement of an opt-out is to be interpreted broadly, whereby a valid opt-out (i) need not refer to a specific work within a larger corpus of content, (ii) need not explicitly be aimed at TDM as a specific use case, but can be aimed at broader uses like web-scraping (which might not always qualify as TDM as it may not involve reproduction of works), (iii) need not reference any enabling legal provision, and (iv) need not be targeted to a specific potential TDM user.

The opposite view noted in literature stresses that the 'expressly' requirement should be interpreted strictly, and should preclude reservations which are not use-specific, content-specific, and can be found on the specific page of online content (Hamann, 2024).

2.2.1.8.2 The 'By the Right holder' Requirement

The second implicit requirement for a valid TDM opt-out is that it is **made 'by the right holder'**. While simple when there is only one 'right holder', the situation may be more complex when copyright protected works are assigned to commercial intermediaries for licensing purposes, and/or licensed to multiple parties holding copyright or neighbouring rights.

Different rights within the 'bundle' of exclusive rights under the InfoSoc Directive may be assigned to (or managed by) different parties. In the case of Collective Management Organisations (CMOs), **some CMOs only represent one specific exclusive right** on behalf of their members, while others represent multiple rights. For example, most European musical work CMOs (e.g., SACEM who publicly declared an opt-out as discussed above) represent both performance rights (right of communication to the public) and mechanical rights (reproduction rights). As the CDSM Article 4 TDM provision is an **exception to the right of reproduction**, it may be the party that manages this right of reproduction in particular (and not the right of communication to the public) that is the relevant party to make a valid opt-out.

An approach that has been discussed in some Member States is the possibility of managing TDM authorisation through a system of **'extended collective licensing'**. Under such a system, a CMO that adequately represents an entire category of rights holders may grant licences for a specific exclusive right, unless a right holder explicitly opts out or objects. Such a licensing mechanism is provided for in CDSM Directive Article 12, when applied 'within well-defined areas of use', where obtaining authorisations from right holders on an individual basis is particularly onerous and impractical.

The most notable example of such a discussion was a November 2024 proposal from the Spanish Government ([42]). The proposal argued that the massive use of copyright protected works to develop AI models is a well-defined use where individual authorisation is so impractical and onerous that individual licensing is unlikely. It was proposed that 'sufficiently representative' CMOs could extend their representation to rights holders that are not members of the CMO in a specific class of rights. Authorised CMOs would be allowed to administer extended collective licences for the reproduction and extraction of works in the context of text and data mining under CDSM Article 4. Rights holders would still maintain a right to object to their works being included in such a licence (i.e., a right to 'opt-out' of the extended management of their 'TDM opt-out rights'). The proposal was withdrawn at the end of January 2025. However, stakeholder interviews reveal that similar discussions were underway in other Member States.

In many cases, the right of reproduction is not assigned to a commercial intermediary but is **licenced directly by the author**. This may include licensing reproduction rights to another commercial intermediary such as a **content aggregator**. In the LAION Case (discussed in *Section 2.3.1*), the Court of Hamburg's obiter dicta comments suggested that reservation statements made by a licensee (through website terms and conditions) likely meet the 'by the right holder' requirement. However, when there is more than one licensee, or one non-exclusive licensee, the question may arise as to which party may make a valid TDM reservation.

([42]) 'Proyecto de Real Decreto por el que se regula la concesión de licencias colectivas ampliadas para la explotación masiva de obras y prestaciones protegidas por derechos de propiedad intelectual para el desarrollo de modelos de inteligencia artificial de uso general' (in Spanish), Spanish Ministry of Culture, 19 November 2024 (accessed 14 March 2025)..

Furthermore, some rights holders representatives have engaged in specific agreements to have **TDM opt-out explicitly delegated to them**. For example, German music CMO GEMA (who has started litigation against OpenAI in November 2024 - see *Section 2.3.1.3*) states that in 2022 its membership authorisation agreement was amended to explicitly have members grant GEMA the authority to declare opt-out ([43]).

In sum, the *'by the right holder'* requirement for a valid TDM opt-out may be very context specific, and guidance on when this requirement is met may emerge through national case law and/or industry practice. Furthermore, as exclusive rights under copyright are **territorial**, there might be different rights holders in different jurisdictions in relation to the same work, leading to an additional layer of complexity. Currently, key observations are that there are various legal principles through which a licensee or other representative may *potentially* make a TDM reservation on a right holder´s behalf, including: (i) an **explicit assignment** of the authority to make an opt-out, (iii) **existing delegated management** of the right of reproduction, or (iii) an **implied authority** to make an opt-out though the agency principles and duties of a licensee.

2.2.1.8.3 The 'Appropriate Means' Requirement

Both Article 4 and Recital 18 imply that TDM can be applied to '*content that has been made publicly available online'* and other cases. For '*content that has been made publicly available online'*, the 'appropriate' requirement for a valid opt-out is more specific and must be made through **'machine-readable means'**. Recital 18 gives **'metadata'** and **'terms and conditions of a website or service'** as examples of appropriate machine-readable means. For other cases (where content is not made available online), Recital 18 gives the examples of ***'contractual agreements or a unilateral declaration'*** as possibilities.

The reference to 'content that has been made available online' highlights that the same work may be made available on several locations. Using the example of REP (*robots.txt*) as a possible reservation mechanism that applies to copyright protected content on a particular website, multiple **location-specific reservations** may be necessary if the copyright owner wishes to broadly opt-out of TDM from all locations in which the content is legally accessible.

([43]) Suno AI and Open AI: GEMA sues for fair compensation, GEMA, 21 January 2025 (accessed 14 March 2025).

In the specific subset of cases where '*content that has been made publicly available online'*, the additional criterion applies that the 'appropriate means' must be 'machine-readable'. While Recital 18 suggests that a 'unilateral declaration' is an appropriate means for non-online cases, such a measure will only constitute a valid opt-out if it is also machine-readable. Furthermore, Recital 18 suggests that the terms and conditions of a website may be appropriate means that are machine-readable. Presumably, both unilateral declarations and website conditions are generally communicated in human-readable language. The question thus arises as to when the 'machine-readable' sub-criterion of the 'appropriate means' requirement for online cases applies. In particular, the **relationship between 'human-readable means' and 'machine-readable means'** is important to understand to determine when a measure constitutes a valid opt-out for online uses. In the LAION Case (see *Section 2.3.1*), the Court of Hamburg's *obiter dicta* statements suggest that reservations made in natural language (human-readable website terms and conditions) likely meet the 'machine-readable' criterion.

It is however noted that not all Member States (e.g., Italy) have explicitly included this 'machine-readable' criterion in national transposition of Article 4 and have only an 'appropriate means' requirement.

2.2.1.9 The Exception for Temporary Reproduction

In addition to the TDM-specific exceptions under CDSM Article 3 and Article 4, other limitations and exceptions to copyright and related rights within the EU legal framework may be relevant to GenAI technologies.

The exception outlined in Article 5(1) of the InfoSoc Directive for qualified acts of **'temporary reproduction'** could be relevant to certain activities associated with GenAI training. Unlike the exceptions and limitations set out in Articles 5(2) and 5(3), this provision is mandatory for Member States to implement in their national laws.

The CDSM Directive explicitly acknowledges that the temporary reproduction exception may apply in the context of TDM use. The CDSM Directive Recital 9 states that the temporary reproduction exception should continue to apply to TDM techniques that do not involve the making of copies beyond the scope of that exception, while Recital 18 notes that legal

uncertainty can arise as to whether TDM use meets all of the requirements for the temporary reproduction exception ([44]).

Information Society Directive, Article 5

1. Temporary acts of reproduction referred to in Article 2, which are transient or incidental and form an integral and essential part of a technological process, are exempt from the reproduction right outlined in Article 2, provided that these acts:

(a) enable a transmission in a network between third parties by an intermediary; or

(b) allow a lawful use of a work or other subject matter,

and have no independent economic significance.

The rationale and scope of this exception are further clarified in the InfoSoc Directive Recital 33, which highlights its application to acts, like browsing and caching, that facilitate efficient operation of transmission systems. This recital specifies that intermediaries must **neither modify the information nor interfere with the lawful use** of widely recognised industry-standard technologies for tracking data usage. While this exception was originally designed for simpler technological activities such as browsing or caching, it may also apply to more complex technologies, provided all required conditions are met.

For this exception to apply, Article 5(1) outlines five cumulative conditions, which have been interpreted by the CJEU in cases such as *Infopaq 1* (C-5/08), *FAPL/Murphy* (C-403/08 and C-429/08), *Infopaq 2* (C-302/10), *Meltwater* (C-360/13), and *Filmspeler* (C-527/15). The conditions are as follows:

- **The act must be temporary:** the storage and deletion of the reproduction must not depend on discretionary human intervention ([45]), although it is not required that the

([44]) Directive (EU) 2019/790 Recitals 9 and 18.

([45]) Case C-5/08, *Infopaq 1*, § 62.

process be entirely automatic in the sense that there is no human intervention in its activation and completion ([46]).

- **The act must be transient or incidental:** its duration should not exceed what is necessary to complete the technological process ([47]), or the reproduction must not exist independently or serve a purpose separate from the technological process ([48]).
- **The act must be an integral part of a technological process:** the technological process must be incapable of functioning effectively without the act of reproduction. This condition allows for scenarios where the process is initiated by a human ([49]).
- **The sole purpose of the process must be *either the* transmission of the work in a network between third parties by an intermediary, *or* a lawful use of the work:** lawful use may either be authorised by the right holder or not restricted by law. Mere reception of satellite broadcasts via decoders in private circles qualifies as "lawful use" ([50]), but not the reception via multimedia players of streaming broadcasts from illegal sources ([51]).
- **The act must have no independent economic significance:** the reproduction must not generate an additional economic advantage on its own ([52]).

Currently, no CJEU decisions directly address the application of the exception for temporary reproduction to GenAI training. However, in the LAION case (discussed in *Section 2.3.1*), the Hamburg Regional Court ruled that this exception does not apply to the reproduction of photographs for creating a filtered, cleaned, and semi-structured dataset for AI training. The court argued that such reproductions fell short of the 'temporary' and 'transient or incidental' conditions.

([46]) Case C-302/10, *Infopaq 2*, § 32 and Case C-360/13 *Meltwater*, § 32.

([47]) *Infopaq 1*, § 64.

([48]) *Meltwater*, § 43.

([49]) *Infopaq 2*, § 30-32.

([50]) Joined Cases C-403/08 and C-429/08 *FAPL/Murphy*, § 171-172.

([51]) Case C-527/15, *Filmspeler*, § 57.

([52]) *FAPL/Murphy*, § 175.

2.2.1.10 Lawful Use of Databases

Besides the exceptions for TDM in the CDSM Directive, the *sui generis* database right is subject to specific limitations outlined in the Database Directive. Particularly relevant to GenAI training activities is Article 8, which states that a 'lawful user' of a database protected under the *sui generis* database right **cannot be restricted from performing 'insubstantial' extractions and re-utilisations** 'for any purpose whatsoever'. This limitation on the database maker's exclusive right is framed as a 'right' granted to the lawful user and **cannot be overridden by contract.**

Database Directive, Article 8

Rights and obligations of lawful users

1. The maker of a database which is made available to the public in whatever manner may not prevent a lawful user of the database from extracting and/or re-utilizing insubstantial parts of its contents, evaluated qualitatively and/or quantitatively, for any purposes whatsoever. Where the lawful user is authorized to extract and/or re-utilize only part of the database, this paragraph shall apply only to that part.

2. A lawful user of a database which is made available to the public in whatever manner may not perform acts which conflict with normal exploitation of the database or unreasonably prejudice the legitimate interests of the maker of the database.

3. A lawful user of a database which is made available to the public in any manner may not cause prejudice to the holder of a copyright or related right in respect of the works or subject matter contained in the database.

Database Directive, Article 15

Binding nature of certain provisions

Any contractual provision contrary to Articles 6 (1) and 8 shall be null and void.

A **'lawful user'** has been defined by the CJEU as a user *"whose access to the contents of a database for the purpose of consultation results from the* ***direct or indirect consent*** *of the maker of the database"*([53]). This includes databases made publicly accessible online, where any member of the internet public qualifies as a lawful user. The rights of a lawful user are subject to the following limitations:

- They are limited to the extraction and/or re-utilisation of only 'insubstantial parts', prohibiting wholesale extraction or re-utilisation.
- The use must not conflict with the normal exploitation of the database, unreasonably harm the database maker's legitimate interests, or otherwise prejudice the holder of any copyright-protected works contained in the database.

The scope of a lawful user's rights remains to be fully clarified, as the database maker's exclusive rights cover the extraction and/or re-utilisation of the **'whole or a substantial part'** of the database's content ([54]). This suggests that the extraction and/or re-utilisation of **'insubstantial parts'** should generally be permissible. This provision may be relevant for web scraping, the automated extraction of information from publicly accessible websites and web pages. As discussed in *Section 3.1.2.2*, web scraping serves as a key source of data for training GenAI. It is important to note, however, that the CJEU, in Case C-30/14 (*Ryanair vs. PR Aviation*), clarified that the provisions on lawful users apply exclusively to databases that qualify for protection under either copyright or the *sui generis* database right. It does not extend to databases that, while meeting the Directive's definition of a 'database', lack either originality or the 'substantial investment' required for protection. The judgment leaves intact the possibility for owners of 'unprotected' databases to establish contractual conditions on the use of these databases, provided that these conditions are valid under national private laws (Borghi & Karapapa, 2015).

([53]) Case C-604/10, *Football Dataco*, § 58 (emphasis added).

([54]) Directive 96/9/EC, Article 7(1).

2.2.1.11 Exceptions to Exclusive Rights in Computer Programs

Since computer programs are generally protected as literary works, the exceptions and limitations applicable to the reproduction of copyright-protected works also extend to computer programs, including the exceptions for TDM. Additionally, other restricted acts in computer programs are subject to specific exceptions laid down in the Computer Programs Directive. Article 5(3) establishes an **exception for observing, studying or testing** a computer program, which cannot be overridden by contract and may hold relevance in the context of GenAI training.

Computer Programs Directive, Article 5 **Exceptions to the restricted acts** […] 3. The person having a right to use a copy of a computer program shall be entitled, without the authorisation of the right holder, to observe, study or test the functioning of the program in order to determine the ideas and principles which underlie any element of the program if he does so while performing any of the acts of loading, displaying, running, transmitting or storing the program which he is entitled to do.
Computer Programs Directive, Article 8 **Continued application of other legal provisions** […] Any contractual provisions contrary to the exceptions provided for in Article 5(2) and (3) shall be null and void.

As stated by the CJEU, the provision aims to ensure that *"the ideas and principles which underlie any element of a computer program are not protected by the owner of the copyright*

by means of a licensing agreement"([55]). Additionally, copyright is not violated when a lawful acquirer *"studied, observed and tested that program in order to reproduce its functionality in a second program"*([56]). Some commentators have suggested that **this exception could encompass TDM activities conducted for scientific research purposes** (Strowel & Ducato, 2021).

2.2.1.12 Overview of Relevant Exceptions and Limitations

The table below provides an overview of the exceptions and limitations available under EU law that are relevant in the context of GenAI training.

<table>
<tr><th rowspan="2">Exclusive Rights</th><th colspan="5">Activities</th></tr>
<tr><th>TDM for scientific research purpose (Art. 3 CDSM)</th><th>TDM for any purpose (Art. 4 CDMS)</th><th>Acts of temporary reproduction (Art. 5(1) InfoSoc)</th><th>Insubstantial extraction of database content (Art. 8 Database Directive)</th><th>Observation, study or testing of computer program (Art. 5(3) Computer Programs Directive)</th></tr>
<tr><td>Reproduction of
• Works
• Fixations of performances
• Phonograms
• Films
• Fixations of broadcasts</td><td rowspan="3">Permitted to
• Research organisations
• Having lawful access
• Applying appropriate security measures
Not overridable by contract</td><td rowspan="3">Permitted unless the right has been
• Expressly reserved
• By the right holder
• In an adequate manner (machine-readable format)</td><td>Permitted if:
• Temporary
• Transient or incidental
• Integral part of a technological process
• It enables transmission or lawful use
• Lacks independent economic significance</td><td>n.a.</td><td rowspan="3">n.a</td></tr>
<tr><td>Reproduction and making available of Press publications</td></tr>
<tr><td>Extraction and/or re-utilisation of Databases content</td><td>n.a.</td><td>Permitted to lawful users. Not overridable by contract.</td></tr>
</table>

([55]) Case C-406/10, *SAS Institute v World Programming*, § 51.

([56]) Ibid. § 61.

Reproduction, translation/ adaptation/alteration of **Computer programs**	*n.a.*			*n.a.*	Permitted to persons having a right to use the program. Not overridable by contract

Table 2.2.2-1: Overview of the exceptions and limitations available under EU law that are relevant in the context of GenAI training.

2.2.2 Contractual Restrictions on Web Scraping

Web scraping, the automated extraction of information from publicly accessible websites and web pages, is a key source of data for training GenAI models (see *Section 3.1.2.2*). Web scraping often depends on web crawling, which involves locating and identifying relevant information online ([57]). Website owners can use the REP to instruct web crawlers not to access their site or specific parts of it (see *Section 3.4.2.1*).

Web scraping can face restrictions under copyright and related rights and **through contractual terms**. The interaction between these two forms of protection becomes particularly intricate when the website in question qualifies as a 'database' under the Database Directive (see the definition in *Section 2.2.1.2*). As discussed in *Section 2.2.1.10*, the 'rights' granted to lawful database users that cannot be overridden by contract apply only to databases qualifying for protection under copyright or *sui generis* database right. In principle leaving **owners of 'unprotected' databases at liberty to establish contractual conditions** on the use of such databases.

Many websites include provisions in their Terms of Service (ToS) aimed at ruling out web scraping. These restrictions may be phrased as general conditions of use (e.g., "You are not permitted to use this website other than for private, non-commercial purposes") or as provisions that specifically prohibit automated extraction, scraping, reproduction or use of data without permission (e.g., "You may not use automated systems or software to download content or to extract data from this website"). A question thus arises regarding the extent to which these ToS can be construed as enforceable contracts which are binding on the scraper.

([57]) See the *Glossary* for the exact definitions of web scraping and crawling and the difference between them.

A contractual agreement is typically formed when ToS are presented as 'clickwrap'' requiring users to click 'I agree' before gaining access to a website or parts of it. If this is not the case, an alternative form of contractual agreement must be established. National courts have tended to dismiss breach of contract claims against web scrapers when there is insufficient evidence that users were aware of (or agreed to) the ToS, including when the ToS link is not prominently displayed on the website. In such cases, the absence of actual or constructive knowledge of the ToS may shield the web scraper from liability for breach of contract. When the ToS link is clearly visible courts may enforce the terms because constructive knowledge on the part of the scraper can be established (Pagallo and Ciani Sciolla 2023).

Ultimately, the enforceability of restrictions on web scraping remains **heavily context-dependent**. Some stakeholders noted that **current voluntary codes** (e.g., the developing GPAI **Code of Practice**) **do not extend to upstream dataset providers or 'in-house' AI development**. This may enable some GenAI developers to **train AI models internally without the transparency obligation** that may otherwise apply to external-facing services. **The current soft-law framework is still developing**, leaving significant portions of AI training activity unregulated. That said, it is noteworthy that the Court of Hamburg in the LAION Case (discussed in *Section 2.3.1.1*) indicated that ToS prohibiting web-scraping can be interpreted as a valid form of 'opt-out' from all-purpose TDM under Article 4 of the CDSM Directive.

2.2.3 The Artificial Intelligence Act

The EU was the first jurisdiction to adopt a **comprehensive legal framework** for Artificial Intelligence, through the legal instrument *Regulation (EU) 2024/1689* (the 'AI Act') ([58]).

([58]) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). OJ L, 2024/1689, 12.7.2024.

2.2.3.1 Key Definitions

Article 3 of the AI Act sets out specific definitions which are important for understanding how the Act defines the different actors within the digital ecosystem, each having specific legal obligations.

Under the AI Act definition of **'General-Purpose AI' (GPAI) ([59])** there are models which have a variety of uses and are trained on large datasets, which may or may not have been acquired using TDM practices. The legal concept of 'GPAI models' thus *generally* corresponds to the technical term **'foundation models'** (Madiega 2023). Several of the important provisions of the AI Act take the form of affirmative obligations on the providers of general-purpose models. The term 'AI system' ([60]) is then a wider term referring to systems based or not on general-purpose AI models. These systems are then put on the market by 'providers' ([61]) (and may be further adapted into subsequent systems put on the market by 'downstream providers' ([62])). AI systems are then utilised by 'deployers' ([63]), effectively those who use and may adapt these systems in a specific use case. Depending on the nature of the use case, the AI system may be meant to interact with a natural person who is not necessarily the deployer. For example, where an AI system is used internally in an undertaking within a business process, the company is the deployer, but there may be no specific natural person outside of the undertaking that directly interacts with the system. On the other hand, GenAI systems may be used by natural persons of the public to generate content, and such natural persons may be described as **an end-user**.

([59]) **'General-Purpose AI Model'** means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market.

([60]) **'AI system'** means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

([61]) **'Provider'** means a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system, or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge.

([62]) **'Downstream provider'** means a provider of an AI system, including a general-purpose AI system, which integrates an AI model, regardless of whether the AI model is provided by themselves and vertically integrated or provided by another entity based on contractual relations.

([63]) **'Deployer'** means a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity.

Figure 2.2.3-1: Key parties in the General Purpose AI Deployment Chain

While the AI Act does not use or define the term **'Generative AI' ('GenAI')**, this term is understood to refer to the specific subset of AI systems which are designed and used for the **creation of machine-generated outputs**. Recital 99 of the AI Act states that *"Large generative AI models are a typical example for a general-purpose AI model, given that they allow for flexible generation of content, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks".* Thus **'GenAI models' are understood to be a subset of General-Purpose AI models** if they fulfil the definition in the AI Act as stated above.

The **categorisation between "provider" and "deployer" is important because it determines the allocation of legal responsibilities** under the AI Act and potential liabilities..

2.2.3.2 Intersections between the AI Act and EU Copyright Law

The AI Act establishes a comprehensive legal framework governing market introduction and deployment of AI systems and general-purpose AI models. It is meant to address risks to health, safety and fundamental rights, while promoting innovation and uptake of AI. This report, however, focuses exclusively on the provisions relevant to copyright and GenAI in relation to general purpose AI models. The AI Act intersects with the broad framework for EU copyright law in three anchor points: (i) Policy to **respect for EU copyright Law** (including provisions on TDM rights reservations), (ii) **Transparency requirements**, and (iii) **Extra-territorial application** of certain provisions.

The AI Act sets out transparency provisions in relation to AI inputs, AI models and systems themselves, and AI outputs. Transparency obligations for inputs and outputs are analysed in this Report, though it is transparency obligations regarding inputs that has the most direct relationship with EU copyright law. By 'extraterritorial application', reference is being made to

Article 2(a)(1) of the AI Act, which stipulates that the legal obligations of the Act apply to AI providers '*irrespective of whether those providers are established or located within the Union or in a third country*'.

With these anchor points in mind, the following sub-sections summarise the key provisions of the AI Act that relate to copyright issues in GenAI.

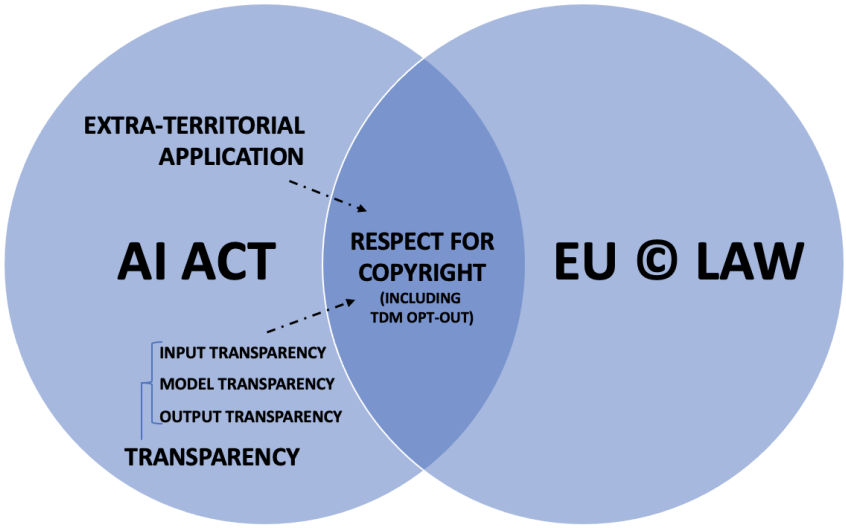


Figure 2.2.3-2: Intersections between AI Act and EU copyright law

2.2.3.3 Obligations Regarding AI Inputs

The relationship between copyright law and the development of AI models is best summarised by Recital 104 of the AI Act, which states:

> *"General-purpose AI models, in particular large generative AI models, capable of generating text, images, and other content, present unique innovation opportunities but also challenges to artists, authors, and other creators and the way their creative content is created, distributed, used and consumed. The development and training of such models require access to vast amounts of text, images, videos and other data. Text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content, which may be protected by copyright and related rights. Any use of copyright protected content requires the authorisation of the*

rightsholder concerned unless relevant copyright exceptions and limitations apply. Directive (EU) 2019/790 introduced exceptions and limitations allowing reproductions and extractions of works or other subject matter, for the purpose of text and data mining, under certain conditions. Under these rules, rightsholders may choose to reserve their rights over their works or other subject matter to prevent text and data mining, unless this is done for the purposes of scientific research. Where the rights to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorisation from rightsholders if they want to carry out text and data mining over such works."

The **policy objective of transparency** is reflected in several provisions of the AI Act, including on the training process and data used as inputs in developing general-purpose AI models. The relationship between copyright law and transparency occurs at two distinct levels: (i) the **policies used by AI developers** to comply with copyright law, including rights holders TDM opt-outs, and (ii) the **details of the actual content** used to train models.

With respect to the actual training data used, there is a tension between this objective of transparency and the fact that information on training data can potentially constitute proprietary trade secrets or confidential business information ([64]). In light of this tension, the obligation for transparency in training data may be met through publishing a **'sufficiently detailed summary'** of the content used for training the general-purpose AI model. Such a summary should be **publicly available** and does **not necessitate a work-by-work assessment** in terms of copyright compliance ([65]).

The disclosed information on training data should be comprehensive enough to enable copyright holders to enforce their rights ([66]). The transparency obligation of the AI provider is thus conceptually connected to the right holder's legal entitlement to exclusive rights. The Recitals of the AI Act suggest that the level of detail in such training data summaries might be met by *"…listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used".*

([64]) This tension is recognised by Regulation (EU) 2024/1689 Recital 107.

([65]) Ibid. Recital 108.

([66]) Ibid. Recital 107.

Article 53 of the AI Act sets out specific obligations for providers of general-purpose AI models. The objectives of transparency in copyright-compliance are integrated into the obligations under this Article. In particular, Articles 53(1)(c) and 53(1)(d) set out certain obligations, with specific reference being made to the TDM opt-out mechanism of CDSM Article 4.

The disclosure and transparency principles of Article 53 apply once a model is placed onto the EU market, '**regardless of the jurisdiction** in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place' ([67]). This ensures that model providers do not gain a competitive advantage by engaging in data collection or model training in jurisdictions that have lower copyright standards than those in the EU.

It is important to note that there is debate in the legal scholarship about how Article 53(1)(c) of the AI Act should be interpreted and applied, and what requirements regarding TDM would apply for models trained outside of the EU.

Perspectives range from a 'minimalist' reading where extra-EU TDM is an act not done pursuant to the CDSM directive due to the territorial limitations of copyright law, an 'intermediate' proposal where the obligations should apply depending on whether scraped content (which formed the basis of training data) was hosted on servers in the EU, to a 'maximalist' approach which would directly require extra-EU trained models to comply with the TDM provisions of the CDSM in order to be legally placed on the EU market (Peukert, 2024; Rosati, 2024; Stieper & Denga, 2024).

Article 53 (Obligations for providers of general-purpose AI models)

1. Providers of general-purpose AI models shall:

…

(c) put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790;

([67]) Ibid. Recital 106.

> (d) draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office.
>
> ….
>
> 4. Providers of general-purpose AI models may rely on codes of practice within the meaning of Article 56 to demonstrate compliance with the obligations set out in paragraph 1 of this Article, until a harmonised standard is published. Compliance with European harmonised standards grants providers the presumption of conformity to the extent that those standards cover those obligations. Providers of general-purpose AI models who do not adhere to an approved code of practice or do not comply with a European harmonised standard shall demonstrate alternative adequate means of compliance for assessment by the Commission.

There appear to be different paths for a general-purpose AI model provider to demonstrate compliance with Article 53. In particular, Article 53(4) allows such providers to either (a) rely on **codes of practice**, (b) rely on **a harmonised standard ([68])** once developed, or (c) develop their **own alternative adequate means** to demonstrate compliance ([69]).

Under Article 56(3), the development of the GPAI Code of Practice is facilitated by the AI Office in a transparent and inclusive process involving more than a thousand stakeholders ranging from GPAI providers, downstream providers and business associations, rights holders, civil society, academia etc. In terms of copyright related measures, the Code should include a dedicated section to operationalise the obligations for providers to put in place a policy to comply with Union copyright law. Commentators have described such codes of

([68]) A harmonised standard is a European standard developed by a recognised European Standards Organisation (CEN, CENELEC, or ETSI) following a request from the European Commission (Regulation (EU) No 1025/2012 on European standardisation).

([69]) Regulation (EU) 2024/1689 Article 56 provides for the development of 'Codes of Practice' whose development is to be facilitated by the AI Office. It includes the issue of 'the adequate level of detail for the summary about the content used for training'. As of the date of this report, the process for the development of such Code of Practice is underway.

practice as sources of soft law and 'meta-regulatory tools' under the AI Act (Bygrave & Schmidt, 2024).

In parallel to this process, the AI Office is developing a **template for the sufficiently detailed summary of training data** that GPAI model providers are required to make public. In the long term, the European Commission is also expected to mandate European standardisation organisations to develop harmonised standards on the obligations of GPAI providers (Art. 40(2)).

Furthermore, while Article 53 sets out general requirements of information disclosure applicable to all providers of general-purpose AI models, Article 53(2) provides that some elements of this Article 53 obligation do not apply to providers of AI models that are released under a **free and open-source licence**. The underlying premise is that certain technical information is already disclosed through the mechanism and definition of open-source licensing. Nevertheless, the specific requirement to disclose a summary of training data used still applies no matter what licensing approach the AI provider adopts (i.e., the training data summary requirement applies even if a model is licensed on an open-source basis). The AI Act suggests that this is necessary because *"...the release of general-purpose AI models under free and open-source licence does not necessarily reveal substantial information on the data set used for the training or fine-tuning of the model and on how compliance of copyright law was thereby ensured..."*. However, it is noted that other frameworks may exist and definitions for 'Open-Source AI', such as the definition from the Open Source Initiative (see *Annex V*), Linux Foundation's proposed Model Openness Framework (White et al., 2024), which might set out relatively high standards for the disclosure of training data information. Furthermore, it is noted that there is **active debate over the implications of 'open source' definitions** and the different obligations that apply in this context, including the role played by partially-open models (Liesenfeld & Dingemanse, 2024).

It is important to note that while general-purpose AI model providers are required to put in place policies to comply with CDSM Article 4 rights reservations, this does not mean that AI providers themselves are delegated the task of developing protocols and standards for rights holders TMD opt-out. The Article 53 obligation rather states that the policies put in place are meant to **'identify and comply with' rights reservations**, implying that the obligation is on the model provider to identify the reservations made in the form appropriately chosen by the right holder (once this form meets the legal requirements for a valid opt-out under Article 4).

Article 53 also states that the AI provider's policy for identifying and complying with rights reservations includes using **'state-of-the-art technologies'**.

In the LAION Case (discussed in), the Court of Hamburg's *obiter dicta* statements suggest that this requirement for AI providers' to use 'state-of-the-art technologies' to comply with rights reservations supports the argument that natural language reservations should be understood as 'machine-readable'.

2.2.3.4 Obligations Regarding AI Outputs

The concept of transparency also relates to the ability of end-users (natural persons that interact with AI systems) to be **aware that they are interacting with AI systems**. As 'GenAI systems' are a subset of AI systems, the notion of transparency is also extended to the ability of natural persons to discern AI generated or manipulated content from human-generated content. Such transparency obligations (regarding both transparency for end-users and for machine-generated content identification) are set out in Article 50 of the AI Act.

Article 50 distinguishes between three types of content that may be outputted by AI systems: (i) general 'synthetic content', (ii) 'deepfakes' ([70]) and (iii) AI-generated or manipulated text that is published with the purpose of informing the public on matters of public interest ([71]). While **'synthetic content'** refers to any AI generated or manipulated content (whether text, code, image, audio, or audiovisual), the term **'deepfakes'** refers to a class of machine-generated content which is essentially a subset of 'synthetic content'. To be qualified as 'deepfake' content needs to meet two criteria, namely (a) the resemblance of the content with actually existing subject matter, and (b) the potential of such content to falsely appear authentic or truthful to a person.

([70]) Regulation (EU) 2024/1689 Article 3(60). The AI Act defines the term 'deepfake' to mean *"AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful."*

([71]) The AI Act distinguishes deepfakes that are image, audio, or video content which requirements are defined in Regulation (EU) 2024/1689 Article 50(4) subparagraph from AI-generated or manipulated text content published with the purpose of informing the public on matters of public interest in Regulation (EU) 2024/1689 Article 50(4) subparagraph 2.

Article 50(2) requires that users be informed that they are interacting with AI systems. Regarding general synthetic content, Article 50(2) places an obligation on AI systems providers to *"ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated"*. These technical solutions are required to be 'effective, interoperable, robust and reliable' as far as technically feasible ([72]). This requirement is motivated by the need to promote integrity and trust in the information ecosystem and mitigate the social risks of misinformation and deception. Furthermore, identification of GenAI output is also useful in the copyright context, given the prevailing view that AI-generated content is not copyright eligible, and debates over the threshold of human involvement necessary for such content to attract protection (Leistner & Jussen, 2025).

Regarding 'deepfakes', Article 50(4) places an obligation on AI systems deployers (specifically those which generate or manipulate deepfakes) to visibly **disclose that the content has been artificially generated or manipulated**. This obligation is completed by two special cases. First, the transparency obligation may be limited where necessary because of the artistic nature of the generated content. Second, text generated for 'informing the public on matters of public interest', which may include some journalistic content, must be disclosed as artificially generated, unless the content has undergone a process of human review or editorial control.

2.2.4 Boundaries between TDM and AI Training

The definition of TDM under the CDSM Directive requires the use to be '*an automated analytical technique*', which is '*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*'. Recital 8 provides broader interpretive context, suggesting that TDM is undertaken with a **'*view to gaining new knowledge and discovering new trends possible*'**.

However, some empirical research has suggested that GenAI systems and LLMs in particular may be able to generate large quantities of content which may amount to verbatim reproductions of works included in their training datasets ([73]) (Carlini et al., 2023). In defining

([72]) Including considerations of the specific nature of different types of content, the costs of implementing measures, and the technical state of the art.

([73]) See the discussion of training-data memorisation in Section 3.2.

the scope of what constitutes TDM, a distinction needs to be made between the different stages in the data processing chain. The general view in the legal scholarship literature is that **TDM can include the process of AI training but not output generation** (Dusollier, 2020; Mezei, 2024; Novelli et al., 2024; Rosati, 2021). This scholarship also highlights that TDM may apply to GenAI model training even though the CDSM Directive was fully drafted and negotiated before the current GenAI developments, pointing specifically to **Article 51(1)(c) of the AI Act** which explicitly requires that general-purpose AI models must comply with CDSM TDM reservations.

The logic is that if an AI model provider is explicitly required to respect Article 4 opt-outs, then the model provider is inherently a potential beneficiary of the Article 4 exception - i.e., a model provider has the capacity to carry out TDM. This is further supported by Recital 105 of the AI Act which states that *"The development and training of such models require access to vast amounts of text, images, videos and other data. Text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content, which may be protected by copyright and related rights.",* and further "*Where the rights to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorisation from rightsholders if they want to carry out text and data mining over such works."*

However, there still remains a view which argues that AI training does not constitute TDM within the meaning of Articles 3 and 4 of the CDSM, based on the underlying technology used and its capacity to process both semantic and syntactic information (Dornis & Stober, 2024). This controversy was also acknowledged by the Court of Hamburg in the LAION Case (see *Section 2.3.1.1*). However, the Hamburg Court opined that the teleological reduction of the TDM exceptions should be rejected and that the TDM exemption should apply to AI training. In particular, the Court suggested that AI training is not distinct from other forms of TDM (and that the view that distinguishes information 'hidden' in data from creative expression is unclear), that the potential for AI output to compete with trained data is irrelevant, that there is no contradictory legislative intention, and that the use complies with the three-step-test under copyright law.

The relationship between the definition of 'TDM' and GenAI model training is also important for understanding input transparency measures. AI Act Recital 107 suggests that the Article 53(1)(d) requirement for GPAI providers to publish a 'sufficiently detailed summary' is linked

to the objective of facilitating the ability of copyright holders to exercise and enforce their rights. However, a key challenge is that there exists TDM users which may act as an intermediary between the rights holders and the AI provider - specifically, dataset developers that are themselves not AI model developers. Such independent dataset developers do not fall within the scope of Article 53's data transparency and copyright policy obligations (as they are not GPAI model providers). Industry best practices are currently evolving to address this issue.

A further challenge also lies in the fact that the Article 53 obligation applies only to 'providers', as per the relevant definition set out in Article 3 of the AI Act. Thus, privately developed models used within the confines of a private environment (i.e., that are not made available to the public or put into service under a commercial name) may not be bound by the disclosure requirements of the AI Act. In such cases, rights holders may lack a mechanism to verify if their content was used (at least under Article 53) to manage and enforce their rights, including compliance with CDSM Article 4 opt-outs.

2.3 Ongoing Legal Challenges and Litigation

The use of copyright protected works as training inputs in GenAI model development has resulted in much public debate and several legal challenges. So far, lawsuits have mainly been initiated in the USA, but cases have also been brought forward in China, Canada, UK, India and the EU. These legal challenges **generally focus on the issue of whether limitations and exceptions to copyright apply to AI training**, and the circumstances under which restrictions stipulated by rights holders must be respected.

2.3.1 Litigation in the European Union

To date, copyright-related **litigations in the EU have been relatively limited compared to the USA**. National cases, particularly in Italy ([74]) and in the Czech Republic ([75]), have addressed the protectability of AI-generated output, generally concluding that synthetic content is not eligible for copyright protection per se, unless a distinct and sufficiently creative human contribution can be identified in the generation process. Regarding cases related to copyright infringement in the context of GenAI training, **four cases have been filed to date: three in Germany and one in France.**

2.3.1.1 Kneschke vs. LAION (Germany)

The first case of litigation in the EU between a right holder and an AI ecosystem actor the German case *Kneschke vs. LAION ('LAION Case')* ([76]). While the Hamburg Regional Court arrived at a final decision based on the **scientific TDM exception** (CDSM Article 3), the Court in its judgement made **several *obiter dicta* remarks which are instructive** for understanding the breadth of possible issues that may arise in applying TDM exceptions in practice.

([74]) Corte di Cassazione (Cassation Court), ordinanza n. 1107 (16 January 2023).

([75]) Městský soud v Praze (Municipal Court of Prague), 10 C 13/2023 (11 October 2023).

([76]) *LG Hamburg, Urteil vom 27.09.2024 - 310 O 227/23*.

LAION is a German non-profit organisation which is active in the AI training dataset market. It offers a database of over five billion image-text pairs which matches hyperlinks of images publicly available on the internet to text information about the image's content. This dataset is based on data from *Common Crawl* - a comprehensive and monthly-updated web archive of publicly available online content (see *Section 3.1.2.1.2*). LAION extracted image URLs from the Common Crawl dataset, downloaded these images, undertook checks to verify the image descriptions, filtered out images where the text content was insufficiently matched, and (re)extracted the location (URL) and description information to create a new dataset. In essence, the LAION dataset is a filtered, cleaned, and semi-structured sub-set of the Common Crawl Dataset, optimised for training GenAI image systems.

The claimant, a photographer, licensed a copyright protected photograph to a stock photo agency. This photo was then downloaded, analysed, and included in the LAION dataset, without the photographer's permission. The photo agency's website contained terms and conditions in plain language which prohibited web scraping([77]).

The dispute centred on the application of various provisions of German Copyright Law (Urheberrechtsgesetz, or 'UrhG'), specifically the articles that implement the TDM provisions of the CDSM Directive. LAION's activity as a reproduction under UrhG §16 was not disputed, the issue was whether any specific copyright limitation applied to such activity. Three possible copyright limitations were raised in this case (i) temporary reproduction, (ii) commercial TDM, and (iii) scientific TDM. The court's findings in relation to these limitations are summarised below.

2.3.1.1.1 Temporary Reproduction Exception

The **temporary reproduction exception** of InfoSoc Article 5(1) is implemented in German Copyright Law in UrhG §44a. The Court found that **LAION's use did not meet the specific conditions** required for this exception to apply. The Court considered that the reproduction was not incidental as *'the deletion was not carried out "independently of the user" but rather*

([77]) Specifically, the website's terms stated "RESTRICTIONS...YOU MAY NOT: (...) 18. Use automated programs, applets, bots or the like to access the ...com website or any content thereon for any purpose, including, by way of example only, downloading Content, indexing, scraping or caching any content on the website."

due to the defendant's conscious programming of the analysis process' ([78]). In addition, since the images were downloaded to be analysed using a specific software, their downloading is not *'just a process that accompanies the analysis being carried out, but a conscious and actively controlled acquisition process that precedes the analysis'* ([79]). The Court considered that LAION was therefore unable to rely on the temporary reproduction exception.

2.3.1.1.2 Definition of TDM

First, the Court determined if **LAION's activities fell within the definition of TDM**. The act of downloading works in order to analyse them by comparing them with pre-existing descriptions (i.e., with the descriptive information originally found on the Common Crawl database) falls within the definition of TDM, as this analysis was done in order to extract information about correlations ([80]). The Court suggested that the filtering of datasets is not a prerequisite for the definition of TDM, which may be relevant to the overall understanding of how TDM provisions may be interpreted in the context of data value chains ([81]).

The Court rejected arguments for a *teleological reduction of the limitation provision,* which propose that the scope of the TDM exception should be limited in light of the intended purpose of a dataset produced through TDM. On this basis, such arguments considered that TDM should not be permissible if the purpose of the produced dataset is to train an AI system where such training itself does not fall within the scope of TDM. Therefore, the TDM exception should not extend to actual AI training ([82]). The Court left unanswered the contended question as to whether the definition of TDM extends to actual AI training.

2.3.1.1.3 Commercial TDM exception

([78]) LAION Case, para 63. See supra sec. 2.2.2.2.2 on the exception for temporary reproduction.

([79]) LAION Case, para 66.

([80]) LAION Case, para 73.

([81]) LAION Case, para 74.

([82]) As argued in the study of Dornis and Stober (2024) cited in the decision.

German Copyright Law implements the general (non-scientific but commercial) TDM exception of CDSM Article 4 in UrhG §44b. The court ultimately ruled that LAION's activities were permissible under the scientific TDM provision, and there was no need to assess whether the commercial TDM exception also applied. Nevertheless, the Court suggested in its *obiter dictum* statements that it appeared doubtful that LAION could rely on this exception.

With regard to TDM opt-outs, if the Court did not make a definitive determination on this matter, it suggested that the photo agency's **website terms probably served as a valid 'effectively declared reservation of use.'** ([83]) The reservation declaration need not be made by the author himself, as he is entitled to rely on a reservation made by the photo agency in the capacity as a licensee ([84]). Furthermore, the Court suggested that a reservation of use written solely in natural language may meet the 'machine-readable' requirement ([85]).

2.3.1.1.4 Scientific research TDM exception

German Copyright Law implements the scientific research TDM exception of CDSM Article 3 in UrhG §60d. The Court ultimately ruled that this copyright exception was applicable and LAION's activities were permissible as TDM for scientific purposes. LAION's activities **constituted scientific research** as they were done in the pursuit of new knowledge. To meet this criterion, it was not necessary to have actually gained any new knowledge, it was sufficient that the activities were **aimed gaining knowledge at a later stage** ([86]). Furthermore, LAION qualified for the scientific TDM exception as their research activities are non-commercial, evidenced by the fact that the resulting dataset is made publicly available for free, irrespective of how the organisation is financed or staffed ([87]).

([83]) LAION Case, para 93.

([84]) LAION Case, para 96.

([85]) LAION Case, para 102.

([86]) LAION Case, para 114.

([87]) LAION Case, para 119.

2.3.1.2 Possible Lessons from LAION: Requirements for Valid Reservation

The Court's *obiter dicta* could be relevant to further the understanding of the requirements for a TDM reservation of rights ([88]). However, it is important to note, that it is premature to draw any conclusions from the statements made in a first instance decision that was subsequently appealed. Additionally, such statements were made on the basis of the provisions implementing the CDSM Directive in German law.

2.3.1.2.1 The 'Expressly' Requirement

The Court of Hamburg noted that the **website terms and conditions *likely* met the requirement for a valid opt-out**. In this case, the relevant terms stated *"YOU MAY NOT: … Use automated programs, applets, bots or the like to access the ….com website or any content thereon for any purpose, including, by way of example only, downloading content, indexing, scraping, or caching any content on the website."* ([89]) These terms and conditions do not explicitly reference the term 'text and data mining,' nor do they cite any supporting statutory provision. The Court observed that this reservation was **formulated with sufficient clarity**, and that it was made explicitly (not implied), and with precision to unequivocally cover a particular content and specific use. Furthermore, the Court opined that this reservation as applied to all uploaded works on a website was *likely* valid.

The Court also suggested that *"For the legal effect of the declaration it is not a requirement that the declaration be made with specific reference to a particular legal provision"*. As such, the Court considered that a valid rights reservation did not need to make explicit reference to a specific legal provision that enables the TDM exception and the opt-out possibility. This is an interesting observation, as it may mean that rights holders that already have some reservation in place, may not have to change or re-declare their opt-out, even if that reservation mechanism was declared without explicitly having TDM and AI training in mind.

Furthermore, the Court states that "*Even a reservation explicitly declared for all works uploaded on a website is clearly definable in its scope and content and is therefore explicitly*

([88]) The requirements are discussed above in *Sec. 2.2.2.2*.

([89]) LAION Case, para 9.

declared"([90]). Based on this view a reservation made for the entire body of content on a website may constitute an expressly made reservation. Here again, this is an interesting observation, as an individual reservation may not be required for every single work contained on a website. The overall implication of these comments is that the 'expressly' requirement of a valid CDSM Article 4 TDM rights reservation **might be interpreted broadly**.

2.3.1.2.2 The 'By the Right holder' Requirement

The Court stated that *"...it is not only the declarations of the original copyright holder that should be considered, but also those of subsequent rights holders, whether they are legal successors or holders of derivative rights from the original author."* The **original author may rely on the reservation made by the stock photo website**, as the stock photo agency is the right holder of the specific photo hosted on their website, and the exploitation took place through the agency. The Court also noted that there was no claim of a conflicting agreement between the agency and the photographer.

The standard Contributor Agreement of Bigstock, the stock photo agency used by the claimant ([91]) required only a *non-exclusive* license from its contributors ([92]). Thus, it may be possible for rights holders to submit the same photo to multiple agencies, which may in turn have different policies on web scraping.

A conceptual distinction may be made between a reservation of rights for a work generally, and reservation of rights for the specific copy of a work as hosted on a specific website (location-based reservation). While a specific licensee may be able to make a rights reservation for the specific digital asset that they are the custodians of, it does not follow that they may make a universal reservation of rights which applies to all copies of a work in all locations.

Overall, it appears that **duly authorised representatives of a copyright owner** (which according to the LAION Case may include specific licensees as 'rights holders') **may make**

([90]) LAION Case, para 99.

([91]) What is allowed when creating AI training data? First day of negotiations in the case against LAION e.V., Alltag eines foto produzenten, 12 July 2024 (accessed 15 March 2025).

([92]) Bigstock Contributor Agreement, Bigstock (accessed 15 March 2025).

valid TDM reservations. However, these reservations may be copy/location-specific. It follows that a copyright owner who wishes to universally opt-out of TDM may have to coordinate the expression of this reservation with the various licensees through which its content is made available to the public. A location-specific reservation, for example through *robots.txt* on a specific website, does not guarantee a valid reservation of rights for the licensed copies of the same content hosted on other websites.

The comments from the Court in the LAION Case suggest that the ability of a licensee to declare a valid reservation on behalf of a copyright owner is **derived from the relationship and duties of agency**. The Court also noted that there was no claim of a conflicting agreement between the agency and the photographer, regarding the reservation of rights. This indicates that while the default position is that the licensee may express a specific reservation on the owner's behalf, it could be modified by a contract regarding TDM reservation capacity.

2.3.1.2.3 The 'Appropriate Means' Requirement

In the LAION Case, the Court suggested that it *"...tends to consider a reservation of use expressed solely in "natural language" as "machine-understandable""*. This suggests that **natural language terms and conditions of a website may meet the 'machine-readable' criterion** and thus the 'appropriate means' requirement for online content.

The Court also noted the obligations under Article 53(1)(c) of the AI Act, which requires general purpose AI model providers to *"...put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790"*. The Court suggests that this provision's reference to **'state-of-the-art technologies'** used by AI developers who engage in TDM may include AI-driven natural language processing capabilities, **which are able to read natural language opt-outs like website terms and conditions**. Following this reasoning, an open question is whether the standard for machine-readability might be different for TDM users who are not AI developers, and are thus not bound by the obligations of Article 53(1)(c) of the AI Act.

Within a conservative interpretation of the 'expressly' requirement, there is a view which supports a narrow interpretation of 'machine-readable', suggesting that natural language

reservations should not be considered to be machine-readable (Hamann, 2024). This interpretation claims partial support by observing the strict definition of 'machine-readable' in the *Open Data and Re-Use of Public Sector Information Directive* ('Open Data Directive'), which defines that *"'machine-readable format' means a file format structured so that software applications can easily identify, recognise and extract specific data, including individual statements of fact, and their internal structure"*([93]). In the LAION Case however the Court expressed doubts about the need to interpret 'machine-readable' in the same manner under the CDSM and the Open Data Directive given that both legal instruments have different objectives.

2.3.1.3 GEMA vs OpenAI (Germany)

GEMA is a German music performance rights organisation and one of the largest CMOs in the world. It alleges that OpenAI has infringed the exclusive rights of German lyricists by using the lyrics of the musical works to train its AI systems, specifically ChatGPT. The case was filed in November 2024 before the Munich Regional Court and is currently ongoing.

GEMA states that it has strategically chosen to file an action based on lyrics (as opposed to musical compositions) as infringements can more easily be established for lyrics (text) than for audio (where there may be more inherent subjectivity in determining similarity) ([94]).

GEMA claims that ChatGPT undertakes unauthorised reproductions of these lyrics when simple prompts are entered by a user, suggesting that the system has been trained on these texts without authorisation. GEMA's press statements reference its AI Charter, which includes the principles of **'Protection of Intellectual Property'** and **'Fair Participation in Value Creation'**. It describes its action as a **'test case'** and *"a model action to clarify AI providers' remuneration obligations in Europe"*.

The organisation states that it is considering lawsuits against other AI providers and that its aim is not to generally prevent the use of works by AI systems, but to obtain licence fees and

([93]) Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast), OJ L 172, 26.6.2019.

([94]) GEMA files model action to clarify AI providers' remuneration obligations in Europe, CISAC 13 November 2024, (accessed 14 March 2025).

ensure **'fair remuneration' for authors** - both for use of works as training data and for reproductions generated from GenAI systems.

A key observation is that GEMA appears to be seeking to enforce its member's rights not just with respect to TDM (i.e., to establish a basis for licensing works for GenAI input use), but also to set a precedential basis for licensing works for output use (see *Section 2.4.5.2*).

2.3.1.4 GEMA vs Suno AI (Germany)

In January 2025, GEMA announced that it had filed a second lawsuit against a GenAI provider, Suno AI, a USA-based company that developed a tool for creating AI-generated audio content. In its press release ([95]), GEMA stated that the lawsuit is based on evidence suggesting Suno AI's tool can be prompted to produce synthetic songs that closely resemble, in terms of melody, harmony, and rhythm, works within GEMA's repertoire. The evidence reportedly submitted to the court includes well-known songs, and a side-by-side comparison of some of these songs with their AI-generated counterparts is available on GEMA's website ([96]).

The lawsuit claims that Suno AI made unauthorised use of musical works for two purposes: **training its music-generating tool** and **creating AI-generated products** that reproduce the works in a 'confusingly similar' manner. To support its claim GEMA points to statements made by Suno AI in USA court proceedings, where the company reportedly admitted to using 'pretty much everything available on the internet'. Additionally, GEMA cites the production of 'confusingly similar' content as further evidence of unauthorised use.

2.3.1.5 SNE vs Meta (France)

In March 2025, the French Publishers' Association (*Syndact national de l'édition* – SNE), alongside the Society of Writers (*Société des Gens de Lettres* – SGDL) and the National Union of Authors and Composers (*Syndicat national des auteurs et des compositeurs* – SNAC),

([95]) Fair remuneration demanded: GEMA files lawsuit against Suno Inc., 21 January 2025, and FAQ on the AI lawsuit, both GEMA (accessed 14 March 2025).

([96]) Audio samples: How Suno copies famous songs, GEMA (Accessed 14 March 2025).

announced that they had initiated legal proceedings against Meta before the Paris Judicial Court. The lawsuit alleges copyright infringement due to the unauthorised use of the claimants' works in Meta's training datasets ([97]).

2.3.1.6 EU Litigation Perceptions and the Role of Competition Authorities

Some interviewed European rights holders representatives have suggested that a possible reason for relatively low litigation rates in the EU is that stakeholders are being **cautious and discreet** with their strategies, while observing the **rollout of the process of implementation of the AI Act**. Thus, some rights holders consider that their interests might be better addressed through **regulatory processes** rather than direct litigation. Some stakeholders have also indicated that they foresee a shift towards **relief through competition law investigations** into the AI sector.

This reliance on competition law seems to be influenced by the decision of the French Competition Authority (Autorité de la concurrence) in March 2025 to fine Google €250 million for failing to comply with previously made binding commitments under a June 2022 Decision ([98]). This case centred upon negotiations with press publishers, and in its March 2024 Decision, the Authority explicitly discussed that the Google 'Bard' AI system ([99]) was trained on press publishers' content without authorisation. The Authority also noted that Google failed to propose a technical solution for press agencies and publishers to opt-out of the use of their content by Bard without affecting the display of such content on other Google services (search engine).

([97]) Authors and Publishers Unite in Lawsuit against Meta to Protect Copyright from Infringement by Generative AI Developers, SNE, 18 March 2025 (accessed 29 March 2025).

([98]) Autorité de la concurrence - Décision n° 24-D-03 du 15 mars 2024 relative au respect des engagements figurant dans la décision de l'Autorité de la concurrence n° 22-D-13 du 21 juin 2022 relative à des pratiques mises en oeuvre par Google dans le secteur de la presse.

([99]) In February 2024, 'Bard' was renamed 'Gemini'. See Google Blog, 8 February 2024 (accessed 14 March 2025).

2.3.2 Litigation in the USA

There are a relatively **large number of ongoing lawsuits** regarding copyright enforcement and AI in the USA. As a result, the issues surrounding the use of copyright protected works in AI training and deployment have been widely covered by general interest news and media outlets.

In January 2023, the first major case was *Andersen v. Stability AI* ([100]). A group of visual artists filed a class-action lawsuit against Stability AI, whose Stable Diffusion GenAI models are deployed by various providers including DreamStudio, Midjourney, and DeviantArt. The artists allege direct and induced copyright infringement. This case intensified public debate in the USA about AI's impact on professional creators as it was filed by independent artists.

The *Anderson Case* was followed by *Getty Images v. Stability AI* ([101]). Getty Images owns a large repository of stock images which it licences to commercial users and media companies. In February 2023, Getty Images filed a lawsuit against Stability AI for allegedly copying more than twelve million images (with associated captions and metadata) which were used to train Stability AI's Stable Diffusion model. This case also received wide coverage in the general media with evidence presented to the public that Stable Diffusion generative outputs sometimes contain digital artefacts which allegedly resemble Getty Images' watermarks.

In mid-2023 to 2024 several lawsuits from copyright owners against AI companies focused on literary works. The most publicly discussed case is probably *New York Times v OpenAI* ([102]). In December 2023, the New York Times (NYT) filed a lawsuit against Microsoft and OpenAI, claiming that OpenAI models are trained on millions of NYT articles. The NYT submitted that AI-driven services, including Microsoft Bing search index, and ChatGPT, provide verbatim excerpts of NYT works. The NYT has asked a federal court to order OpenAI to identify all of NYT content that has been used to train its models.

An overview table of the current lawsuits in the USA related to copyright and GenAI is available in *Annex VI*.

([100]) *Andersen et al v Stability AI Ltd. et al* 23-cv-00201-WHO (N.D. Cal. Aug. 12, 2024).

([101]) *Getty Images (US), Inc.* v. *Stability AI, Inc.*, 1:23-cv-00135-JLH.

([102]) The New York Times Company v Microsoft Corporation and OpenAI Inc. *et al, Case 1:23-cv-11195.*

As these cases are ongoing, key legal issues are yet to be resolved. Nevertheless, it is evident that the key question is whether – or rather under what circumstances might – TDM and/or AI training fall **within the fair use defence of USA Copyright Law** ([103]).

A notable development is a 11 February 2025 (revised) summary judgement from the USA District Court of Delaware, in the case *Thomson Reuters v. Ross Intelligence* ([104]). This case involved the defendant's use of case documents from the plaintiff's Westlaw databases, which included copyright protected legal headnotes, to train a competing AI-driven legal research platform. In its summary judgement, the Court found that the defendant's actions did not constitute fair use. While this case involved using copyright protected works for AI-training, it does not specifically relate to a GenAI use case, with the Court noting that *"because the AI landscape is changing rapidly, I note for readers that* ***only non-generative AI is before me today****."*

A broad assessment of the various USA copyright and AI cases leads to a few generalised observations ([105]):

- GenAI legal disputes often involve claims of copyright infringement, **unfair competition**, misappropriation, trade mark dilution, and breach of publicity rights. Breach of contract claims are also common, with some rights holders claiming that web-scraping violate the terms and conditions of websites that prohibit such practices. Additional claims are sometimes also based on the unauthorised removal of rights-management information ([106]).

- Most litigation **concerns text and literary works** protected by copyright, with several cases brought by book publishers, and to a lesser extent press publishers. Formal legal disputes regarding images are also relatively more frequent, while disputes over music or audio-visual works are less frequent.

([103]) 17 U.S. Code § 07

([104]) United States District Court for the District of Delaware, Thomson Reuters Enterprise Centre GMBH and West Publishing Corp. (Plaintiffs) v Ross Intelligence Inc (Defendant), Case 1:20-cv-00613-SB Document 770, Memorandum Opinion, 11 February 2025.

([105]) It must be stressed that these points are stylised observations on the body of litigation, and do not amount to any opinion on the respective facts or legal issues.

([106]) 17 U.S. Code § 1202 - Integrity of copyright management information.

- Rights holders actions often **combine claims of infringement** triggered by unauthorised use of works as training data **(input claims)**, and claims that GenAI systems produce infringing content **(output claims)**. The identification of specific works that have been infringed is critical for rights holders to have successful claims. Broad claims of infringement are generally dismissed unless the infringement action is rooted in the exclusive rights on specific works that are explicitly identified.

- A key challenge of 'input claims' is that rights holders often cannot **prove on a factual basis that GenAI systems have ingested their works**. In some cases, rights holders' legal arguments rely on referencing public documentation in which an AI developer cites the training datasets that it uses (e.g., in a technical white paper), and the inclusion of their works in these datasets where such information is public.

- Several cases concern not just the use of content obtained from web scraping, but copyright-protected material sourced from **'shadow libraries'**, which are extensive online collections that aggregate known unauthorised content.

- In other cases, rights holders' claims are based on **reasonable inferences about the training data used**, by demonstrating that specific prompts lead to (potentially infringing) outputs which can only be generated if specific works were ingested as training data. In this way, the supporting evidence of input claims and output claims are directly linked.

- Difficulties in rights holders being able to ascertain whether their works have been used in AI training datasets, including during the litigation proceedings, has been **driving the discourse on potential obligations on training data disclosure**. There are at least two legislative proposals in this regard – the '*AI Foundation Model Transparency Act of 2023'* (Rep. Beyer, 2023), and the *'Generative AI Copyright Disclosure Act of 2024*' (Rep. Schiff, 2024). There is also legislation at the level of State legislatures, such as California Bill *AB-2013 on 'Generative artificial intelligence: training data transparency'*, which was adopted and will come into force in January 2026 ([107]).

([107]) California State Bill *AB-2013* *Generative artificial intelligence: training data transparency* *(2023-2024).*

- A key legal question is the **interpretation of 'copies' under copyright law**, as this term has a statutory definition under USA copyright law ([108]). The definition of a 'copy' and how this term is interpreted in the context of GenAI model development may have a major impact on the liability of AI models. A very broad interpretation may lead to a GenAI model being deemed to constitute a 'copy' meaning that infringement may arise not only in terms of unauthorised reproduction or works (through data mining or infringing output), but even through unauthorised distribution (when the model is commercially deployed).

- Claims are sometimes dismissed where rights holders are **unable to prove actual harm** incurred from unauthorised use. This may disadvantage smaller rights holders who claim that their works are used without authorisation for AI training, but do not necessarily have strong claims in relation to infringing output which competes with their original works.

2.3.3 Litigation in other non-EU Countries

Aside from the EU and USA, there is publicly available information about litigation between rights holders and AI companies in China, the United Kingdom (UK), Canada, India, and South Korea.

2.3.3.1 Guangzhou Internet Court (China)

The decision of the Guangzhou Internet Court on 8 February 2024, centred on a dispute between Shanghai Character License Administrative Co., Ltd. (SCLA) and a Chinese AI company operating a platform supporting text-to-image GenAI services. SCLA held exclusive rights over the character Ultraman. The allegation was based on the use of Ultraman's images in training the AI model and their unauthorised reproduction via the company's platform. SCLA argued that when prompted the platform generated images substantially similar to Ultraman

([108]) 17 USC §101: '"Copies" are material objects, other than phonorecords, in which a work is fixed by any method now known or later developed, and *from which the work can be perceived, reproduced, or otherwise communicated*, either directly or with the aid of a machine or device. The term "copies" includes the material object, other than a phonorecord, in which the work is first fixed.' (emphasis added).

characters and monetised this feature through membership fees and 'computing power' purchases. The AI company denied liability, asserting it ceased operations upon notification of the case, lacked intent to infringe, and that the image generation was conducted by a third-party provider. Furthermore, it argued that there was no proof of direct profits or deliberate copyright infringement.

The Court ruled based on Copyright Law of the People's Republic of China (PRC) and the Interim Measures for the Administration of Generative Artificial Intelligence Services, issued in August 2023 by the Cyberspace Administration of the PRC. It discussed the alleged violations of the rights of reproduction, adaptation, and dissemination via information networks. The court found that the AI-generated images were **substantially similar to Ultraman's copyright-protected features**, constituting unauthorised reproduction and adaptation. However, it chose not to assess the infringement of network dissemination rights, as the other two rights adequately addressed the infringement issue.

The defendant was deemed a 'generative AI service provider' under Article 22(2) of the Interim Measures and was held accountable for ensuring the cessation of infringing activities. The Court found that keyword filters implemented by the defendant to prevent the generation of infringing content were insufficient, users could still generate Ultraman-like images using alternative prompts. Consequently, the court ordered the defendant to adopt **more robust preventive measures**. However, it rejected SCLA's request to delete copyright materials from the training dataset due to insufficient evidence of the defendant's involvement in model training.

Regarding civil liabilities, the court identified deficiencies in the defendant's operations that exacerbated the infringement. The absence of complaint mechanisms, user warnings about potential copyright violations, and explicit labelling of AI-generated content were noted as significant oversights. The court emphasised the importance of **transparent AI practices to protect intellectual property and user awareness**. These failings justified awarding damages to SCLA for its economic losses and enforcement expenses.

While the Court held the AI company accountable, it acknowledged the challenges of balancing copyright protection with GenAI development. The judgment deliberately refrained from addressing whether the use of copyright-protected material for AI training constitutes infringement. Acknowledging that such a determination could disproportionately hinder the AI

industry, the court chose to **focus its ruling on content generation (output) rather than the training process (input)**.

2.3.3.2 Getty Images vs Stability AI (UK)

Getty Images filed a lawsuit against Stability AI in the High Court of Justice in London. It is important to note that UK Copyright does provide for a TDM exception which allows copies of works to be made *'in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose'* ([109]). This provision is similar but not equivalent to the Article 3 exception under the CDSM Directive.

This case is interesting for several reasons. First, legal disputes between Getty Images and Stability AI are making their way through two courts in different jurisdictions with different legal systems – the District Court of Delaware (USA) and the High Court of Justice in London (UK). These two cases are a test for how **different legal systems** with distinct copyright laws will adjudicate between the same parties on a similar set of facts. This case may highlight the differences between the USA law's fair use framework and the UK's TDM exception as is incorporated into its Fair Dealing framework.

Second, the case involves questions of **private international law**. This is a critical dimension given the international nature of the AI ecosystem, where TDM practices, model development, and model deployment may take place in different jurisdictions. Stability AI claims that Stable Diffusion training and development took place in the USA, while Getty contends that some infringing activity took place on servers in the UK.

Third, a contested issue in this case is the interpretation of the term 'article' as it is used in the UK copyright act (CDPA 1988), particularly in the context of the statutory provisions on infringing copies and secondary infringement, and how this applies in the GenAI context. This parallels similar legal debates in the USA regarding the interpretation of 'copies' under USA Copyright Law.

([109]) Copyright, Designs and Patents Act 1988, Section 29A.

2.3.3.3 CanLII v Caseway AI (Canada)

The Canadian Legal Information Institute (CanLII) is a non-profit organisation funded by the Federation of Law Societies of Canada. CanLII manages a freely accessible public database of Canadian legal documents, which holds approximately 3.5 million documents, and is widely used by Canadian researchers and legal practitioners. Caseway AI is an AI start-up company founded in Canada but incorporated in Ireland, who has developed an AI-chatbot to assist in legal research.

In November 2024 CanLII filed a lawsuit in the Supreme Court of British Columbia, alleging that Caseway AI scraped its database to train their chatbot. In doing so Caseway AI had violated the terms of use of the CanLIIs' database and infringed copyright. The Notice of Civil Claim filed by CanLII (on 4 November 2024) states that: *"CanLII expends significant time, resources and expertise to review, analyse, curate, aggregate, catalogue, annotate, index and otherwise enhance the Data prior to publishing its original work product (being the CanLII Works) on the CanLII Website"* ([110]). The breach of contract is based on the terms of use stated on the CanLII website, which include a prohibition on *"bulk or systematic downloading of the CanLII Works, including by way of programmatic means or by way of hiring human resources to manually download the-CanLII Works*". CanLII claims that Caseway's unauthorised use and subsequent publication and distribution of the copied materials (through its AI services) amounts to a violation of copyright.

2.3.3.4 Canadian News Media Companies v OpenAI (Canada)

In November 2024, a coalition of leading Canadian media companies and news publishers brought a claim against OpenAI before the Ontario Superior Court of Justice ([111]). The media companies claim that OpenAI infringed copyright when it scraped their websites without authorisation, ignored copyright restrictions in their websites´ terms and conditions, bypassed

([110]) Supreme Court of British Columbia, Court File No. VLC-S-S-247574 – Notice of Civil Claim.

([111]) Ontario Superior Court of Justice, Court File No.: CV-24-00732231-00CL - Statement of Claim.

paywalls and circumvented technological protection measures in its scraping activities. OpenAI submitted that its models are trained on publicly available data.

2.3.3.5 Asian News International v OpenAI (India)

The Indian news agency company, Asian News International (ANI), brought a lawsuit against OpenAI before the Delhi High Court in November 2024. ANI claims that OpenAI has used its news content to train ChatGPT without authorisation and that OpenAI is also responsible for harm to ANI's reputation due to fabricated news stories generated by ChatGPT and falsely attributed to ANI. OpenAI submitted that ChatGPT is trained on publicly available data, that its use of data represents facts not protected by copyright, and that it has respected the requests of ANI to cease training on its content by blocking its domain. OpenAI further argued that the Indian Court does not have jurisdiction to hear the matter since neither OpenAI nor its servers are based in India.

In the first hearing, the Court framed four key issues under consideration: (i) whether the storage of ANI's data for training amounts to copyright infringement, (ii) whether the use of the data to generate user responses amounts to infringement, (iii) whether the use qualifies as 'fair use' (fair dealing) under the Indian Copyright Act, and (iv) whether the Courts in India have jurisdiction in the matter given that OpenAI and its servers are located in the USA. In January 2025, the Federation of Indian Publishers and the Digital News Publishers Association (DNPA) and the Indian Music Industry (IMI) filed pleas to intervene in the case. In February, two further parties, the Indian Governance and Policy Project (IGAP) and Flux Labs AI, sought to intervene in the case on public policy grounds.

2.3.3.6 Korean Broadcasters v Naver (South Korea)

In January 2025, several South Korean news outlets reported that three South Korean territorial broadcasting organisations (KBS, MBC, and SBS) filed a lawsuit against the South Korean tech company Naver. The broadcasters state that Naver used copyright protected news articles without authorisation, for training its AI platform. Public information on this case is currently limited.

2.4 Licencing and Dataset Markets

In parallel to the set of litigations initiated by rights holders against AI developers, there has been an **emergence of an evolving market for training data**. The actors in this market include rights holders who may licence their content for TDM use, and **TDM users who acquire data and create training datasets** for use by downstream developers.

As discussed in *Section 3.1*, a large proportion of resources in the AI value chain are dedicated to developing training datasets, including data curation and processing, both at the general-model training (or pre-training) and fine-tuning (post-training) levels.

Figure 2.4-1 below shows the significant increase in overall private investment in the AI ecosystem between 2022 and 2023. Investment in 'creative, music, and video content', the investment category specifically linked to GenAI systems, has reduced. It should be taken into consideration that most categories of investment (except notably 'data management and processing') have decreased, with investment shifting towards 'AI infrastructure, research, and governance' which creates returns relevant to all AI focus areas. *Figure 2.4-2* highlights that estimated training costs for models have continuously increased over time, with newer models associated with higher training costs. However, exceptions exist, such as DeepSeek, which has been reported to require significantly lower computing power, potentially shifting the cost-efficiency dynamics of model training (see *Section 3.1.8*). *Figure 2.4-3* shows that estimated training costs are correlated with the necessary training compute.

Overall, it is difficult to reliably estimate the investments made in data acquisition and processing for training AI models and systems. The observed patterns all point towards the importance of large quantities of data in the AI ecosystem and **the importance of investment in training data**, creating potential for robust training data markets.

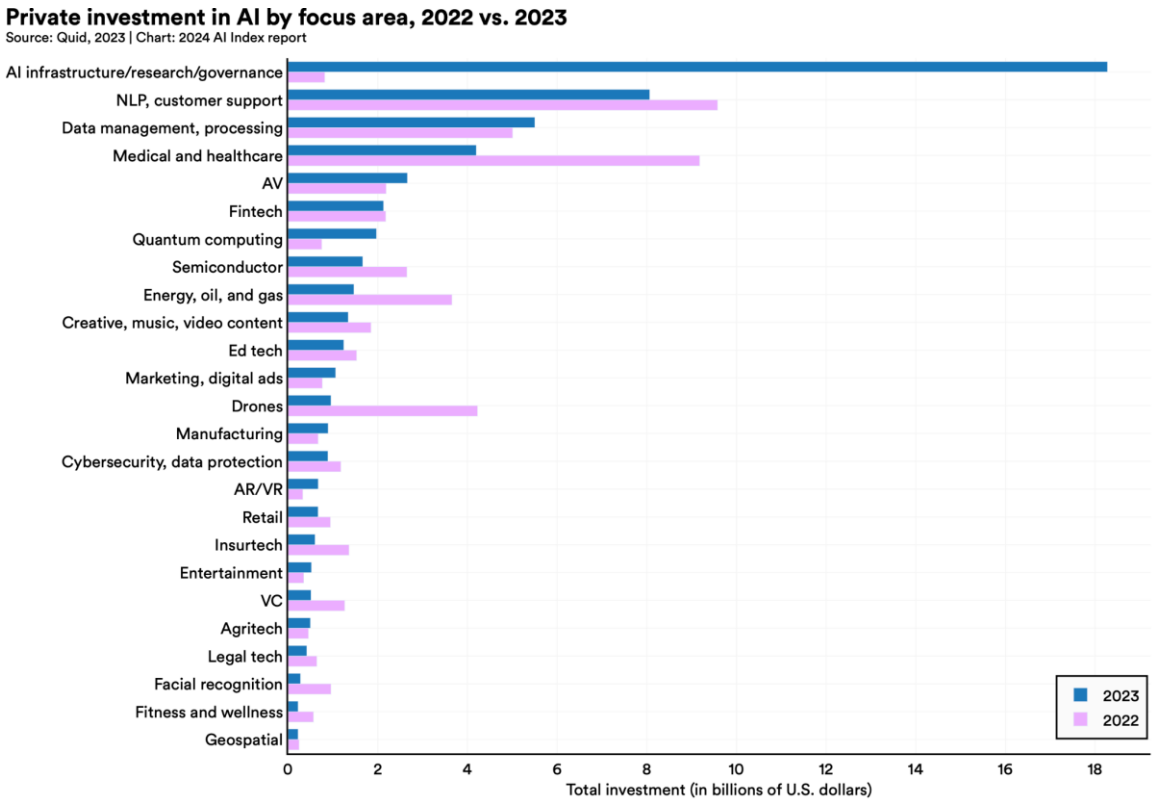


Figure 2.4-1: Private Investment in AI ([112]).

([112]) The AI Index 2024 Annual Report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024. ('*Stanford AI Index Report, 2024*'); p 254 (accessed 14 March 2025).

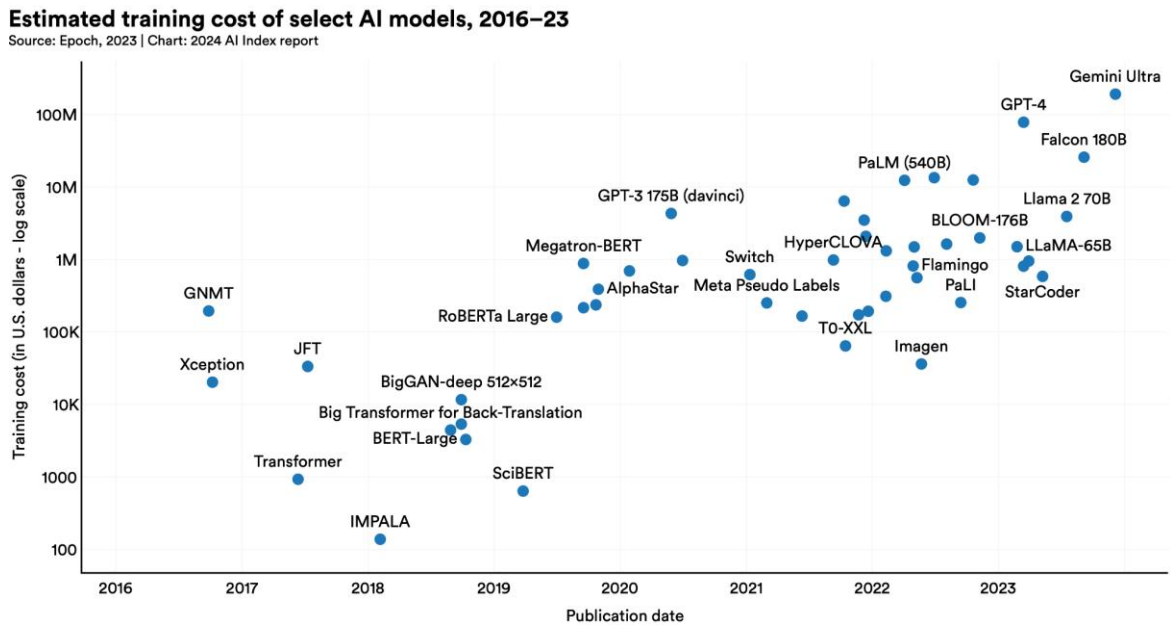


Figure 2.4-2: Estimated Training Costs ([113]).

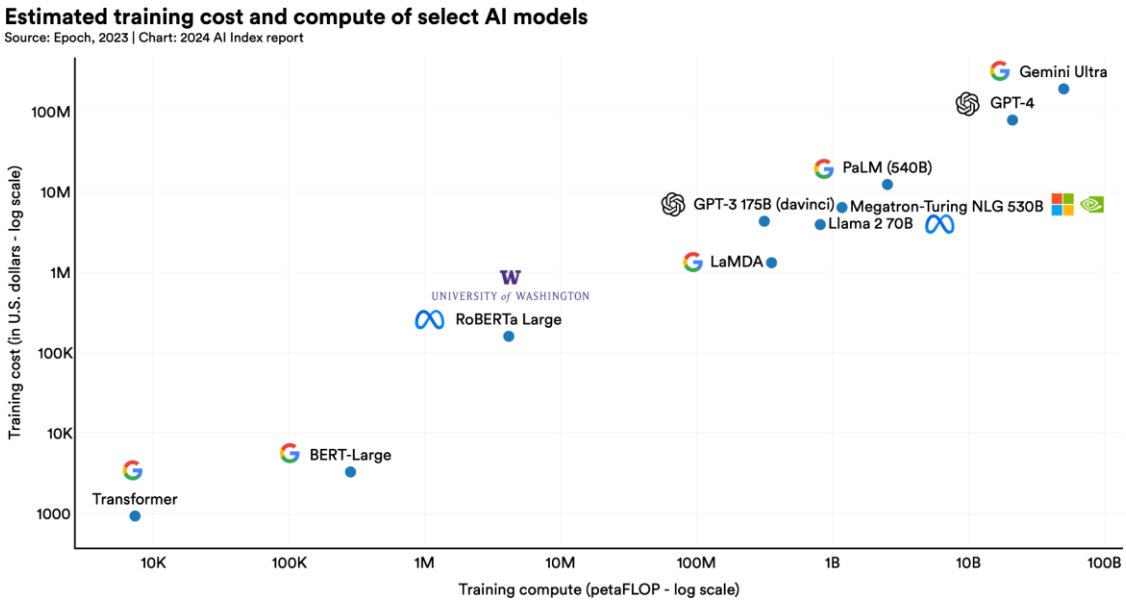


Figure 2.4-3: Estimated Training Costs and Compute ([114]).

([113]) Ibid. p. 56

([114]) Ibid. p. 65.

Given the critical role played by training data, there are **different sources for training datasets** with distinct markets and technical factors. The sub-sections below discuss some of the key aspects of the training data market.

2.4.1 Datasets Development

Training datasets may include a variety of information and data from different sources, and constituent data elements may or may not be subject to IP rights. Furthermore, it is critical to note that even where data sources are **'freely available' on the open internet**, it does not mean that this content is free of intellectual property rights. Even when a dataset is made available and openly accessible, this does not mean that there is automatic authorisation to use that dataset and the content within it.

Early development of AI systems was generally based on carefully sourced, curated, and labelled datasets. The evolution of AI technologies has given rise to **demands for increasingly large training datasets**, which has led to the importance of datasets derived from a variety of sources. Content scraped from online sources has become a critical component of the AI ecosystem. *Section 3.1.2.1* provides further details on commonly used training datasets, and the typical structure and organisation of such datasets.

To understand the AI value chain and the role that data plays within it, it is critical to stress that data curation and processing activities themselves may involve different actors, many of whom - at least conceptually - undertake some form of TDM. Data scraping and processing may be done by TDM users, **specialised in producing datasets,** not necessarily AI developers. For example, the LAION datasets of image-text pairs used Common Crawl archive data as a starting point (see chapter on Common Crawl in *Section 3.1.2.1.2*). This data was filtered and processed to improve its quality and suitability as a training dataset for image-based AI systems, with the completed dataset being distributed freely to the public in the form of a spreadsheet of hyperlinks and text descriptions of images.

Raw data, such as that scraped from the publicly accessible internet, may just be a starting point for AI training datasets. The data-processing stage sometimes involves the annotation of data to make it more suitable for training purposes. The result is that **datasets that undergo some pre-processing,** through steps like filtering and annotation, may themselves be

protected by intellectual property rights. If there is originality due to the arrangement and selection of the data then copyright may apply. Even where copyright does not apply, the substantial investments undertaken in data processing may possibly meet the threshold for *sui generis* database protection (see *Section 2.2.1.2*). This is in addition to rights that might exist in **added metadata** contributions for individual data elements. These rights in a compiled dataset and its metadata may be another layer of IP that exists on top of any applicable IP in specific content contained in the individual data elements. Furthermore, even when datasets are compilations of mere facts which are not protected by any exclusive rights, **the dataset itself might be protected by some form of IP** (such as database rights), and these datasets are usually distributed under some specific usage terms.

Thus, the **terms of dataset distribution** as well as applicable TDM provisions may need to consider the relevance of both layers of rights (if they exist). There is a dynamic of more than one TDM use relevant to the dataset market. First, data may be scraped from the internet (or sourced through some other mining process) and compiled into raw datasets through a TDM process. These datasets may then be annotated to create supervised datasets which themselves may be protected (by copyright or *sui generis* database rights). Subsequently, these supervised datasets are used to train AI models through another TDM process.

A specific legal challenge with supervised datasets is also that **annotations might be created with the assistance of AI systems**. This may result in potential violations with the terms of use of such systems, as many foundation models are released with terms that stipulate that the model cannot be used for creating competing models ([115]).

As noted previously, the legality of TDM for AI training is a critical open question in the USA legal system. The context of fair use may be different for using copyright protected works, and using databases created specifically to serve as training resources for AI models. This is because US fair use considers both the purpose and character of a use (and whether it is transformative in relation to the purpose of the original work), as well as the effect of the use on the market for the original work ([116]). The TDM exceptions under EU copyright law do not

([115]) For example, Anthropic's Consumer Terms of Service states that: *"You may not access or use, or help another person to access or use, our Services in the following ways: ...To develop any products or services that compete with our Services, including to develop or train any artificial intelligence or machine learning algorithms or models or resell the Services"*.

([116]) 17 US Code §107.

make this distinction, meaning that once the purpose of use constitutes TDM, and relevant legal criteria are met, a TDM exception may likely apply equally to protected creative works (used to create training datasets) and the actual datasets (if protected) for TDM during AI training. However, **'lawful access' to the protected work or database is a pre-requirement** for benefiting from the Article 4 TDM exception under EU copyright law. The terms and conditions of accessing a protected dataset are relevant to the question of whether lawful access exists, and whether the TDM exception will apply when using such a dataset.

2.4.2 Datasets Distribution

Platforms and community networks where datasets are shared and openly distributed play an important role in this ecosystem. Two of the best-known platforms are Hugging Face (a private company focused on promoting open-source approach to AI development, and self-described as 'on a mission to democratize good machine learning') ([117]), and Kaggle (a platform for data scientists owned by Google). These platforms are important actors in the AI value chain, as they create the **distribution framework for dataset dissemination** and widespread use. The datasets hosted on these platforms are not only those sourced through scraping, but also include datasets curated or developed by dataset creators of digital assets (including synthetic data). These platforms provide spaces in which open data practices effectively facilitate downstream training activity by AI developers, where **investments in data acquisition and processing** are made by actors who are not necessarily AI developers.

To assist developers, there is a growing space in the data value chain for independent tools that provide a meta-analysis of datasets, which can include statistical metrics on content diversity, analysis of potential bias, as well as guidance on copyright compliance. To some extent, such tools are being more and more integrated into dataset distribution platforms, including increasing details in **'dataset cards'** which outline metadata information attached to specific datasets.

Empirical research auditing 1858 datasets hosted on major dataset platforms (GitHub, Hugging Face, Papers with Code) has found that these platforms are often prone to mislabelling the licences attributed to these datasets (Longpre et al., 2023). Often, the

([117]) Huggingface website (accessed 14 March 2025).

mislabelled licences suggest use which is more permissive than what was presumably intended by the dataset licensor. Dataset platforms were found to have a large proportion of datasets missing licences. According to this research even where licences are indicated, the licence information on these dataset platforms was sometimes mislabelled, whith GitHub, Hugging Face and Papers with Code, each labelling license use too permissively in 29%, 27%, and 16% of cases respectively. The research suggests that in many circumstances this is not due to intentional mislabelling of licences, but rather platform contributors mistaking licences attached to open-source code for licences attached to data.

When a dataset has an 'unspecified licence' it would be unclear to a potential user whether this is intentional, and the dataset has been released without a licence, or whether this is a shortcoming of the aggregation platform. As a result, whether these datasets are used by an AI developer depends on the developer's own perceptions of relative risk, and their level of risk aversion and tolerance. As risk aversion may naturally differ between the size of an undertaking and its ability to navigate and negotiate potential legal challenges from rights holders and dataset creators, this can lead to a distortion in the market for uses of AI training datasets.

Mislabelled datasets are also problematic as they may result in model developers **incurring liability for violating the true terms of dataset use**. This liability *may* potentially be passed downstream to users who integrate these models into AI systems and deploy the systems in various use cases. One response from some companies in the AI sector has been to guarantee indemnification to users of their models, in case of future legal liability due to unauthorised training data (see *Section 4.8*). However, this again is a strategy which may only be viable for larger undertakings who are able to navigate these legal issues and internalise the respective risks into their business strategies.

Thus, mitigating potential liability for training data, facilitating orderly development of dataset markets, and ensuring a balanced competitive environment in the AI ecosystem require attention to be paid to the terms on which datasets are licenced, and the mechanisms through which these terms are communicated. As suggested in the technical literature and by many interviewed stakeholders, part of this challenge is **adapting existing open-source licensing tools specifically for training dataset distribution**. Problems with using standard open-source software licenses for licensing training datasets include that some licences may contain prohibitions on the creation of derivative works, and it may be claimed by some that an AI

model is a derivative work of an entire dataset. Furthermore, open datasets are often processed in different ways (e.g., filtered or annotated for specific training use cases), and multiple datasets (with possibly conflicting licensing terms) may be consolidated into larger distributed training datasets. Possible solutions that have been proposed include a new standard open licence specifically designed for AI training datasets (e.g., BigScience Responsible AI License (RAIL)), and a modification of existing open licences such as the MIT licence (Ioannidis et al., 2024). Another possibility is that this issue resolves itself overtime - at least for datasets which are intended to be openly licensed - as uptake of Open Data Commons licensing schemes increases ([118]).

Critically, such solutions **only deal with the distribution of datasets themselves** and not any copyright-protected content that might be included as specific data elements. The CDSM TDM exceptions are related to the right of reproduction only, and **only reproduction pursuant to the TDM process itself** (or in the case of scientific research TDM, secure repositories for verifying research activities). TDM under the CDSM exceptions does not permit any reproduction in the form of copies of works included in training databases which are then distributed beyond the actual TDM user, nor do these exceptions permit any communication to the public which occurs through such distribution. Even when dataset elements are not reproduced directly in a dataset but are distributed to potential dataset users in the form of hyperlinks (as in the case of LAION image-text pairs), there may still be a potential copyright relevant act taking place, given the CJEU jurisprudence on hyperlinking and the exclusive right of communication to the public (Rosati, 2021).

It is important to understand the role that upstream licensing terms play throughout the TDM value chain, starting with the terms of use for content on the open internet, content scraped from large public datasets such as Common Crawl (see *Section 3.1.2.1.2*), and the terms of distribution of training datasets. These contractual terms are important for determining how a dataset may be distributed to users (including AI developers), even when the legal basis for initial legal TMD is clear.

The usage terms of **Creative Commons licences** and Common Crawl data are two important examples, given the importance of these two instruments in the data marketplace. First, Creative Commons has publicly stated that its standard open license (on which a substantial

([118]) Open Data Commons website (accessed 14 March 2025).

portion of open content on the internet is licensed, e.g., Wikipedia and Wikimedia Commons) **should not be construed as a CDSM Article 4 opt-out**([119])**.** Creative Commons licensed content is free to be used in TDM processes, and subsequently distributed in accordance with the licensing terms (which, important for AI training purposes, may include restrictions on commercial use). The extent to which open license regimes (in particular Creative Commons licences) continue to be used on a widespread basis in the post-AI-boom phase of the internet, may be an important factor that shapes training data markets. Second, the **terms of use of Common Crawl** explicitly contain provisions which prohibit users of Common Crawl datasets from **violating the intellectual property rights of third parties** (including rights that relate to protected content in the database itself).

2.4.3 Direct Licencing Markets

In addition to data obtained through crawling and other TDM processes, AI dataset development, training, and use may also be based on **content licensed directly from rights holders**, or agents representing rights holders in emerging training data markets.

Content licensed from rights holders is most often used in either post-training/fine-tuning of AI models or Retrieval Augmented Generation (RAG) applications (see *Section 4.1.2*). If data scraping techniques as a form of data acquisition are often the basis of general-purpose model developments where the quantity of training data is a key factor, fine-tuning requires small but higher quality datasets suitably adapted for specific use cases. The **market for licensing of content appears to be rapidly growing** with a number of rights holders entering into agreements with AI providers and even more signalling their willingness to enter into negotiations. Below are a few selected examples of major publicly announced licensing agreements between rights holders and AI developers.

- Stock image website Shutterstock, which claims to manage more than 530 million digital assets([120]) has entered several licensing deals with major AI Developers. The company has stated that its training data licensing agreements with 'anchor customers' are worth approximately USD$10 million in annual revenue, with customers including

([119]) Creative Commons website (accessed 14 March 2025).

([120]) See Shuttershock website (accessed 14 March 2025).

Meta, Apple, Amazon, Reka AI, and OpenAI with whom the company has a six-year licensing deal ([121]). Licensing content to AI companies has produced an estimated USD$104 million in annual revenue in 2023, accounting for a roughly estimated 12% of the company's overall revenue ([122]).

- AI-driven search engine ('answer engine') provider Perplexity AI has launched a 'Perplexity Publishers Program', for which the first batch of online partners include major news publishers such as TIME, Der Spiegel, Fortune, Entrepreneur, The Texas Tribune, and WordPress.com. Perplexity has indicated that it will be adding advertising to its platform, and when a publisher's content is referenced a revenue-share system will be in place. Interestingly, Perplexity has publicly signalled that it is open to developing a flat-fee subscription model for public users which would bundle Perplexity's services with subscriptions to participating publishers ([123]). It should be noted that Perplexity is a RAG system in which the traditional model of an LLM (developed through vast training data) is complemented by retrieving external sources which provide context for prompts to increase output reliability and quality. It is not explicitly clear from public statements whether Perplexity's agreement with publishers includes access to content for training data, or if the agreement is limited to enabling Perplexity's RAG functionality.

- OpenAI has secured agreements with a growing number of major media companies, particularly in the press publications sector. Publishers with which OpenAI has agreements include: Associated Press (AP) ([124]), Dotdash Meredith ([125]), FT Group

([121]) Shutterstock's AI-Licensing Business Generated $104 Million Last Year; Bloomberg, 4 June 2024; Reka Announces Partnership with Shutterstock; Shutterstock, 4 June 2024; Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data, Shutterstock, 11 July 2024 (all accessed 14 March 2025).

([122]) Shutterstock Reports Full Year 2023 and Fourth Quarter Financial Results, Shutterstock, 21 February 2024 (accessed 14 March 2025).

([123]) Introducing the Perplexity Publishers' Program, Perplexity, 30 July 2024 (accessed 14 March 2025).

([124]) ChatGPT-maker OpenAI signs deal with AP to license news stories, The Associated Press, 30 July 2024 (accessed 14 March 2025).

([125]) Dotdash Meredith Announces Strategic Partnership with OpenAI, Bringing Iconic Brands and Trusted Content to ChatGPT, PR Newswire, 7 May 2024 (accessed 14 March 2025).

(Financial Times) ([126]), Axel Springer ([127]), News Corp ([128]), Vox Media ([129]), The Atlantic ([130]), and European Publishers Prisa Media and Le Monde ([131]). While the terms of these agreements are generally not publicly disclosed, it is known that many of them specifically include authorisation to access and use works (particularly press publications) for the purpose of RAG.

- Several direct-licencing deals have also been secured not only with large rights holders (and platforms aggregating content on their behalf), but also platforms and networks whose content repertoire consists largely of user-generated content. For example, OpenAI has secured an agreement with Stack Overflow, a knowledge sharing platform for software programmers and repository of community know-how regarding coding practices ([132]). Google has an agreement with Reddit, a news and content aggregation platform whose content largely consists of user contributions which are ranked through a community feedback system ([133]).

There appears to be an absence of direct licensing agreements between prominent movie and television production studios and AI developers, although the audiovisual production is an economically valuable content sector. One possible explanation raised in industry discourse is that the film industry is defined by many creative agents and complex contractual agreements. As a result, there are a number of overlapping rights (in particular the image rights of actors), which would first require clearance, to develop new licensing markets.

While the above examples are just a few cases of the many recent direct-licencing agreements that have been concluded, **the terms of these agreements are largely not disclosed to the**

([126]) Financial Times announces strategic partnership and licensing agreement with OpenAI, Financial Times, 29 April 2024 (accessed 14 March 2025).

([127]) Axel Springer and OpenAI partner to deepen beneficial use of AI in journalism, Axel Springer, 13 December 2024 (accessed 14 March 2025).

([128]) A landmark multi-year global partnership with News Corp, Open AI, 22 May 2024 (accessed 14 March 2025).

([129]) Vox Media and OpenAI Form Strategic Content and Product Partnership, Vox Media, 29 May 2024 (accessed 14 March 2025).

([130]) *The Atlantic* announces product and content partnership with OpenAI, The Atlantic, 29 May 2024 (accessed 14 March 2025).

([131]) Global news partnerships: Le Monde and Prisa Media | OpenAI, Open AI, 13 March 2024 (accessed 14 March 2025).

([132]) Stack Overflow and OpenAI Partner to Strengthen the World's Most Popular Large Language Models, Stack Overflow, 6 May 2024 (accessed 14 March 2025).

([133]) Google expands partnership with Reddit, Google, 22 February 2024 (accessed 14 March 2025).

public. Taking a broad view of the data value chain and the overall AI ecosystem, several potential drivers for licensing markets can be identified. These potential drivers are summarised in the following sub-sections.

2.4.3.1 Rights Reservation as a Market Pre-Condition

The CDSM Directive explicitly states that TDM under the Article 3 exception for scientific research is to be without remuneration to rights holders ([134]). As the Directive does not explicitly state that remuneration is required for use of works under Article 4, the potential for remuneration arises from the possibility for rights holders to opt-out their works under Article 4. Subsequently, rights holders may licence the use of such works, following a general principle of copyright law ([135]), and the norms of standard contractual arrangements under which authorisation to use a work is granted against negotiated remuneration. The structure of **Article 4 creates the conditions for a possible market for licensing permissions** for commercial TDM uses, contingent on rights holders exercising their right to opt-out of TDM usage.

The market for direct licensing of copyright-protected content to AI developers is enabled by the Article 4 opt-out mechanism, making it a copyright infringement for AI developers to use opted out works that may be available for license. There are indications that a growing number of rights holders engage in public declarations of reservations of rights as a pre-requisite for potential licensing negotiations. **A well-functioning system for TDM opt-out is thus a pre-condition for a well-functioning market for legitimate content licensing**. This also creates **a market for technical solutions** for managing access to content (particularly in online settings) and administering TDM rights reservations.

Using opt-outs for strategic positioning is most important for rights holders whose content is most likely to be acquired through scraping and then included in training datasets. The strategic positioning of a specific right holder group thus depends on the various ways through which the protected content is accessible to the public, and the extent to which being made 'publicly available on the internet' is central to the rights holder's distribution model. This is a

([134]) Directive (EU) 2019/790 Recital 17.

([135]) Directive 2001/29/EC Directive Recital 30.

possible reason for which the majority of publicly announced content licensing agreements have been in the area of text and press publications (which are inherently more susceptible to being scraped on the publicly accessible internet). In addition to this, the viability of licensing agreements and hence strategic position of rights holders also depends on the extent to which **access control is undermined by the availability of unlawful sources** such as shadow libraries.

2.4.3.2 Data Drought

The intensity of data demand, to develop AI models, has led to concerns within the AI community that future generations of computer scientists will run out of data to scale and improve AI systems leading to a slowing down of machine learning progress. One projection estimates that the stock of high-quality language data will be exhausted by 2026, low-quality language data by 2030 - 2050, and image data by 2039 - 2060 (Villalobos et al., 2024).

The implications of data scarcity may vary significantly across different content sectors:

- **News Publishing**: This sector produces high volume of content in an extremely dynamic way, but the value of content is often time sensitive. News publication datasets face other challenges including layered copyright and press publisher's rights, and interactions with public interest exemptions in copyright law (e.g., reporting exceptions).
- **Creative Industries**: Artistic and entertainment content, while more static in nature, often comes with complex licensing agreements. These sectors are especially sensitive to data scarcity as originality and emotional resonance are difficult to replicate.
- **Technical/Scientific Documentation**: Proprietary datasets, sometimes confidential, are less accessible due to strict licensing and security concerns, making data scarcity a significant issue for industries relying on domain-specific knowledge.

Increased data scarcity also raises the potential value of direct licensing, incentivising rights holders to withhold granting permission to use their works in TDM to extract greater value at a future date. On the other hand, as machine learning and AI technologies progress, the incremental value of works as training data may either increase or decrease depending on the specific sector. For example, the per-token value of works for AI developers is likely to vary across different types of content, affecting willingness to pay for training data accordingly. This dynamic however relates specifically to licensing content for use as primary training data, and not necessarily the use of content in RAG applications. In the case of RAG licensing, market value is often due to the up-to-date nature of content (news publishing in particular) which decreases over time.

This change in value may also be affected by the extent to which synthetic data becomes a viable substitute for real data in training processes. Synthetic data, while promising, presents challenges in replicating the nuanced quality and diversity of real-world data, particularly in sectors like news publishing and creative industries. Moreover, questions remain about its ability to meet domain-specific requirements, as in technical or scientific fields.

2.4.3.3 Demand Driven by Data Quality

The **need for high quality data at the fine-tuning level** of AI development is also an important driver for licensing markets. There may be cases where content is 'publicly available online' but scraping may result in low-quality data ([136]) on which substantial processing needs to occur. Sourcing content and digital assets directly from rights holders may be associated with **higher quality metadata**, and lower risks of duplication. Thus, datasets licensed directly from rights holders may represent an economically efficient transfer from AI developers to rights holders, of the resources that would otherwise be allocated to data filtering, labelling, annotation, and pre-processing. In the image and photography sector, an important driver in

([136]) As outlined in *Section 3.1.2.2* on Web Scraping, the collection of data from web pages necessitates extensive curation. If this final step is not executed effectively, it may result in issues like those discussed in *Section 3.1.2.1.2* on Common Crawl. This serves as an example of a widely used scraping-derived dataset that, if not properly curated, could potentially lead to copyright infringements.

addition to image quality is access to **images without visible watermarks** which reduce the effectiveness of training image-recognition and generation systems.

Furthermore, targeted licensing agreements with specific rights holders is itself a form of data curation and filtering, where a right holder's specifical catalogue or repertoire is known to be aligned with a developer's training needs. More generally, datasets licensed directly can be important sources of data used to (re)train and fine-tune models which have been determined to generate biased results because of low-quality pre-training data ([137]).

A specifically identified area in which there is a need for targeted datasets is **multilingual text**. As noted above, Common Crawl archives are often used as a starting point for training datasets. Common Crawl's own analysis which can identify 160 different languages, found that some 43% of documents in its crawl archive are in English, with the second most common language being Russian - accounting for 6.2% ([138]). Public crawling of text tends to generate datasets which are highly skewed towards English language content, to the detriment of developing AI systems (particularly LLMs) trained in other languages. This is a specifically critical issue in the EU where there are 24 official languages (and many more others at sub-national community levels), and there have been active discussions on the use of AI policy for people speaking minority languages to participate more actively in public life and to avoid linguistic discrimination (Gerkem, 2022). Well-functioning licensing markets for AI training data can thus be seen as one of the ways for AI deployment to further the interests of European linguistic diversity.

The need for high quality data not only relates to metadata, but also the technical characteristics of digital assets themselves. From a technical standpoint, the quality of raw text licensed from a press publisher may be comparable to other publicly available texts scraped online. However, other types of works may differ in **technical quality and resolution** depending on the means through which they are sourced. Digital assets - specifically audio, images, and video - are often compressed for online distribution, and may not be as suitable as high-resolution data assets for specific training use cases. This puts the rights holders of categories of works for which the quality of digital copies may vary, in a stronger negotiating

([137]) For a discussion on how wider TDM exceptions can be used to address bias issues in AI training and performance, see: Levendowski, A. (2018).

([138]) Statistics of Common Crawl Monthly Archives, Common Crawl (accessed 14 March 2025).

position for the licensing of these works. Again, this dynamic is specific to the use of works as primary training data, as opposed to use of works in RAG applications where timeliness is the primary characteristic of data quality, particularly for news publications.

While data quality may be a driver of licensing dynamics in fine-tuning use cases, the current paradigm for **training GPAI models remains 'quantity over quality',** with some interviewed stakeholders suggesting for example, that a major AI developer would not be willing to engage in licensing negotiations with an audiovisual content provider for a repertoire of less than 50,000 hours of audiovisual content.

On this point, some interviewed stakeholders explained that the commercial audio-visual sector may have a unique opportunity given the nature of film, television, and video production. The formation of audiovisual content typically involves the creation of quantities of data and content much larger than what appears in a final commercial product. Audio-visual producers often have large archives of hundreds of hours of multiple recorded takes, b-roll, and raw footage that is unused or unusable. Therefore, direct licensing for AI training represents a new path to monetising content which might otherwise just be costly archiving material to preserve.

2.4.3.4 Market Evolution and Risk Aversion

While there are still various uncertainties regarding the nuances of AI regulation, pioneer developers emerged in a pre-AI Act market where there was even higher legal uncertainty. The strategic choices of subsequent market entrants are driven by different conditions in the post-AI Act environment. Later entrants may be **more risk-averse** in their approaches to TDM and selection of training data, **which may be another factor driving demand for licensing content from rights holders**. Furthermore, public discourse over copyright and AI issues may also be affecting investor and consumer attitudes towards the GenAI sector and driving investment to and demand for AI services following **'ethical AI' business practices** (beyond regulatory pressures), including revenue sharing with rights holders whose contents are used in training and content generation processes. The existence of a certification scheme like 'Fairly Trained' - with an organisation administering certifications for GenAI models that only use explicitly authorised training data – is evidence of **increased attention to data licensing as a dimension of AI ethics**. Additionally, some AI models (such as Bria.ai text-image

service) use the claim that they are 'fully trained with licenced data' as a differentiating branding strategy.

Interviewed rights holders also suggested that entering into licencing agreements might be perceived as an admission of legal liability (on the part of AI companies) that authorisation from copyright owners is required. In a similar vein, there is a perception that AI companies are reluctant to enter into licensing agreements as the **negotiated terms may become reference points for damages** in the context of on-going and future litigations. Some interviewed rights holders have also suggested that inertia in licensing agreements may be related to the lack of transparency on the side of AI companies that ingest copyright protected works as part of the model training process. This lack of transparency undermines the **ability of potential licensees and licensors to negotiate on equal terms** with comparable information.

2.4.3.5 Synthetic Data

Within the machine learning community there seems to be increased discourse on the use of synthetic data as training data (see *Section 3.1.2.3*). While the appropriateness of synthetic data varies between different model use cases, the potential shift towards synthetic data may be driven by concerns of an increase in the cost of natural datasets in the face of stronger copyright compliance obligations and rights holders reservations. Currently, the view appears to be that overreliance on the synthetic test data by GenAI models may result in declining output quality and a phenomenon known as model collapse (Alemohammad et al., 2023; Shumailov et al., 2024).

2.4.3.6 Internal resources within Rights holders Organisations

Stakeholder interviews found that negotiating direct licences **requires significant internal resources** within a rights holders organisation. In many cases, negotiations can take several months of dedicated staff work. Licensing for basic conversational LLMs appears to require the least resources, while special use cases and sandboxes (innovative or frontier AI applications) require substantially more time and expertise. In this context larger rights holders

with access to internal resources and existing institutional roles for licensing strategy and negotiations are better positioned to enter into direct licensing markets.

At the same time, if many content companies do not necessarily have the resources to anticipate and embrace technological developments, the **potential demand for direct licences is driving new internal processes**. This involves formalising various operational aspects such as the digitisation of contracts, renegotiating terms with rights holders (for example in the case of publishing companies), as well as cataloguing and storing content in multiple formats and with relevant metadata. This represents a major paradigm shift for some companies in the way they operate technically and commercially. In practice, **approaches to data management have shifted** to take potential AI applications into consideration. In commercial terms, the paradigm shift is that content protection measures which are traditionally about loss prevention, are now reframed as creating new avenues for revenue generation. These operational shifts are both a driver and a result of direct licensing markets.

2.4.3.7 Content Licencing Aggregators

The development of content licensing markets has led to the emergence of new actors in the AI ecosystem. These actors serve as **content aggregators**, as new types of **intermediaries between rights holders and AI developers**. Notable examples of such aggregators include Datarade, Created by Humans, and Protoge Media (formerly called Calliope Networks).

This has also led to new roles for existing content distribution intermediaries. For example, digital music distribution platform TuneCore (a service largely used by independent music artists to distribute sound recordings to online streaming services) has introduced a 'AI and Data Protection Program' ([139]). This programme is opt-in (currently by invite only) and allows the digital distributor to manage rights reservations on a participating artist's behalf, as well as licence their content for AI applications.

Typically, rights holders in the best licensing negotiation positions are those who control large, centralised repertoires of commercially valuable works, such as CMOs, large production companies, media conglomerates, and publishing houses. Content licensing aggregators are

([139]) TuneCore's AI & Data Protection Program, Tune Core (accessed 14 March 2025).

likely to gain increasing importance as the AI ecosystem develops, as they facilitate access to the training data licensing markets to all rights holders, including those who have rights on a limited number of works. This **benefits smaller independent rights holders**, as well as types of content not traditionally managed through CMO representation. Such rights holders might otherwise not have the **scale and bargaining power to access certain licensing opportunities**.

Interviewees from the creative industries observed that while **CMOs** will be instrumental in facilitating and administering remuneration from AI training agreements—particularly in ensuring that **smaller creators** receive equitable compensation—**participation in such collective frameworks should remain voluntary**. In practice, even where **large rights holders groups** negotiate licensing agreements with AI firms, **CMOs may be necessary** to ensure the fair and transparent distribution of payments to **individual authors and performers**. However, stakeholders emphasised that any collective licensing model **must preserve the rights holders' ability to opt in**, rather than imposing **mandatory participation schemes**.

An impact of the increasing presence of content aggregation platforms is that they sometimes leverage subscription-based licensing regimes (as opposed to a one-off negotiated licence). Subscription models are valuable for AI developers as they allow access to a dynamic pool of training data, as the aggregator's content repertoire expands, while rights holders are provided with a potential ongoing revenue stream.

Some content aggregators specialise in aggregating content from user-generated-content (UGC) platforms such as social media networks. This represents a new revenue stream for online content creators who, in the past, may not have been commercially oriented (such as high-volume-posting non-commercial social media users). For example, it has been reported that Troveo (a content aggregation and AI training licensing platform), has processed around 1 million hours of video (from 1,300 licensors), twenty-five percent of which has come from YouTube, TikTok, and Instagram creators ([140]). The market for monetising UGC content in this way may be affected by whether social media platforms licence users content for AI training

([140]) How Creators Are Licensing Content to Train AI Video Models (Paywall), Variety, VIP+ Variety Intelligence Platform, 14 March 2025 (accessed 15 March 2025).

themselves, and or offer tools for their users to opt-out their content (or even licensing opt-in opportunities).

2.4.3.8 Mutual Access to Assets and RAG

Some direct-licencing agreements appear to include **access to AI developers' technical assets and services through APIs** as part of the counterpart for content usage.

This approach seems to be part of some agreements with online press publishers in an attempt to explore new ways to engage with their readers. These forms of technical capacity counterparts may also be attractive for academic and scientific content publishers seeking to **develop new interfaces and tools for researchers and readers** who use their content platforms. A key challenge for press publishers and rights holders of literary works is that many leading LLMs have already been trained on a large corpus of text mined from the open internet, with part of their content already included in widely distributed datasets. This may undermine their negotiating position for licensing their content for the purpose of AI training. On the other hand, high-quality academic, scientific, and news content is often behind subscription paywalls, and cannot be text and data mined under Article 4 CDSM as the 'lawful access' pre-condition is not met.

The negotiating position of such rights holders may also be strengthened by demand for up-to-date factual content, specifically in the field of scientific, academic, and news content. The emergence of RAG (see *Section 4.1.2*) technologies which provide AI developers an alternative to frequently retraining or re-fine-tuning models also provides rights holders with a **new form of authorisation to pursue through licensing**. This results in an emerging demand for direct licencing not just for training data but for RAG deployment, which is specific for this sub-sector of literary works at this stage.

2.4.3.9 Linguistic Framing of Agreements

An interesting point raised in stakeholder interviews is the importance of **language framing in licensing agreements**. There is evidence that at least some major tech companies refer

to agreements with copyright owners as ***'data access agreements'*** or simply **'*partnerships*'** rather than *'licensing agreements'.* This terminology may reflect that large AI companies are **reluctant to explicitly acknowledge** that the use of copyright protected works for AI ingestion is a copyright-relevant act. This framing is consistent with the public claims from major AI developers that their training is based on use of 'publicly available information', and that ingestion of data is not necessarily a copyright-relevant act. In contrast, there is a tendency to frame agreements to access data for RAG purposes as 'licences', which may reflect on the fact that through the AI ecosystem, a **broader consensus that RAG is a copyright relevant act** (as opposed to standard model training) is emerging.

2.4.4 Pricing Dynamics

While a growing number of rights holders are positioning themselves for potential licensing negotiations, the licensing market is still undermined by **opaque pricing signals**. This is common to many new markets in their early stages, such as copyright licensing for user-generated content, streaming, and certain forms of collective management. At this early stage, the exact terms for direct-licencing agreements between rights holders and AI developers are not publicly known, so the market lacks reference points and benchmarks for the terms of such agreements.

Nevertheless, on the issue of pricing dynamics in training data licensing markets generally, a number of key issues should be considered.

2.4.4.1 Market Rates and Annotation Costs

While specific market rates for training data assets are not known, some reference points have been disclosed through investigative journalism sources. A Reuters article has reported that image hosting platform Photobucket discussed proposed rates of $0.05 - $1 (USD) per photo (with price varying depending on licensee and types of images), while stock image platform Freepik licensed the majority of its archive of 200 million images at $0.02 - $0.04 (USD) per

photo ([141]). Reuters also cites one content licensing intermediary that claims AI developers are willing to pay $1 - $2 per image, $2 - $4 for short-form video, and $100 - $300 per hour of longer films. This source also claims that the market rate for text is $0.001 per word. However, certain types of sensitive content which need to be handled carefully and used for training GenAI filters (such as images of nudity) may cost $5 - $7 per image. In a similar vein, a Bloomberg article claimed that Adobe purchased video clips for AI training at an average rate of $3 per minute ([142]). Another source notes that *"the market hadn't yet settled into a standard, though reported figures have ranged $1 to $2 or as high as $6 per minute of video"* ([143]). Video aggregator licensing platform Calliope however lists out a price of $6.25 per minute for high-definition video content (with an additional premium for 4K or 3D content) ([144]).

In addition to licensing content from rights holders, an AI developer may **also have to pay significant costs in data labelling and annotation**. As a benchmark example, Amazon SageMaker Ground Truth (an Amazon service for building training datasets for machine learning) has published recommended prices for using a crowdsourcing platform operated by Amazon Web Services for data labelling services ([145]) at $0.012 per object for basic image classification, $0.012 for text classification, $0.036 for boundary box labels, and $0.84 for semantic segmentation (in addition to a $0.08 per-object per-month charge for under 50,000 objects).

Therefore, data labelling and annotation can in some cases cost more than the costs of licensing the unlabelled training data itself. This suggests a **potentially valuable commercial market for rights holders**, for not just licensing their works for AI training, but also provide **data annotation and metadata (at in-house or content aggregator level)**, to extract greater economic value from licensing agreements. This potential is linked to the observation in *Section 2.4.3.6* that new licensing markets may be driving digitisation and cataloguing efforts within rights holders organisations. The development of **industry-defined dataset standards**

([141]) Inside Big Tech's underground race to buy AI training data, Reuters, 5 April 2024 (accessed 14 March 2025).

([142]) Adobe Is Buying Videos for $3 Per Minute to Build AI Model , Bloomberg, 10 April 2024 (accessed 14 March 2025)

([143]) How Creators Are Licensing Content to Train AI Video Models (Paywall), Variety, VIP+ Variety Intelligence Platform, 14 March 2025 (accessed 15 March 2025).

([144]) Training AI With TV & Film Content: How Licensing Deals Look (Paywall), Variety, VIP+ Variety Intelligence Platform, 6 August 2024 (accessed 15 March 2025).

([145]) Human in the loop – Amazon SageMaker Ground Truth Pricing, AWS (accessed 14 March 2025).

for specific content sectors could significantly enhance these opportunities. This is supported by interviews with AI developers that suggested that a lack of standardisation in data labelling and dataset structures leads them to prefer the licensing of raw data.

2.4.4.2 Tokenisation as a pricing metric

A core issue is the **basis on which remuneration is calculated and based**. The value of work when used as training data can be a function of the **quantity of data within the work** that may be the basis of extracting information and correlations. This is unlike traditional markets for licensing a large-repertoire of copyright protected works, such as CMO blanket licences where remuneration is based often on a **per-work per-use basis**. For example, for a musical works performance rights organisation, the remuneration for a similar use by the same user does not depend on the length of the musical work (i.e., a three-minute song performed on the radio does not necessarily attract a different royalty rate than a four-minute song).

In the case of AI training data however, copyright-protected content is dissected into tokens meaning that larger works (and works embodied in higher-resolution formats) translate into a **larger number of tokens and are inherently more valuable as training inputs**. Thus, in training data markets it is possible that norms for licensing emerge which frame **tokenisation as a pricing metric**.

If and when the terms of major licensing agreements for AI training data become known to the public, stakeholders will be able to observe trends in the basis of remuneration and revenue-share distribution. This basis may have an impact on the way market pricing for licenses emerge and evolve over time, and the relative commercial value of different types of works.

2.4.4.3 Impact of International Legal Developments

While the focus of this analysis is the interface between GenAI deployment and copyright within the EU, both the GenAI value chain and content industries are highly internationalised. As many of the major players in the AI ecosystem are American companies, **developments**

in the USA market may have effects on EU licensing markets. This is amplified by the fact that the rate of litigation between rights holders and AI developers regarding training data is higher in the USA than it is in the EU.

As noted in *Section 2.3.2*, one of the key issues that may affect the USA market going forward is how the 'fair use defence' will be interpreted in the case of AI training. This creates an uncertainty in the US market that is likely to reduce once legal precedents are set through case law. This uncertainty is an important factor in direct licensing negotiations, as both rights holders and AI companies position themselves based on their levels of risk aversion and expectations for a legal precedent.

If USA courts set precedents ruling out fair use of copyright-protected works for AI training, they will have to make determinations on the remedies awarded to rights holders. A notable feature of the American IP jurisprudence is that injunctive relief is based on the principle of equity which require consideration of balance of hardships and effects on the public, and injunctions are neither automatic nor as commonly granted as in the past (Samuelson, 2021). In the instance that rights holders are successful in their litigations against AI developers, the basis of an award of damages by a court could indirectly create benchmarks for licensing remuneration, especially if some form of ongoing reasonable royalties are granted without injunctive relief (amounting effectively to a judicially-granted statutory licence).

As previously noted, general-purpose AI developers who train their models outside of the EU must still adhere to the AI Act's provisions on copyright compliance once their models are placed on the market in the EU. Questions may arise on whether AI developers whose models are trained in the USA would be able to claim compliance with EU law based on judicially-determined remuneration granted to rights holders in the USA. More importantly, given the internationalised nature of the GenAI value chain, judicially-determined remuneration in one jurisdiction – especially a major market like the USA – may serve as a remuneration benchmark for direct-licencing in the EU market. This issue may be further complicated by another unique feature of USA copyright law - statutory damages for infringement - which could potentially delink damage awards from estimated market value ([146]).

([146]) 17 U.S. Code § 504.

2.4.4.4 Types of Licences

A review of several licensing pricing from content aggregation and licensing platforms shows that it can depend on different factors, including:

- ***Volume:*** lower per-work licence rates based on higher volumes of licenced content
- ***Resolution:*** higher licencing rates based on resolution (particularly for video and images)
- ***Augmentation:*** higher licencing rates for additional access to variations for content (e.g., zoomed, inverted, or colour variations for images)
- ***Tag Modification:*** premiums for the ability to customise tags and labels

Furthermore, several licensing platforms base licensing **prices on the specific use case**. In some cases, prices are differentiated for use of content for training general purpose AI application or generative AI (which comes at an additional premium). Price and licensing terms may also be differentiated **for using training data for AI systems generating synthetic data**. A premium is sometimes charged for this specific use case, possibly to account for the fact that synthetic data would be used as a partial substitute for real-world (or human created) content in the training process.

Despite the lack of public information on licensing terms, there are indications of rights holder-led efforts to develop **standardised licensing approaches** for content used in AI training. For example, Dataset Providers Alliance (DPA) is a consortium of data aggregators and AI licencing intermediaries (including Rightsify, Global Copyright Exchange (GCX), vAIsual, Calliope Networks, ado, Datarade and Pixta AI) from different content sectors. Part of the DPA's mission is to *"Promote transparency and standardization in the licensing of intellectual property content for AI and ML datasets"*([147]). It has published a position paper on AI Data Licensing, which foresees a number of licensing models, specifically:

- **Usage-Based Licensing:** Fees based on the volume of data used and the scale of AI model deployment

([147]) See Data Providers Alliance website (accessed 15 March 2025).

- **Outcome-Based Licensing:** Royalties tied to the commercial success of AI models trained on the data
- **Subscription Model:** Tiered access to datasets with regular updates and support
- **Hybrid Licensing:** Combining upfront fees with performance-based royalties
- **Domain-Specific Licensing:** Tailored terms for different industries (e.g., healthcare, finance, entertainment)

The DPA also seeks to endorse defined dataset standards, specific to content sectors. This includes an 'Image Dataset Standard' based on the International Press Telecommunications Council Photo Metadata Standard (IPTC), and a 'Music Dataset Standard' called 'BigMusic' proposed by Rightsify.

2.4.4.5 Contractual Periods

A challenge with direct licensing agreements for AI training is the interpretation of **standard contractual concepts** such as length of contractual periods and termination. Some interviewed rights holders groups have expressed concerns over the potential interpretation of such terms in existing direct licensing agreements. Copyright protected content is licensed to AI developers to train models, but data may be subsequently used to (re)train future versions of models. Thus, a concept of 'subsequent training uses' might be a more practical concept than traditional time-defined contractual periods. Concerns have also been expressed on how the concept of **'contract termination'** should be interpreted once licensed data has been ingested and incorporated into the functionality of a model. Rights holders pricing decisions may need to consider the **value of data for** initial model training and for **potential memorisation and recursive learning**.

2.4.4.6 RAG Snippet Length

As noted previously, a key dimension of direct licensing agreements between text publishers and AI providers is **reciprocity**. These agreements often ensure that AI generated answers

to users questions cite and link the original sources on which answers are based, driving traffic to the licensor's original online locations. While the extent of end-user click-through rates depends on content types, a key variable in licensing terms is the maximum length of the content extract (or snippet) that can be used. An **inverse relationship may exist between snippet length and click-through rates**. This dynamic is important for press but also for science and academic publishers. Longer snippet lengths may justify higher licensing fees but may reduce the users' interest in consulting the original source, resulting in lower click-through rates and lower traffic to a provider's own content services. Shorter snippet lengths may justify lower licensing fees but may increase click-through rates and traffic to a provider's content services. Therefore, an important pricing factor for rights holders, particularly in the context of RAG, may be optimising the allowed snippet lengths within their comprehensive business model. An important pricing factor for rights holders, in the context of RAG, may be leveraging the allowed snippet lengths to optimise revenue derived from licensing their content services.

The importance of snippet lengths, outside of any direct licensing agreement, is illustrated by some search engine providers with AI-driven retrieval and snippet capabilities, introducing measures for webmasters to control snippet length. For example, Microsoft Bing allows websites to define maximum-text-lengths of snippets in search results, using robots-meta-tags.

2.4.4.7 Monetising Data Governance Expertise

An interesting concept that has been raised by some stakeholders (particularly in the libraries and archive sector) is that direct licensing agreements between rights holders and AI developers can also cover the **sharing of expertise in data management**. Large rights holders agencies and AI developers both employ data scientists. Data scientists on the content provider side have specific experience and expertise in **data stewardship and curatorial ethics**. Given the increasing demands from AI companies in terms of data governance obligations, data governance knowledge is a valuable asset in a direct licensing agreement.

2.4.5 Input-Output Licensing Linkages

The use of works as training data is quite often single use, although it may be the case that direct licences with rights holders are framed as giving an AI developer (or other TDM agent) permission to reproduce the work as part of a TDM process. Once information and correlations are extracted from the works and used to train a specific AI model, the works are - in many instances - no longer needed. The basis for the TDM licence may therefore be a **one-time payment for authorised use**, rather than successive payments based on use or, periodic payments for use over a prolonged period of time. As discussed previously, ongoing payments are the norm for RAG licensing, which is a separate commercial and technical concept.

Some rights holders are positioning themselves to negotiate remuneration past the one-time use of their works for training purposes, including seeking **remuneration linked to GenAI output**. Two examples are outlined below.

2.4.5.1 Musical AI Example

Musical AI (formerly Somms.AI), a platform used for licensing music to GenAI providers focusing on audio generation, has a unique business model. Musical AI secures agreements with owners of sound recordings (phonogram producers) or other intermediaries such as digital distributors (with whom phonogram producers have agreements). It then aims to license the authorised catalogues to AI developers who use these sound recordings as training data. When this results in an AI system being deployed in the market, the licensing agreement in place requires reporting of generated content made by the GenAI system. Musical AI claims that it has a proprietary software system to determine how the generative outputs may be attributed to specific training inputs ([148]).

In an August 2024 press report of an agreement made between Musical AI and digital music distribution platform Symphonic, it was claimed that *"Licenses made between AI companies, Musical AI and Symphonic will vary, but ultimately that license will stipulate a certain percentage of revenue made will belong to rights holders represented in the dataset. Musical AI will create an attribution report that details how each song in the dataset was used by the*

([148]) How it Works , Musical AI (accessed 14 March 2025).

AI company, and then AI companies will either pay out rights holders directly or through Musical AI, depending on what their deal looks like." ([149]) This licensing approach is comparable to the agreements used by some music streaming platforms, where a portion of the revenue generated is distributed among rights holders based on the number of times their content is streamed.

2.4.5.2 GEMA Example

As noted before (*Section 2.3.1.3*), German music CMO GEMA has recently filed legal action against OpenAI. GEMA has explicitly indicated that one of the purposes of the lawsuit is *"...to specifically refute the AI system providers' contention that training with and subsequent use of the generated content is free of charge and possible without the rights holders' authorisation. GEMA wants to establish a licence model on the market in which systems training using copyrighted content, generation of output based on that and the further use of the output must be licensed."* ([150])

In September 2024, GEMA introduced its licensing model for GenAI based on two components. First, GEMA seeks to ensure that its members participate in all economic benefits of AI providers, with the model setting a standard royalty rate of 30% of net income of the GenAI service provider, with a minimum royalty related to the amount of AI output produced. Second, GEMA seeks to ensure that its members participate in the economic benefits arising from the subsequent use of generated music (at least to the extent that would apply if the music were human-created). GEMA states that this model creates a reliable licence basis for both training and subsequent use of generative content ([151]).

([149]) Symphonic Opens Up Catalog to Train AI Models Through Musical AI Partnership, Billboard, 20 August 2024 (accessed 14 March 2025).

([150]) Suno AI and Open AI: GEMA sues for fair compensation, GEMA, January 2025 (accessed 14 March 2025).

([151]) Two components - one goal: Music creators shall receive fair shares through effective AI licensing, GEMA, 17 October 2024 (accessed 14 March 2025).

2.4.5.3 The Shapley Royalty Share Framework Model

A study by Wang et al. (2024) suggests a framework for a mutually beneficial revenue-sharing model between AI developers and copyright holders. A major challenge in developing a **revenue-sharing model for generative AI** lies in the "black-box" nature of model training and content generation, making the traditional, straightforward pro rata methods unsuitable (Wang et al., 2024). For this reason, the main contribution of Wang et al. was to introduce the **theory of Shapley value** to compute the contribution of each right holder to the GenAI model's output ([152]).

This method has been tested producing positive results, including:

- Indicating a higher contribution for copyrighted sources which styles closely resemble the output against a specific prompt;
- Indicating a higher contribution for copyrighted sources specialised on domains closely related to the output against a specific prompt;
- Indicating the hierarchy of contributions among the data sources,

However, it still has some computational problems, including the need to retrain the model more than one time. The authors state that this framework fits better when the model is trained by involving few copyright holders and they recognise that further research is needed.

2.4.5.4 EKILA Example

Another example is EKILA that is developing a synthetic media provenance and attribution for generative art (Carlini, Ippolito, et al., 2023), while integrating several innovative features to tackle attribution and compensation in generative AI. At its core, it uses the **C2PA** (*Section 4.3.1.1*) standard to embed **detailed metadata into synthetic images**, enabling users to trace their origins back to the generative model and specific training data.

([152]) The "Shapley value" is a concept from game theory that fairly distributes the total gains (or costs) among cooperative participants based on their individual contributions to the overall outcome.

The framework also leverages **Non-Fungible Tokens (NFTs)** ([153]), extending their functionality beyond simple ownership to include usage rights and attribution. This allows dynamic licensing and supports automated royalty payments through tokenised rights tied to smart contracts. A standout feature is EKILA's advanced **visual attribution model**, which identifies the specific training data most responsible for a synthetic image, outperforming existing approaches like CLIP ([154]). Additionally, EKILA supports dynamic ownership updates by integrating **blockchain-based** NFTs, ensuring that provenance is maintained even as assets change hands.

The **decentralised** structure of the blockchain reduces reliance on centralised authorities, enhancing trust and resilience for users. However, there are still some **drawbacks**, including the complexity of the implementation and general scepticism about blockchain technologies and cryptocurrencies. In addition, since the attribution is based on correlation (i.e., computed similarity between images) rather than direct causation, it may deliver some fairness issues.

2.4.6 Linkages between TDM users

Discourses on TDM exceptions tend to assume that CDSM Article 3 and Article 4 exceptions are fundamentally different in their objectives, beneficiaries, and policy bases, and do not intersect.

This assumption can be questioned in view of the complexity of the training data market and the fact that while some AI providers engage in their own TDM, many upstream TDM users create training databases which are then licensed for AI training. If various users can handle upstream database development under different institutional frameworks, Articles 3 and 4 may provide alternative methods for creating TDM-derived AI training datasets.

([153]) A Non-Fungible Token (NFT) is a unique digital asset stored on a blockchain that represents ownership of a specific item, such as artwork, music, or other digital content. Tokens are unique identification codes created from metadata via an encryption function. These tokens are then stored on a blockchain, while the assets themselves are stored in other places. The connection between the token and the asset is what makes them unique. Unlike cryptocurrencies, NFTs are indivisible and cannot be exchanged on a one-to-one basis, making them ideal for verifying the authenticity and provenance of digital creations. Cryptocurrencies are tokens as well; however, the key difference is that two cryptocurrencies from the same blockchain are interchangeable—they are fungible. See Non-Fungible Token (NFT): What It Means and How It Works, Investopedia (accessed 29 November 2024).

([154]) Contrastive Language-Image Pretraining (CLIP); see the *Glossary* for more details.

While research organisations benefit from a broader TDM exception, they may have relationships with other types of users within the TDM ecosystem. A research organisation is explicitly defined in the CDSM Directive as an entity whose research activities are conducted either (i) on a not-for-profit basis, (ii) by reinvesting all the profits in its scientific research, or (iii) pursuant to a recognised public interest mission ([155]). However, the research organisation itself does not necessarily have the technical capacity to undertake TDM, and may rely on private partners to carry out such technical activities. According to CDSM Recital 11, *"While research organisations and cultural heritage institutions should continue to be the beneficiaries of that exception, they should also be able to rely on their private partners for carrying out text and data mining, including by using their technological tools."* Thus, it is important to distinguish between **the *purpose* of the TDM activity**, and **the *entity* carrying it out**, as, in principle, this is the basis on which the applicability of the Article 3 exception should be determined.

While a research organisation can rely on a private partner to carry out TDM on their behalf on the basis of Article 3, a private commercial entity cannot benefit from Article 3 by delegating its TDM activities to a public research institution, once the TDM is conducted pursuant to a commercial purpose. The commercial entity must thus rely on the, relatively, more restrictive Article 4 for commercial TDM. However, this distinction between the *purpose* of TDM which differentiates Article 3 from Article 4 does not translate into a **differentiation of the *use* of the outputs that result from TDM**.

This relates to what some commentators have alleged to be a form of **'data laundering'** (Jiang et al., 2023). With commercial AI developers liaising with academic and other research organisations to undertake TDM in order to benefit from the wider legal exceptions applying to these non-profit organisations. AI developers may be tempted to resort to such data laundering practices as TDM under Article 4 may be more costly and less valuable with the need to incorporate mechanisms to respect opt-out, and to not use or license content effectively opted out.

The CDSM sets out a definition of 'research organisation' that is broad and acknowledges the diversity of forms, operation structures, and mandates that might characterise such

([155]) Directive (EU) 2019/790 Article 2(1).

organisations ([156]). Furthermore, it acknowledges that research organisations – while public or non-profit in orientation – may have **institutional relationships with private entities**, but this should not be associated with preference access to research or decisive influence on the organisation ([157]).

The LAION case brought these issues to light, as LAION was deemed by the Hamburg Court to benefit from the broader Article 3 exception for scientific research purposes, even though the datasets it developed were used for downstream commercial purposes (see *Section 2.3.1.1*) and it was funded by private organisations like Hugging Face and Stability AI (Schuhmann et al., 2022). The Court of Hamburg ruled that LAION qualified as a research organisation as the purpose of its TDM activities were directed towards the generation of new knowledge. Importantly, the court found that LAION was to be considered a research organisation despite its funding and organisational structure, because external commercial interests neither had a decisive influence, **nor benefited from preferential access** to its research results. The court also took into consideration the fact that LAION chose to **openly license its dataset to all potential users**.

This underscores several points about the training data market. First, clear distinctions need to be made between the **TDM undertaken by dataset providers** that may benefit from Article 3 or 4 depending on their institutional settings and the **TDM undertaken by AI commercial developers during model training**. While upstream dataset development may be carried out on scientific research basis (e.g., on the basis of CDSM Article 3), this does not necessarily mean that the same would apply when TDM is practiced by a commercial AI developer (who may need to rely on CDSM Article 4).

The challenge is that downstream Article 4 TDM would require filtering out opt-out protected works from the dataset before use. However, since a different entity carried out the initial TDM through which the dataset was developed, and this was done under Article 3 which does not require consideration for rights holders' opt-outs, the dataset does not necessarily contain the required information for the commercial AI developer to filter and use the data in compliance with Article 4. Another question would be whether AI providers have lawful access to the datasets, when these have been made publicly available online. In such cases, it is important

([156]) Directive (EU) 2019/790 Recital 12.

([157]) Directive (EU) 2019/790 Article 2(1).

to note that the act of making available to the public is not covered by the TDM exception. In this context, concerns were raised by rights holders on the AI Act's disclosure obligations on the sources of training data that apply to GPAI model providers, but not to upstream database developers who are themselves not model providers. The result is that moving from database development based on Article 3 to AI training based on Article 4 is associated with **potential liability for the developer**, and a precarious position for the rights holders. Addressing these issues may pass by practices and procedures which allow non-profit research databases to be used for commercial purposes without bypassing rights holders' rights reservations.

This may also require a better understanding on the **relationships between research organisations and commercial funders**, to ensure that the Article 3 exception is not used as a legal basis for TDM by entities not meeting the definition of 'research organisation' (either because of commercial entity decisive influence or preferential access). The intentional misuse of Article 3 by bad actors, or unintentional misuse by downstream database users who may be unaware of the data provenance, have the potential to undermine the effect of Article 4 reservation mechanisms, and the underlying development of a data licensing market by rights holders.

In addition, as previously noted, the Article 3 mechanism creates a different relationship of duties and responsibilities between the TDM user and rights holders which stem from the fact that research organisations may need to retain copies of works for verification of scientific research results ([158]). This comes with an obligation to store the copies in a secure environment, and the possibility for rights holders to apply measures to ensure the security and integrity of their systems in the face of a potential high number of requests to access them. The CDSM Directive also sets out that Member States themselves have a role in crafting best practices for system and database security when copies of works are stored by research organisations beyond the TDM process itself ([159]). The development of such good practices may help ensuring that the storage copies of works by research and cultural organisation under Article 3 do not undermine rights holder's efforts to restrict access and license the use of their content under Article 4.

([158]) Directive (EU) 2019/790 Recital 13.

([159]) Directive (EU) 2019/790 Recitals 15 and 16, and Article 3.

2.5 Mapping the GenAI Ecosystem

The above sections have identified and analysed the respective roles of the various actors in the GenAI ecosystem. In this section, the term **'GenAI ecosystem'** refers to the various legal and commercial stakeholders in the development and use of GenAI technologies, as well as the legal, commercial, and institutional relationships that arise from their interactions, and that are summarised in *Figure 2.5-1.*

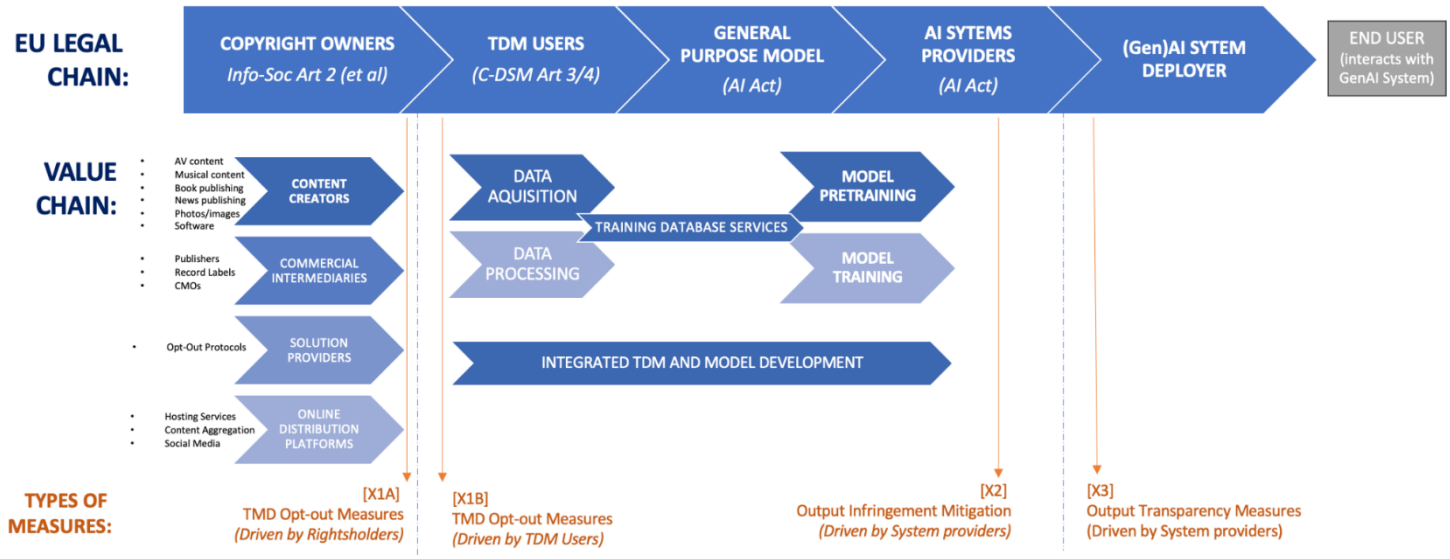


Figure 2.5-1: Mapping of GenAI Ecosystem.

The **top layer** ('EU Legal Chain') lists the entities that form part of the value chain, based on the relevant EU legal provisions applying to their activities, including:

- **Copyright owners** on works to which the exceptions for TDM Activities under CDSM Articles 3 and 4 apply.

- **TDM users**, who undertake text and data mining in accordance with CDSM Articles 3 and 4. As explained they may be AI developers themselves, or upstream dataset providers.

- **AI system providers**, While the AI Act distinguishes between 'General Purpose Model Providers' and 'Providers of AI systems', this distinction is relevant for this study's analysis only insofar as different legal obligations apply to these actors. For the purpose of understanding the AI value chain, both General Purpose AI model providers and 'Providers of AI Systems' are users of training data. While 'Generative AI' (GenAI) models or systems are not explicitly defined in the AI Act, this analysis considers them as a sub-set of General-Purpose AI Models.

- **GenAI system deployers**, providing GenAI services to the general public. Deployers may licence models / systems from upstream providers or may be model developers themselves.

- **End-users**, who are not explicitly defined in the AI Act but who are understood to be the natural persons who interact with GenAI systems once deployed.

The **middle section** ('Value Chain') details the different categories of actors throughout the value chain falling under categories defined in the top layer.

- **Copyright owners** may be subdivided into various subcategories based on the type of content they create including **audio-visual, music, book publishing, press publishing, photography and images, as well as software** content. A number of **commercial intermediaries** exist in each of these content sub-markets, such as publishers and record labels, who engage in production, financing and support of creative activities. Commercial intermediaries also include CMOs who manage certain exclusive rights on behalf of copyright owners, in specific content markets.

- '**Solution Providers**' are a new group of market actors that has emerged after the CDSM and is growing in importance with the development of AI. These are agents who provide technical assistance to copyright owners by developing tools and protocols for the rights reservation mechanism under CDSM Article 4. The developments of these solutions are the focus of Chapter 3.

- **Online intermediaries and web-distribution platforms,** hosting and making available digital content that can be text and data mined, such as websites hosting copyright protected works, social media, and content aggregation services, have an important role in this ecosystem and the practical implementation of TDM reservation measures.

- **Integrated TDM and model development,** covers the different functions that form part of TDM processes, with data acquisition, database development, and model training, that are **distinct subsections of the value chain**. In practice, there are economic agents that specialise in the development of datasets (data acquisition and processing) which are subsequently used by AI model/systems developers. Some AI model/systems developers integrate both steps by developing training dataset in-house ([160]).

The **bottom layer** of the diagram ('Types of Measures') identifies the various measures that are the focus of Chapter 3 and 4 of this report, and contextualises where measures are applied in the value chain. The focus is on three sets of possible measures, summarised in the table below.

<table>
<tr><td rowspan="2">[X1]</td><td rowspan="2">Measures related to managing copyright works within GenAI training input datasets. These measures are divided into two subsets.</td><td>[X1A]: Measures implemented by rights holders to exercise their rights to opt-out of TDM in accordance with CDSM Article 4(3).</td></tr>
<tr><td>[X1B]: Measures used by AI model developers (in particular general purpose model providers) to comply with rights holders opt-outs, as required by AI Act Article 53(1)(c).</td></tr>
<tr><td>[X2]</td><td colspan="2">Measures related to mitigating the risks of generating output which might infringe on third party copyright.</td></tr>
</table>

([160]) This practical distinction recognises (but takes no position on) the debate on which parts of the value chain fall within the legal definition of TDM (i.e., whether the TDM exceptions extend from the process to data acquisition, to database development, and actual model training).

[X3]	Measures implemented by AI system providers to ensure transparency in the nature of generative output (as required by Article 50 of the AI Act).

Table 2.5-1: Definitions of the categories of measures ([161]).

Building upon this mapping, the successive technical steps associated with both the input and output components of GenAI systems can be associated with the relevant measures that can be implemented at these different steps. The *Figure 2.5-2* provides a detailed representation of this approach, with each of the technical measures identified indicated as either **X1A, X1B, X2** or **X3**, in line with the mapping above.

Considering each technical step of the GenAI cycle, the relevant steps and associated measures are:

- The **input data collection**, which is performed by the datasets developers, must adhere to the opt-out reservations defined by rights holders through:
    - **Reservation measures (X1A)**;
    - Embedding into the digital assets via either **provenance tracking solutions (X3)** or **watermarking (X3)**;
    - Association with the digital asset using **fingerprinting-based solutions (X3)**.

From the point of view of the datasets developers, this process can be automated by using existing **libraries for parsing rights declarations (X1B)** (e.g., for correctly parsing robots.txt files). Moreover, some AI developers are offering **online services for rights holders to express their opt-out reservation (X1B)**, that are directly related to their data collection. If reservation expressions are not enough, website owners can install **crawler blockers (XA1)** to prevent AI scraping;

([161]) This categorisation should not be considered as very strict, as in some cases measures on the Input side may also apply to the Output side and vice-versa.

- During **input data cleaning**, that takes place as part of the data processing procedure and before the data is used for training, checks for the presence of **watermarking (X3)** or **provenance information (X1B)** and can prevent unauthorised ingestion of copyrighted content. Additionally, the filtering mechanisms can reference the information provided through **reservation solutions (X1A)**, specifically if they are **asset-based** (i.e., opt-out declarations tied to content identifiers rather than location-based (e.g., URL-based) exclusions).

- **Model pre-training** consists of training the **foundation model** that serves as a basis for the GenAI system. This phase requires vast amounts of data. If certain rights holders introduce **data poisoning techniques (X1A)** as a way to protect their works, AI developers may face significant challenges when developing their model based on such content. In such cases, they may need to filter out purposely poisoned data and, potentially, re-train the model from scratch. Additionally, this step can be influenced by model developers' adoption of **unlearning techniques (X1B), model editing (X1B)** and **revenue sharing (X1B)** techniques;

- **Fine-tuning** enables the evolution of a general-purpose foundation model into a more specialised model, optimised for specific tasks. This process is similar to pre-training, but relies on smaller, higher-quality datasets. Thus, it interacts with the same technical measures.

- **Reinforcement Learning** is an additional model adjustment technique designed to align the system's responses with human preferences, many times through reward-based optimisation. During this step, often requiring manual human intervention, the system's outputs are tested against specific input prompts. If these prompts include **poisoned data (X1A)**, then the procedure may be compromised. Moreover, **fingerprinting-based technologies (X3)** can be leveraged to adjust the model towards generating outputs having a certain degree of dissimilarity with existing fingerprinted works. These technologies can also be leveraged to address **model editing (X2)** or **unlearning (X2)** requests;

- When a model is deployed, it is integrated into a technical environment, such as a server or cloud infrastructure, to ensure stability, security, and accessibility for end users across different deployment scenarios. The model may be integrated with other

software solutions supporting content **provenance tracking (X3)**, **watermarking (X3)**, and output **filtering (X2)** to tag and control generative outputs. This integration effort may also support **revenue-sharing mechanisms (X1B)** to facilitate proportional revenue distribution among contributing rights holders;

- **Real-time inference** occurs during model output generation, integrating additional information beyond the model's training dataset. This data retrieval process may involve structured external databases or unstructured web scraping and is facilitated by Retrieval Augmented Generation (RAG) technologies. Because real-time inference involves active data collection, applicable copyright and **reservation measures (X1B)** must be observed. Additionally, real-time data gathered can be used for **revenue-sharing mechanisms (X1B)** and **output filtering (X2)**. AI-powered inference crawlers can systematically navigate web pages to extract relevant information. However, these crawlers can be **manipulated through data poisoning (X1A)** techniques, which introduce misleading or adversarial data to disrupt AI inference;

- Finally, during output generation, provenance **tracking solutions (X3)** and **output filters (X2)** may be applied, alongside secondary neural networks **facilitating model editing (X2)** and **unlearning (X2)** processes. **Watermarks (X3)** can be embedded into model outputs to identify AI-generated content. Additionally, **fingerprinting (X3)** and deepfake detection technologies can be used to assess similarity with existing works and detect potentially fraudulent activity.

These various technical steps and measures, typical of a GenAI development process, are summarised in the figure below.

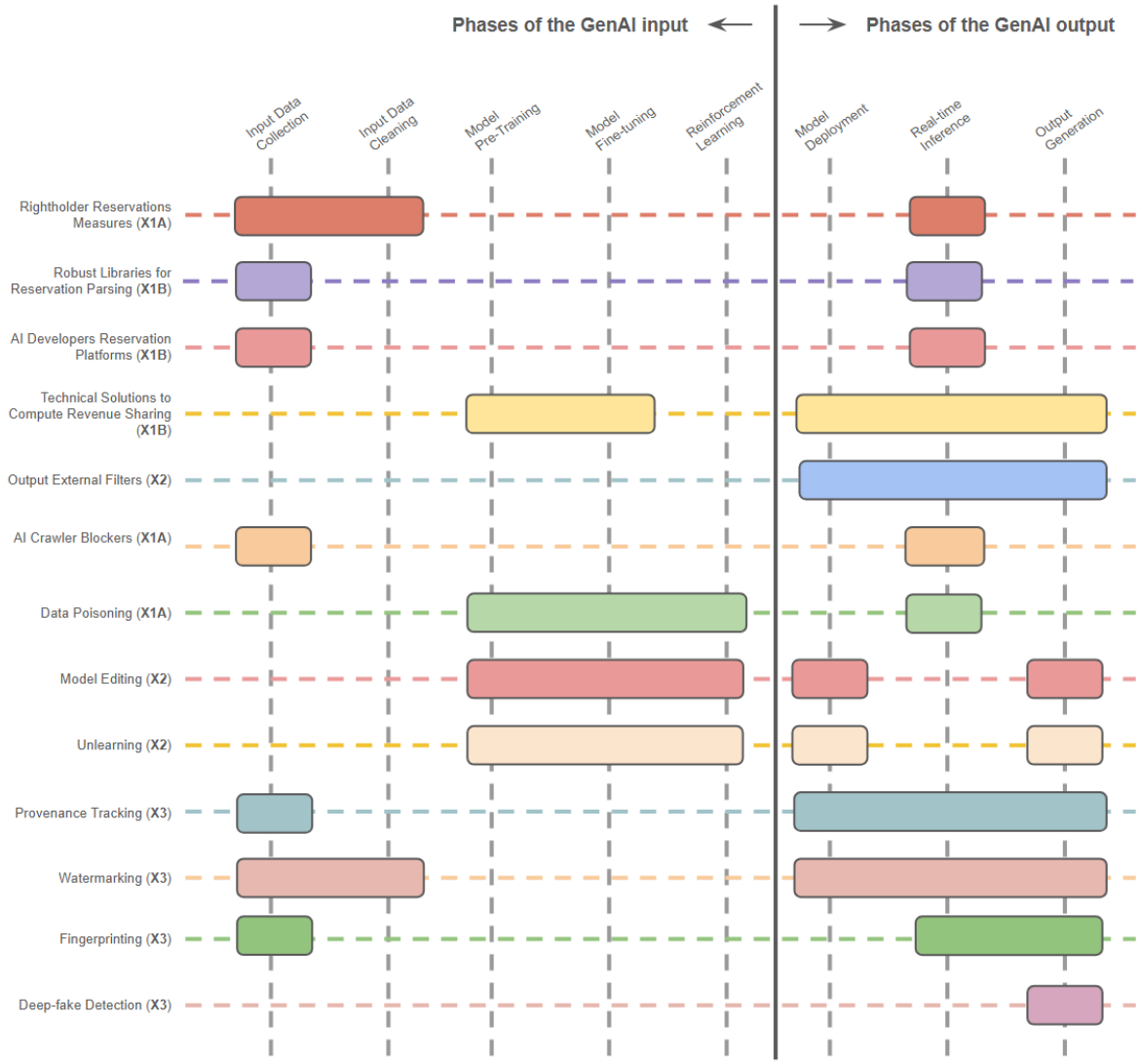


Figure 2.5-2: Mapping the GenAI Development Process.

3 Generative AI Input

3.1 Training methods and practices

This section provides an overview of the GenAI training process, focusing on the input of data. It builds upon the background provided on **foundation models** as discussed in *Section 2.1.2.2*. These general-purpose models are trained on very large datasets before going through a **fine-tuning** process to optimise the execution of specific tasks. Fine-tuning typically consists of making small changes to the model to achieve a desired output or performance, by utilising labelled, more specific data. Additional input data is often needed for the final phase of the training process, named **reinforcement learning**. After that, the final GenAI model (or GenAI system) is developed for potential deployment. *Figure 3.1-1* outlines the main steps involved in the training process. This section analyses each of these steps, highlighting differences between different types of training content (e.g., text, audio, images…) and indicating which tasks need to be performed by humans.

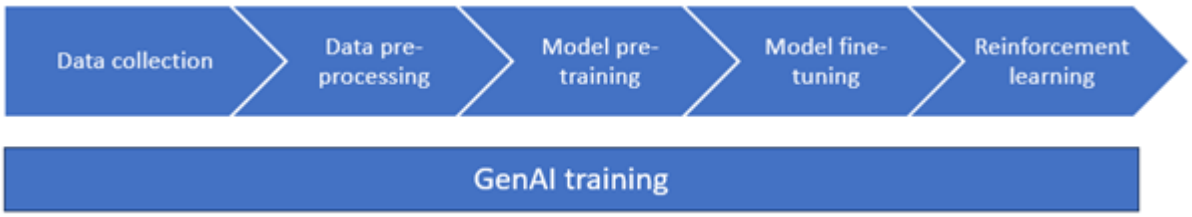


Figure 3.1-1: Main components of the GenAI training process

3.1.1 Training Data Schema

Data is available in a **variety of media types**, including commonly used formats like images, videos, text, PDFs/documents, HTML, audio, time series, 3D/DICOM, geospatial data, sensor fusion, and multimodal content.

A **training data schema** is the overall representation of labels, attributes, spatial information and their relation to each other. It is used to **encode the training data in a structured way** and to handle its complexity. Training data schema should be treated similarly to database schemas. Whatever the type of training data, it can be described through **labels and attributes** to map the human meaning to technical terms (Sarkis, 2023).

Data labelling, also referred to as **data annotation**, is the process of assigning target attributes to training data, thereby enabling a machine learning model to learn the expected predictions. This procedure constitutes a fundamental stage in the preparation of data for supervised machine learning ([162]). Public datasets designed for **AI training purposes** often include pre-labelled data.

Labels are the "top-level" of semantic meaning. In the base case, they represent only themselves. In most cases, though, labels organise a set of attributes. **Attributes** are mostly treated as strings and can also have constraints (Sarkis, 2023). Labels and attributes can also be assigned to specific **portions** of a single data item (e.g., a single image) using segmentation techniques.

When dealing **with specialised information** such as medical data, **accurate annotation usually requires specialised knowledge**. To ensure consistency, annotators rely on and maintain guides, which then define the training data. However, different experts may have different opinions on appropriate annotation decisions. Since there is some level of subjective judgement when it comes to technical data annotation, this may amount to a human-made intellectual contribution that could potentially attract some form of intellectual property protection (Sarkis, 2023) ([163]).

There is a trade-off regarding the **schema complexity**: machine learning derived from a detailed schema is smarter but more difficult to manage. For example, higher level schemas are required to prevent social bias: if the offensive data is not labelled, then it would be impossible to train a model to distinguish it. The media type may affect the trade-off greatly, with complexity rising progressively from text to images and then to videos (Sarkis, 2023; Publio et al., 2018).

([162]) How to Label Data for Machine Learning: Process and Tools, AltexSoft, 16 July 2019 (accessed 14 March 2025).

([163]) Intellectual property in AI, Gemmo.AI, 2022 (accessed 14 March 2025).

It is worth noting that some persons use the term "**metadata**" to refer to any form of annotation. However, metadata specifically refers to information about the dataset that is not directly utilised by the model. Examples of metadata include details such as the date of dataset creation or the identity of the creator. It is important to **distinguish annotations from metadata**, as annotations are an integral part of the primary training data structure (Sarkis, 2023).

3.1.2 Data Collection and Access

This section describes the different methods to gather input data for GenAI systems. As confirmed by interviews with AI developers, data collection often occurs **simultaneously from multiple sources**. These sources include **proprietary and public datasets, publicly available data, contracted APIs,** and **synthetic data**.

Inadequate provenance and attribution often originate during the early stages of data collection and annotation (GenAI input). As the process advances through model training and deployment, these issues tend to grow more complex and become harder to address effectively (Zhang et al., 2024). Thus, it is important to pay attention to how the processes of data collection and access happen.

In addition, AI service providers may not verify if the use of content by their client is copyright compliant ([164]).

3.1.2.1 Large Public Datasets

The latest wave of language models, both open source and proprietary, largely derive their abilities from the **diversity and richness of large training datasets**, including pre-training corpora, fine-tuning datasets compiled by academic researchers, data synthetically generated by models, and aggregated by platforms. Increasingly, widely used dataset collections are

([164]) For example, AWS emphasised that while it offers tools and 'responsible AI' guidelines to assist customers with data governance, it does not pre-screen customer-uploaded content for copyright compliance as this would involve monitoring customer workloads, violating AWS' core commitments to customer privacy and security.

treated as monolithic, instead of a **lineage of data sources**, scraped (or model generated), curated, and annotated, often with **multiple rounds of repackaging** (and re-licensing) by successive practitioners (Longpre et al., 2023).

The **Data Provenance initiative (Longpre et al., 2023)** has made an effort to improve transparency in this context. This research declared that:

> *"Notably, we find that 70%+ of licences for popular datasets on GitHub and Hugging Face are "Unspecified", leaving a substantial information gap that is difficult to navigate in terms of legal responsibility. Second, the licences that are attached to datasets are often inconsistent with the licence ascribed by the original author of the dataset—our rigorous re-annotation of licences finds that 66% of analysed Hugging Face licences were in a different use category, often labelled as more permissive than the author's intended licence. One especially important assumption in cases where datasets are based on data obtained from other sources is that dataset creators actually have a copyright interest in their dataset. This depends on the data source and how creators modify or augment this data, and requires a case-by-case analysis. Our empirical analysis highlights that we are in the midst of a crisis in dataset provenance and practitioners are forced to make decisions based on limited information and opaque legal frameworks." (Longpre et al., 2023)*

As a result, finding precise, publicly available information about the flow of data into the main GenAI training datasets is considerably challenging.

The following sub-sections highlight a selection of major public datasets and platforms distributing such datasets that are shaping the GenAI ecosystem.

3.1.2.1.1 Hugging Face

Hugging Face is a **platform hosting AI models and datasets**, where a wide range of users can download and upload both ([165]).

([165]) See Huggingface website (accessed 14 March 2024). During stakeholders' interviews (conducted in January 2025) it emerged that Hugging Face hosts over 1 million models and approximately 200 datasets (though this figure evolves). The self-governed community of users ranges from large corporate teams to individual developers and small research labs.

The platform fosters **open science** but **also supports licensing options** that may include usage restrictions, like a required user key or a terms-of-service acceptance from a third-party site. Users choose a licence or usage restriction for their uploads. The platform encourages thorough documentation in “Dataset Cards” and “Model Cards.” While Hugging Face tracks the raw download statistics, it **does not track how a dataset or model is used** (fine-tuning, commercial vs. non-commercial, etc.). Instead, Hugging Face adopts a “notice and action” approach. If a user **flags infringing data**, the dataset owner typically removes it or corrects the licence. If unresolved, the company’s moderation team can intervene.

Overall, Hugging Face provides **infrastructure and partial moderation** but does not consider itself an ‘enforcement agency’ for copyright.

3.1.2.1.2 Common Crawl

As mentioned in *Section 2.4.2* Common Crawl is the **largest freely available collection of web crawl data** ([166]) and a **foundational building block for LLM development**, and subsequently generative AI products built on top of LLMs.

LLM builders **train their models on filtered samples** of Common Crawl’s archive.

Typical filtering techniques include:

- **Keywords and simple heuristics**: It is common to remove pages that contain keywords considered harmful in the URL or anywhere within the page.
- **AI classifiers**: A reference dataset considered high quality (for instance OpenWebText2 ([167])) is used to **train a text classifier**. This classifier is used to filter out everything from Common Crawl that does not meet an adjustable similarity threshold (Baack, 2024).

There are a **small number of filtered versions that are reused frequently**, especially Alphabet’s *Colossal Clean Crawled Corpus* (C4) and EleutherAI’s *Pile-CC*. The most popular

([166]) See Common Crawl website (accessed 14 March 2025).

([167]) OpenWebText2 is a high-quality reference and contains the text of all URLs shared and upvoted at least three times on Reddit until April 2020.

filtered Common Crawl versions were **created by LLM builders themselves** as a step towards their actual goal: training LLMs. This **restricts the amount of time and energy that can be dedicated to the filtering effort**, and it means that the filtering techniques are not updated after the publication to take criticism and feedback into account.

Currently, there exists **no well-developed ecosystem of dedicated filtering intermediaries** that could be tasked with continuously filtering Common Crawl, in transparent and accountable ways (Baack, 2024).

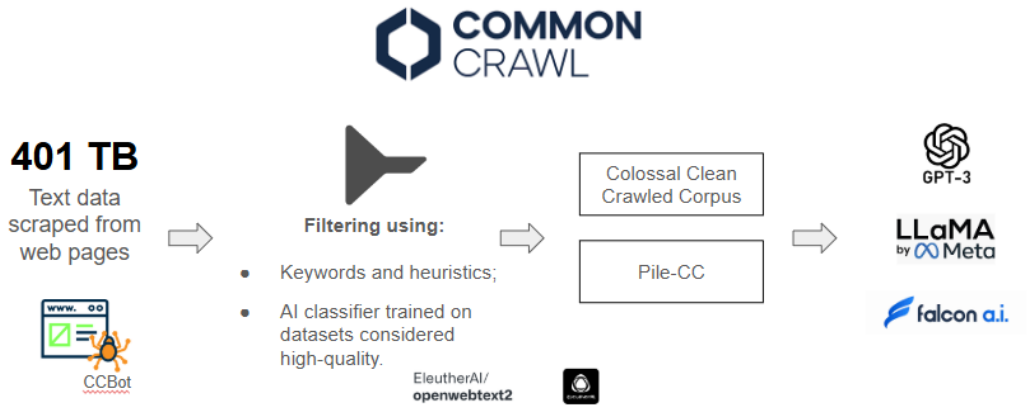


Figure 3.1.2-1: Schema of the data path from Common Crawl to the main foundation models trained using its content.

High-quality filtering is crucial as **Common Crawl includes material protected by copyright and related rights**. It has faced accusations, from entities such as **The New York Times ([168]),** that it is a "highly weighted dataset" in training models allegedly using copyright-protected content without authorisation ([169]).

([168]) The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work (Paywall), The New York Time, 27 December 2023 (accessed 14 March 2025).

([169]) Publishers Target Common Crawl In Fight Over AI Training Data, Wired, June 2024 (accessed 12 November 2024).

The non-profit organisation states in its Terms of Use ([170]) that it is willing to **remove any copyright-protected content** from its archive upon receiving a legitimate notice. However, the solution of **putting more effort into filtering might be preferable** when considering those points highlighted by Common Crawl defenders:

- The datasets are a valuable resource to track web history ([171]);
- Damaging Common Crawl to the point it is no longer useful as a training data source risks further empowering leading AI companies, which already have a scraping system (Baack, 2024).

As a result of the sharp rise in demand to redact data, Common Crawl's web crawler **CCBot** is also increasingly thwarted from accumulating new data from rights holders (see *Section 3.1.2.2* on Web Scraping).

3.1.2.1.3 Image Datasets

Datasets of images used for generative AI training may **include images directly or via URLs** ([172])**.** However, to enable text-to-image machine learning, the **contextual data** associated with these images must always be directly embedded within the dataset. This may include:

- **Metadata** provided by the image's creator (e.g., the creator's identification, the camera model used, and the location where the photo was taken);
- **ALT text** ([173]) from the webpage where the image was sourced;
- **Labels** assigned by dataset curators;

([170]) Common Crawl, 'Terms of Use' (accessed 12 November 2024).

([171]) Jeff Jarvis, a journalist professor and a Common Crawl defender, pointed out: "I'm very troubled by efforts to erase web history and especially news. It's been cited in 10,000 academic papers. It's an incredibly valuable resource".

([172]) Some examples of datasets which don't directly contain the images but only the **links** to them are *Open Images Dataset V7* and *Multimedia Commons*. They've been created respectively by Google and Yahoo and contain the references to 9 million and 99 million images and videos published with a Creative Commons licence on Flickr, an online photo and video sharing platform that allows users, often photographers, to upload, organise, and share media.

([173]) ALT text (alternative text) is a brief description of an image used to improve accessibility and provide context for those who cannot see the image and are using, for example, screen readers.

- General **text descriptions**.

When downloading datasets **referencing images through links**, the actual images aren't downloaded, only the URLs and associated metadata. Therefore, the dataset user is expected to retrieve the original image by following the link. The absence of the actual image in the dataset becomes evident, when the original image is deleted, the link becomes invalid ([174]).

There are also datasets that **contain the images** themselves. However, this is more common when the number of images is **not very large** ([175]).

ImageNet

The ImageNet dataset is organised using the **WordNet ([176]) hierarchy**. Each node in the hierarchy represents a category, and each category is described by a synset (a set of synonymous terms). The images in ImageNet are **annotated** with one or more synsets, providing a rich resource for training models for the recognition of various objects and their relationships ([177]). ImageNet aims to populate the majority of the 80,000 synset of WordNet with an average of 500-1000 images each (Deng et al., 2009). The content of ImageNet is **human-annotated**.

([174]) AIGen: Come Sono Fatti i Dataset Delle Immagini per l'addestramento (in Italian), AI4Business, 2 July 2024 (accessed 14 march 2025).

([175]) Ibid. Examples include the *Cityscape Dataset*, which is restricted to academic use, and the *Oxford-IIIT Pet Database* which contains approximately 7,000 images, a relatively small number compared to the datasets referencing images through links.

([176]) WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. WordNet is also freely and publicly available for download.

([177]) Ultralytics, 'ImageNet', accessed 20 December 2024.

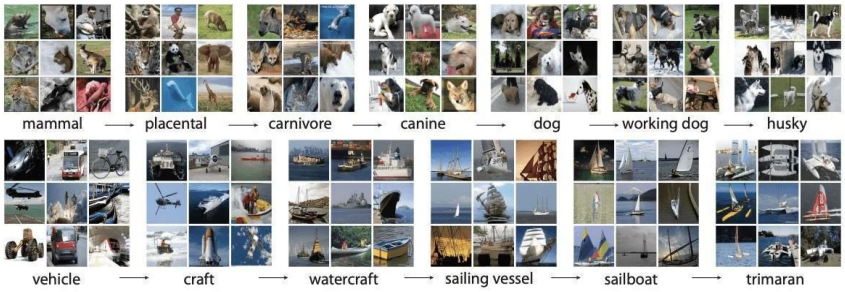


Figure 3.1.2-2: Example of labelled images according to the WordNet hierarchy ([178])***.***

ImageNet **does not own the copyright of the images**, instead it only compiles an accurate list of web images for each synset of WordNet.

LAION-5B

The LAION-5B dataset consists of approximately 5.85 billion text-image pairs indexed specifically for AI training. However, the term 'consists' is used loosely, as it **does not physically store these images but rather provides links to their original locations on the web**. These images were "collected" from the web, as specified in the project's FAQ: "LAION datasets are simply indexes to the internet, i.e., lists of URLs to the original images **together with the ALT texts** found linked to those images." ([179])

LAION-5B was constructed using distributed processing of the Common Crawl dataset.

Each record of the dataset contains the following fields ([180]):

- URL of the image;
- Text description;
- Picture height;
- Caption's language;
- **'pwatermark'**: probability that an image contains a visible watermark, determined using an internally developed watermark detector. The method reportedly employs

([178]) Ibid.

([179]) LAION, 'FAQ' (accessed 21 December 2024).

([180]) LAION, 'LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS' (accessed 26 December 2024).

computer vision techniques to assess visual artefacts typically associated with watermarking, such as opacity patterns or embedded text overlays;

- **'punsafe'**: probability of being an unsafe image, computed using the CLIP ([181]) based detector.

The current crop of image generators, primarily those based on Stable Diffusion, are pre-trained on LAION-5B or its variants. Although the datasets are not available for browsing, various artists have reported finding their works without their consent or attribution ([182]).

3.1.2.1.4 Other examples of Public Datasets

Books3 is a dataset containing 196,640 books in text format by authors including Stephen King, Margaret Atwood, and Zadie Smith, that is used to train language models. It was created in 2020 by open-source advocate Shawn Presser and made available as part of The Pile open source dataset for LLMs developed by EleutherAI ([183]). In 2023, the **Danish Rights Alliance** spearheaded an effort to remove this dataset from the internet, highlighting that some of the books included were sourced from websites that aggregate "pirated" content ([184]).

The Pile is an 886 GB, open-source dataset of English-language text created to help train LLMs. Developed by **EleutherAI** and publicly released in December 2020, it **consists of 22 smaller datasets**, including Books3, BookCorpus and YouTube Subtitles, plus 14 new datasets. Originally created to train EleutherAI's GPT-Neo models, The Pile has since been utilised in training numerous other models, including Microsoft's Megatron-Turing Natural Language Generation, Meta AI's Open Pre-trained Transformers, LLaMA, Galactica, Stanford University's BioMedLM 2.7B, and Apple's OpenELM ([185]).

([181]) Contrastive Language–Image Pretraining (CLIP); visit the *Glossary* for more details.

([182]) Using a tool built by **Simon Willison** which allowed people to search 0.5% of the training data for **Stable Diffusion** V1.1, i.e. 12 million of 2.3 billion instances from **LAION 2B**, artists have discovered that their copyright-protected images were used as training data without their consent (Baio, 2022).

([183]) Books3 AI training dataset, AIAAIC (accessed 13 November 2024).

([184]) Publishers Target Common Crawl In Fight Over AI Training Data, Wired, June 2024 (accessed 12 November 2024).

([185]) The Pile dataset, AIAAIC (accessed 13 November 2024).

The *World Intellectual Property Organization* (WIPO) ([186]) conducted a text mining analysis of the open access subset of the GenAI corpus (34,183 articles out of a total of 75,870) in an attempt to find the actual used datasets in this complex scenario. WIPO found that in the top 20 publicly cited datasets 14 were **image-based**, with detailed results shown in the table presented in *Annex VII*.

3.1.2.2 Web Scraping

The process of **'web scraping'** is a form of data collection used in TDM processes and is **central to the current AI ecosystem**. As commonplace as data scraping practices are, there is no single widely accepted definition. A broad definition, suggested by the *Organization for Economic Cooperation and Development* (OECD, 2025) stresses three general features of data scraping:

- **Automation** - Data scraping typically involves the use of software tools or scripts designed to quickly and efficiently harvest or otherwise aggregate data with minimal human intervention;
- **Scalability** - Data scraping is often used to collect or make accessible large amounts of data that would be impractical to aggregate manually. In addition, the tools and techniques employed can be scaled up to extract data from numerous sources simultaneously;
- **Lack of coordination** - Data scraping is often done without coordination between the data scraper and the entity hosting the data.

The OECD has advocated for an international 'data scraping code of conduct' which would set out voluntary guidelines for scrapers, data aggregators, and AI data users. Such a code could complement standard contract terms and standard technical tools.

To collect the vast amount of data used to create large datasets and to directly train models, a specific type of software called "**web crawler**" – or synonymously designated as "**bot**", "**agent**" or "**spider**" – has been used.

([186]) 'Patent Landscape Report - Generative Artificial Intelligence (GenAI)', chapter 1, WIPO, 2024 (accessed 16 November 2024).

Crawling and scraping, while technically similar, differ significantly in their purpose and their legal implications (see discussion in *Section 2.2.2*). Crawling typically involves the systematic browsing of publicly accessible web pages to index metadata and content for search engines. This process is generally considered less intrusive and primarily serves discoverability functions. Scraping, by contrast, involves the extraction of specific data or content, often targeting more granular information and reproducing portions of works in a way that is more likely to raise copyright concerns.

The crawlers programmed for web scraping are essentially the same as those employed for search indexing: they systematically explore the web by starting with a set of seed URLs and following hyperlinks to discover new pages. This similarity means that a content host **may not easily distinguish between a crawler used for search indexing and one used for GenAI data ingestion**; moreover, some crawlers serve both functions.

Theoretically, **a bot should always identify itself** when interacting with a website. However, interviewed content providers highlighted the growing challenge of managing **non-declarative bots**, which fail to disclose their presence. This issue imposes significant resource costs on content providers as they attempt to enforce their rights and protect their data.

AI crawlers can be identified like other bot traffic as they typically exhibit high bounce rates and low session durations, with the caveat that their traffic often originates from a subset of common IP addresses associated with GenAI vendors (Jiménez, J. & Arkko, J., 2024). However, a significant portion of AI-related crawling is likely for **real-time inference** (i.e., when GenAI models generate content based on information appositely retrieved after the user's input; See *Section 4.1.2* on RAG technologies for more details). While traditional crawlers are designed for **massive data retrieval**, the ones designed for **real-time inference** more closely resemble a **human user browsing the web**: those AI-enhanced crawlers can "understand" web pages, making them more difficult to be detected by crawler blockers (see *Section 3.8.2* on software for managing bot traffic). According to the AI detection startup **Originality AI**, more than 44% of the top global news and media sites block Common Crawl's CCBot ([187]). ([188])

([187]) Publishers Target Common Crawl In Fight Over AI Training Data, Wired, June 2024 (accessed 12 November 2024).

([188]) CCBot is not the only operating crawler: some leading AI companies, like **Alphabet**, **Microsoft**, **Meta**, and more recently also **OpenAI**, have their own crawlers to collect web data themselves (Baack, 2024). For example, Meta declares to train its GenAI models using data coming from two main sources: licensing agreements with some

In addition to not declaring their identity, other **problematic bot behaviour** exists, such as significant resource consumption for website owners ([189]), failure to collect all necessary metadata linked to copyrighted content ([190]), and/or scraping of "pirate" sites ([191]).

Figure 3.1.2-3 represents how some known crawlers can be subdivided into the categories of interest for this study.

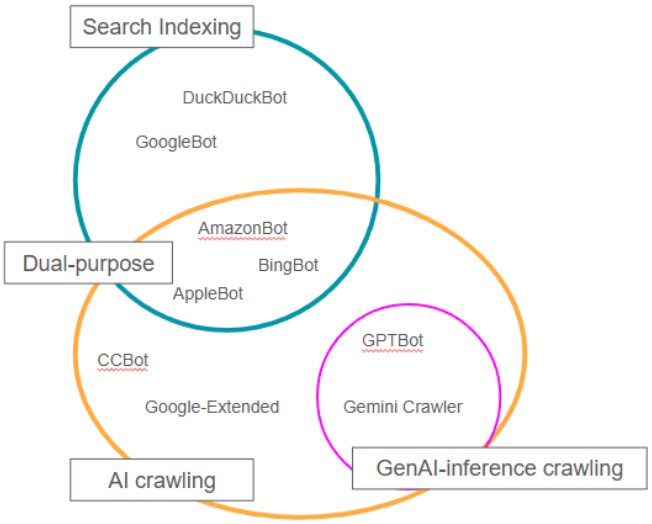


Figure 3.1.2-3: Classification of selected crawlers based on their purpose.

suppliers and public data crawled from the internet. The latter may also include some personal information, like a public post. See *How Meta uses information for generative AI functions and models*. Facebook – Privacy – GenAI (accessed 31 October 2024).

([189]) **Read The Docs**—a company offering a platform designed to simplify the process of building, hosting and managing software documentation, thus storing a **large amount of text data**—reported an increasing **abuse from AI crawlers**. They allegedly have cost them a significant amount of money and caused them to spend a lot of time dealing with abuse, with peaks of 10Tb of downloaded content in a single day resulting in an expense of about $5.000 due to the bandwidth charge.

([190]) As some interviewed stakeholders from civil society noted, one issue with crawlers is that they need to be better programmed to avoid scraping unauthorised content, but also to **ensure they collect all necessary metadata to accompany the content**. Failing to do so would severely limit traceability, complicate compliance efforts, and hinder accurate revenue distribution.

([191]) More copyright-related problems can arise when scrapers fail to identify **"pirate" sites**, such as web platforms that gather and make available large amounts of content without authorisation from the rights holders. It is unlikely that those platforms adopt any protection against AI crawlers. Thus, even if the software of the crawler is properly designed to respect reservation protocols, it may inadvertently scrape copyright-protected content from those "pirate" sites. Meanwhile, interviewed AI developers reported confidence in the data collection processes they have in place, declaring that "pirate" sites are not part of their training data sources.

3.1.2.3 Training on Synthetic Data

As an alternative to real-world data, models can be trained on synthetic data, which is **annotated** information typically **generated** by other GenAI models (which were trained on real-world data or pre-existing creative works). Synthetic data is designed to simulate the characteristics of real-world datasets. At a very high level, the process can be schematised as in *Figure. 3.1.2-4.*

Figure 3.1.2-4: Key stages in the workflow for training Generative AI models on synthetic data.

Some examples extracted from synthetic datasets are reported below.

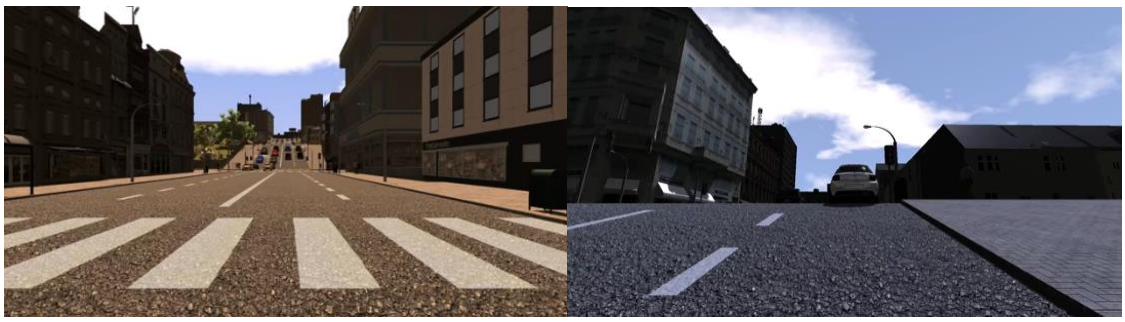


Figure 3.1.2-6: Image samples taken from the Synthia synthetic dataset ([192]), which has been designed to be used for the training of autonomous driving AI systems.

([192]) Download the SYNTHIA Dataset – The SYNTHIA Dataset, The SYNTHIA Dataset (accessed 4 February 2025).

Figure 3.1.2-7: synthetic samples from the Artbench dataset ([193]).

Training predominantly on synthetic data is a **growing practice** but it is not reflective of the common practices in today's GenAI systems. Indeed, there are concerns that training on synthetic data can seriously compromise **model quality**. However, recent work shows that **reduction in model quality can be avoided with extensive data curation** (Gunasekar et al., 2023). Some AI developers assert that, while this technology is sufficiently advanced to enhance training data using data augmentation techniques and to establish benchmarks for evaluating GenAI models, it is **not yet capable of supporting the complete training of new models**. In general, synthetic data is **very useful** when data is non-existent, incomplete or lacking in accuracy. It is also a viable solution when the training data needs to be anonymised for privacy purposes (for example, in case of medical data).

The development of synthetic data is possible through a process called **label-efficient learning** ([194]). Label-efficient learning, which relies on a reduced annotation effort, introduces another advantage related to training on synthetic data ([195]). Synthetic data could allow for the training of GenAI models on **high-quality data not linked to copyright-protected works**, potentially mitigating copyright related issues. Interviewees stressed the need for AI

([193]) Liaopeiyuan/Artbench, Github, 17 January 2025 (accessed 14 March 2025).

([194]) Labelling data is an important step in training many AI models. Traditionally, labelling data involved humans who annotate data with the desired information, which is a time-consuming and expensive process, especially for large datasets.

([195]) Patent Landscape Report - Generative Artificial Intelligence (GenAI), WIPO, 2024 (accessed 16 November 2024).

developers to **clearly distinguish between synthetic data and real copyrighted works** in their training datasets. Rights holders require this transparency to ensure that their opt-out requests are being honoured for genuine content and not obfuscated by claims of 'synthetic' data. There remains a **need to pay attention to the copyright-safety of synthetic data generator**, as there must be enough non-copyrighted data to train the generator.

According to the report 'Recommendations on the Use of Synthetic Data to Train AI Models' (De Wilde et al., 2024), the role of **education** is important to effectively enable the widespread use of synthetic data, given its limitations and risks. Reported below is a comparative analysis highlighting the distinct features of synthetic and real-world data.

Aspect	Synthetic Data	Real-World Data
Source	Generated using algorithms, models, or simulators, often derived from mathematical rules or AI systems.	Collected from physical, social, or online environments through direct observation or user interactions.
Quality Control	Allows complete control over the quality, distribution, and noise within the dataset.	Quality varies and often requires extensive cleaning and preprocessing to remove inconsistencies.
Bias Mitigation	Can be tailored to reduce or eliminate biases inherent to real-world data.	Inherently reflects real-world biases, which can propagate or even amplify into model outcomes.
Scalability	Easily scalable, enabling the generation of large datasets without practical constraints.	Limited by resource availability, regulatory restrictions, and accessibility of source data.
Regulatory Concerns	Minimises copyright and privacy issues as it does not directly use real-world entities or events. However, this may	Subject to privacy laws, copyright, and ethical considerations concerning data usage and retention.

Aspect	Synthetic Data	Real-World Data
	not apply if copyright-protected data has been used at some point during synthetic data generation.	
Training Value	May lack nuanced patterns or unexpected anomalies present in real-world data, reducing authenticity. Synthetic data also need to be regularly updated to adequately represent real-world data and related changes.	Rich with natural variation and complexity, offering greater contextual accuracy for specific tasks.
Cost Implications	Low production costs once generation systems are established, especially at scale.	Acquisition and processing of real-world data can be expensive, particularly for large datasets.
Practical Applications	Ideal for testing and validation environments where controlled variables are crucial.	Crucial for tasks requiring high fidelity or where data authenticity is non-negotiable.

Table 3.1.2-2: Comparison of Synthetic and Real-World Data.

While synthetic data provides significant advantages in scalability, bias mitigation, and compliance, real-world data remains irreplaceable for applications requiring authenticity and complexity. However, the **combined use of both forms of data is emerging as a strategy to maximise the benefits of each**, particularly in applications like AI training, testing, and validation.

3.1.3 Data Pre-Processing

High-quality data preparation is the foundation upon which successful GenAI applications are built. Clean datasets enhance the performance of AI models. For instance, a study by Telmai ([196]) showed that increasing noise levels in datasets led to a drop in prediction quality from 89% to 72%.

Some **key attributes of data quality** include ([197]):

- **Accuracy**: The data correctly represents the real-world condition;
- **Completeness**: All necessary data points are present;
- **Consistency**: Data does not contradict itself across different datasets;
- **Timeliness**: The data is up-to-date;
- **Relevance**: The data is directly related to the task the GenAI model is designed to perform.

Before any model training begins, **data cleaning** is the essential first step. It involves ([198]):

- Removing **duplicates**;
- Handling **missing data**: Use methods such as mean imputation or predictive models to fill in incomplete records;
- Removing **noise**: Random errors or inconsistencies can be identified and corrected through various techniques like data smoothing or filtering;
- Handling **outliers**: Extreme values can skew model results. Outliers can be removed, transformed, or capped at a specific percentile depending on the context;

At this stage, it is also possible to employ **watermarking** detectors (see *Section 4.3.3.1*) to filter out copyrighted content and **fingerprinting** solutions (*Section 4.3.3.2*) to identify assets and verify whether their use in machine learning is compliant. However, the latter is feasible only if a storage system is available to maintain the relationship between fingerprints and the

([196]) Data Quality's Role in Advancing Large Language Models, Telmai, 20 September 2023 (accessed 14 March 2025).

([197]) Ibid. See also Data Preparation For Generative AI: Best Practices And Techniques, Xite blog, 24 October 2024 (accessed 17 November 2024).

([198]) Ibid.

relevant rights information associated with the content, as exemplified by the solution proposed by Liccium (see *Section 3.4.2.6*).

After the data cleaning, some pre-processing is needed to **transform raw data into a suitable format** for model training. This step includes encoding categorical variables, scaling numerical data, and splitting the data into training, validation, and test sets ([199]).

Then, a **data augmentation** procedure is often performed to enhance the volume and variety of the training data, especially when data is scarce. This technique involves generating new data points from existing ones by applying transformations, such as rotation, flipping, or random cropping of images ([200]).

To ensure that all features have the same scale, data **normalisation**, and **standardisation** are applied. These techniques are particularly important in neural networks, where differences in scale can impact the convergence of the model ([201]).

Feature engineering involves selecting, modifying, or creating new variables (features) from raw data to improve model performance. One of these techniques is Interaction Features ([202]).

Finally, adhering to privacy regulations like GDPR and **ensuring sensitive information is anonymised or encrypted** protects both the company and individuals from legal and ethical issues ([203]).

An example of a training data curation process can be found in *Annex VIII*.

([199]) Ibid.

([200]) Ibid. For example, in image generation tasks, rotating or flipping an image can provide additional data without altering its core properties. In textual applications, synonym substitution or slight rephrasing helps create more diverse training samples.

([201]) Ibid.

([202]) Ibid. Interaction Features consists in creating new variables by combining existing ones to capture relationships between them. This is useful, for example, in a GenAI model generating house images, where engineered features like architectural styles, colour palettes, or room spatial relationships can enhance results.

([203]) Ibid.

3.1.3.1 Text Tokenisation

Tokenisation is a **pre-processing** procedure applied to **text data** before feeding language models (the subset of GenAI models devoted to natural language generation).

During **tokenisation**, Natural Language Processing (NLP) algorithms split the text into individual words, word parts, numbers and punctuation. The text can then be processed in a way that makes it machine-readable: each token is assigned a unique numerical value.

Embeddings assign each token to statistically calculated semantic fields of meaning using embedding vectors. This allows the model to generalise based on meaning and not just match exact word patterns ([204]).

3.1.3.2 Convert Image Data in a Form Suitable for Training

In **GenAI for images**, analogous techniques have been developed to process and represent visual data in a form suitable for machine learning models, particularly transformer-based architectures. While many recent models adopt approaches analogous to tokenisation, like patch-based tokenisation or vector quantisation**,** to align with transformer-based architectures, others use alternative strategies or avoid tokenisation altogether. Details of such approaches can be found in *Annex VIII*.

3.1.4 Model Fine-Tuning

Fine-tuning is the process of taking a pre-trained general-purpose model (in the case of GenAI it can be a foundation model) and further training it on a **specific, smaller dataset** tailored to a particular task or domain. This approach allows the model to adapt its general purpose to meet the needs of a specific application, improving performance while requiring **less**

([204]) The embeddings are computed during training and serve to capture semantic similarities between tokens (Example: Apples, pears and bananas will most likely be found in the proximity of word fields like fruit, food, etc.). Each token is thereby assigned one or usually multiple numerical embedding vectors. These vectors comprise a list of hundreds to thousands of numbers representing the statistical-semantic characteristics of the respective tokens. Tokens with similar meanings receive similar embedding vectors. The embeddings for each token are concatenated together into one long input vector in order to convert a passage of text into a form that the neural network can process.

computational effort compared to training a model from scratch. In practice, it consists of making **small changes to the base model's parameters** until the target behaviour is obtained. To reach an adequate level of control over this operation, the training data used is often labelled, i.e., annotated with labels and attributes (see *Section 3.1.3* on Data Pre-Processing).

While fine-tuning is a technique normally applied to general foundation models, certain fine-tuning methods—such as Textual Inversion, DreamBooth, and Custom Diffusion—allow individual users to **incorporate personalised concepts** into base models with minimal data and computational resources. These developments have raised increasing concerns among copyright holders due to the potential for fine-tuning to **generate outputs that closely replicate protected works without authorisation** (Zhao et al., 2024).

Users of image generated art can **mimic an artist's style** by fine-tuning models on specific artists' images using services appositely offered by some companies ([205]) (Jiang et al., 2023).

3.1.5 Example: OpenAI's ChatGPT training

To better describe all the data-involving training phases, OpenAI's ChatGPT can be considered as an example. In *Figure 3.1.5-1* the relative building steps are outlined.

([205]) An example of those companies is Wombo, which allows users to fine-tune Stable Diffusion.

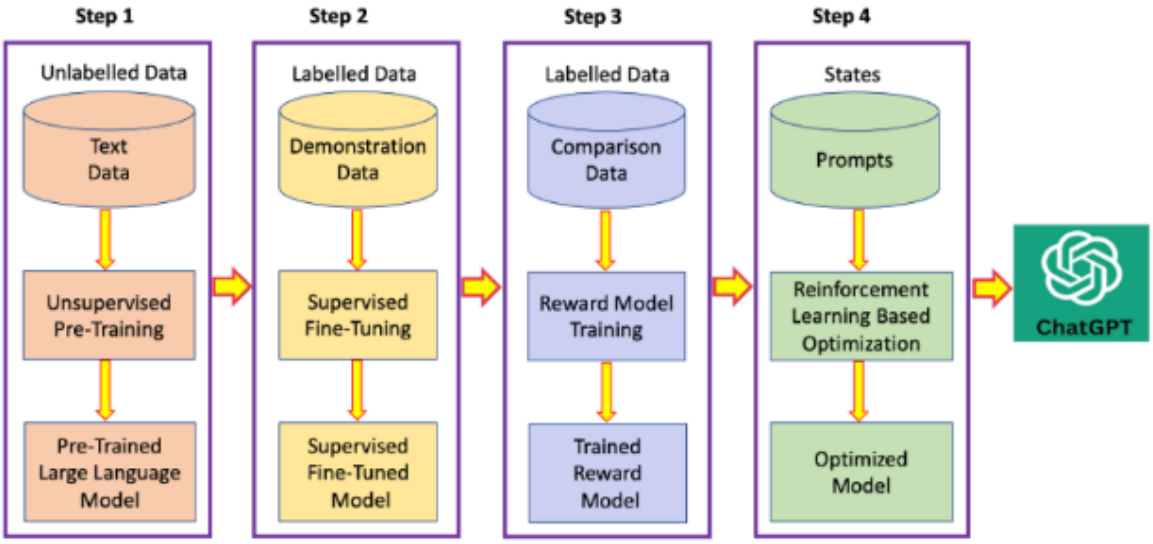


Figure 3.1.5-1: Building steps of ChatGPT (Naik et al., 2023).

The first stage is the one that **consumes the most data** since it aims to build the large **pre-trained foundation model** (GPT). Its training is fed with the input vectors coming from data-preprocessing (see above *Section 3.1.3* on 'Tokenisation') and consists of calculating billions of **weights** by **statistically** determining which token is most likely to follow the respective preceding token. The weights are refined using so-called **back-propagation**: the predicted next token is compared to the actual next token from the original text and errors (losses) are back-tracked in order to adjust the weightings and improve predictions. These training loops are repeated, also with **Human Feedback (HF)**, until developers find the error rate acceptable.

Next, HF techniques are utilised once more for **fine-tuning** (Step 2). In this phase, a **different set of training data** is labelled by human AI experts, enabling **supervised** learning to take place. This additional training phase is used to make the model capable of managing the specific tasks desired (for example translation and paraphrasing). Thus, this step needs **filtered** input data related to the target task.

Step 3 describes the training, which again involves HF, of a secondary model: the **reward model**. In particular, it is trained to be capable of distinguishing which GenAI's outputs are better than others. The 3H evaluation metric is used: Honest, Helpful, Harmless.

In step 4, the GenAI and the reward models are put together to conduct the **reinforcement learning**. This is done by trying to maximise the output of the reward model when evaluating the GenAI's generated outputs.

Thus, for the entire procedure, **both generic and specific data is needed**. Moreover, the input data for the reward model's training highly influences the ChatGPT's quality and adherence to predetermined principles.

OpenAI itself declares on its webpage ([206]) that this huge amount of content is gathered mainly from three types of sources:

- *"Select **publicly available data**, mostly collected from industry-standard machine learning datasets and web crawls, similar to search engines. We exclude sources we know to have paywalls, primarily aggregate personally identifiable information, have content that violates our policies, or have opted-out;*
- ***Proprietary data from data partnerships**. We partner to access non-publicly available content, such as archives and metadata. We don't pursue paid partnerships for purely publicly available information;*
- ***Human feedback** from AI trainers, red teamers, employees, and users whose data control settings allow model improvements."*

3.1.6 How Training Data is Represented Inside the Models

After training, the **model's "knowledge"** is distributed among its parameters and its **representation depends significantly on the GenAI model's architecture**:

In **Generative Adversarial Networks (GANs)** - much used models for generating images, sounds or videos - the **distribution** of training data's features is coded into the model's parameters (which are the result of the adversarial training process). These values enable reconstructing a latent space whose distribution represents a model of the initial database.

In a **Variational Autoencoder (VAE)**, a type of model commonly used for generating content (particularly images), the model learns and internally represents the **data distribution** of the

([206]) Our Approach to Data and AI, OpenAI, 7 May 2024 (accessed 7 November 2024).

training dataset. In other words, it learns how the data points are spread across different values or features (for example, in a dataset of images of handwritten digits, the data distribution includes information about the shapes of digits, their sizes, variations in handwriting, etc.).

In **Diffusion Models (DMs)**, widely recognised for applications such as text-to-image generation (e.g., DALL-E), the model processes noised versions of the training data which has been produced by introducing Gaussian noise. The training process aims to estimate a **latent space that captures all possible variations of the images**. The model subsequently learns to reverse this noising process, progressively denoising samples from the final distribution of the noised data to generate new content. Although the latent representation of the training dataset is effectively stored in the model, this occurs indirectly through its learned parameters. These parameters encode **patterns, correlations, and features** extracted from the training data. The latent representation is a compressed, generalised abstraction that captures the key characteristics of the dataset, rather than an explicit copy.

In **Generative Pre-trained Transformers (GPTs)** and **LLMs**, used to generate natural language, the model learns the **probability distribution** over a sequence of words (see *Section 3.1.5* for the description of ChatGPT's training). Consequently, generation consists of sampling based on this distribution. In summary, such architectures rely on distributed representations, where training data is encoded as patterns within the model's parameters and transformer attention mechanism's weights.

3.1.7 GenAI Training Technical Costs

The computing and energy cost for training LLMs is substantial and rises with increasing **model size**, like the **number of its parameters** (Hoffmann et al., 2022) ([207]).

AI Index collaborated with researchers from Epoch AI to estimate the training costs for some of the well-known GenAI models:

- OpenAI's GPT-4 was trained in 2023 with an estimated cost of $78.4 million;
- Meta's Lama 2 70B required about $4 million in 2023;

([207]) For example, Kaplan et al. found that increasing the model size by 5.5 times and the number of tokens by 1.8 times requires a tenfold increase in the computational budget (Kaplan et al., 2020).

- In the same year, Google's Gemini Ultra is estimated to have required up to $190 million for its training.

Those estimations are based on cloud compute rental prices and consider the model's training duration, the hardware's utilisation rate and the value of the training hardware ([208]).

Hardware includes GPUs, high-performance servers, networking and orchestration equipment (to allow collaboration between the different bunches of GPUs), the storage system (for the training data), as well as the cooling and power infrastructure ([209]).

As pointed out previously in *Section 2.4*, those **elevated costs often lead GenAI startups to partner and develop agreements with major tech companies**.

3.1.8 DeepSeek's Training Strategies

In January 2025, DeepSeek released DeepSeek-R1 ([210]) – a model that many ([211]) consider a pivotal step in the next evolution of GenAI. This model has been released as **open-source**, although its **training dataset remains undisclosed**, and **has been trained at a fraction of the cost** required for other models achieving comparable performances.

This efficiency is the result of **new machine learning strategies**, which **maximise computational efficiency** while leveraging **state-of-the-art model architectures**. Details of these technologies can be found in *Annex X*. The company associated those techniques with the **optimisation of the hardware infrastructure** and utilisation to further improve training efficiency.

DeepSeek's approach reportedly demonstrates that high-performance AI models can be trained efficiently without requiring **prohibitively expensive computational resources**. By

([208]) Visualizing the Training Costs of AI Models Over Time, Visual Capitalist, 4 June 2024 (accessed 14 March 2025).

([209]) Ayres, L, Estimating the Infrastructure and Training Costs for Massive AI Models, LinkedIn, 19 May 2024 (accessed 19 November 2024).

([210]) Deepseek-Ai/DeepSeek-R1, Github (accessed 15 February 2025).

([211]) DeepSeek: A Problem or an Opportunity for Europe?, CSIS, 14 February 2025 (accessed 14 march 2025).

integrating **Mixture of Experts architectures** ([212])**, MHLA, FP8 low-precision training, and infrastructure optimisations**, DeepSeek may have set a new **cost-performance benchmark** for the AI industry.

([212]) A *Mixture of Experts* (MoE) is a machine learning architecture that divides tasks among multiple specialised models ("experts") and uses a gating mechanism to dynamically select the most relevant experts for a given input. This approach improves efficiency and adaptability by focusing computational resources on the most relevant parts of a model.

3.2 Training Data Memorisation

3.2.1 Memorisation in Large Language Models

Prior research (Carlini, Ippolito, et al., 2023) have shown that LLMs memorise and **regurgitate** potentially private information, as well as long sequences (which could be copyright-protected) from their training sets. Those memorised elements can be emitted **verbatim (or nearly verbatim)** when the model is prompted appropriately (e.g., by prompting it with a piece of the memorised string). **This effect is unpredictable** because of the presence of randomness in the generation process (see *Section 4.1.3*).

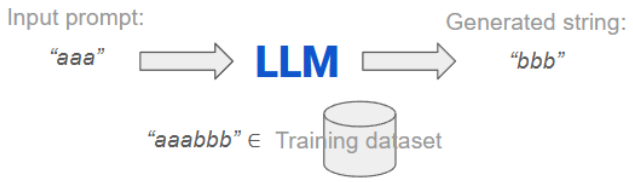


Figure 3.2.1-1: Graphical representation of the definition of 'verbatim memorisation'.

In 2023, Carlini et al. conducted a study (see *Annex IX* for more details) to identify the factors contributing to this phenomenon. They found that the probability of verbatim training data regurgitation increases with (a) the **model size** (i.e., the number of parameters), (b) the **length of the text given as input prompt** and (c) **the frequency of the sequence** within the training dataset.

While some **AI developers** claim that the memorisation phenomenon affects only a **negligible portion of the training set** and is primarily observable in **controlled laboratory settings**, other researchers state that it should not be underestimated. The extent of memorisation in models could grow with: (a) the trend toward larger model sizes, (b) the enhanced capabilities for processing an increasing number of tokens simultaneously, and (c) the increased complexity of datasets, which makes managing data duplication more challenging.

3.2.2 Memorisation in Image Models

Another study (Carlini, Hayes, et al., 2023, see *Annex IX* for more details) reveals that memorisation also occurs in **Stable Diffusion** and **Imagen**, causing the generation of images that closely resemble those in the training data. This phenomenon may occur both unintentionally and intentionally, as the probability of occurrence increases significantly if the **input prompt** used to guide generation partially overlaps the **caption associated with an image in the training set**.

The study demonstrated that the **level of memorisation is affected by the way the model is trained**. This phenomenon is less frequent in models based on the **Generative Adversarial Network (GAN)** architecture than in Diffusion Models, possibly because GANs' generators are only trained using **indirect information about the training data.**

In addition, results possibly suggest that **some characteristics of the data point itself can influence the degree of memorisation**. The most frequently extracted images were those that **differed significantly from the rest of the dataset** in terms of image features (in other words, the most 'original' images). This may lead to the conclusion that the more a content creator is original, the more likely their works are to be memorised by text-to-image GenAI models.

3.2.3 Mitigations Against Memorisation

Differentially private training, e.g., using Differential Privacy (DP) ([213]) stochastic gradient descent, is the gold standard for training models which likely do not memorise individual training examples. However, **in practice, these techniques result in less qualitative generative models** thus, LLMs aren't currently trained with DP (Ippolito et al., 2023).

([213]) Differential privacy in generative AI works by injecting statistical noise into the training process, ensuring that the influence of any single data point is limited. It minimises the risk of the model reproducing copyrighted content verbatim by adding noise or other techniques to obscure specific training data while preserving the model's overall functionality.

Instead, **data deduplication ([214])** has arisen as a pragmatic countermeasure against data memorisation. Nonetheless, deduplication alone does not guarantee that a model will avoid memorising unique (deduplicated) content (Ippolito et al., 2023). Findings of studies indicate that **deduplication significantly reduces memorisation but does not eliminate it entirely** ([215]).

3.2.4 Copyright Implications of Training Data Memorisation

First, it is necessary to distinguish between **verbatim** and **approximate memorisation**. Although the second is defined using the concept of similarity, it is still relevant from a copyright perspective because in cases of approximate memorisation the GenAI output strongly resembles the training data.

Ippolito et al. (2023), in their study used both the **BLEU score** and the **length-normalised character-level Levenshtein similarity** to detect **approximate memorisation**. The study focused on **Copilot** ([216]), which to prevent generating memorised code adopts a filtering mechanism that blocks model outputs from being suggested if they overlap significantly (approximately 150 characters) with a training example. This is a practical example of a **filter** (i.e., implemented outside the model) that aims at preventing perfect verbatim memorisation.

However, Ippolito's study shows that Copilot's filter can easily be bypassed by prompts that apply various forms of "**style-transfer**" to model outputs, thereby causing the model to produce (approximately) memorised outputs ([217]). Ippolito conducted similar experiments on

([214]) **Data deduplication** is a preprocessing technique that removes duplicate instances of data points from a dataset before training. This reduces redundant exposure to identical or highly similar content, thereby lowering the likelihood of memorisation in machine learning models. However, it does not eliminate memorisation entirely, as unique or rare training samples may still be retained in the model's parameters.

([215]) For example, Carlini et al. (2023) measured the difference in memorisation when a model is trained on a deduplicated training dataset. When they randomly generated 2^{20} images across different versions of the same model—one trained on a deduplicated dataset and the other on the original, non-deduplicated dataset—they found that the model trained on the deduplicated dataset produced 986 memorised samples, whereas the non-deduplicated version generated 1,280.

([216]) Copilot is a code auto-complete service which is trained on GitHub code. Copilot is built using the Codex language model designed by OpenAI (GitHub Copilot · Your AI Pair Programmer, 2024).

([217]) For example, by only requesting to translate the variables' names into another language it is possible to induce the model to output almost literally the input training code (for further examples see Ippolito et al., 2023, appendix F).

GPT-3 and was able to obtain training material by prompting a partial and modified version of it (for example, a text in all uppercase). This emphasises the **importance of models' training set compositions and training methods on memorisation tendencies**.

A widely cited paper on memorisation summarised this issue for a legal audience in the abstract except below.

> *…We say that a model has "memorized" a piece of training data when (1) it is possible to reconstruct from the model (2) a near-exact copy of (3) a substantial portion of (4) that specific piece of training data. We distinguish memorization from "extraction" (in which a user intentionally causes a model to generate a near-exact copy), from "regurgitation" (in which a model generates a near-exact copy, regardless of the user's intentions), and from "reconstruction" (in which the near-exact copy can be obtained from the model by any means, not necessarily the ordinary generation process).*
>
> *Several important consequences follow from these definitions. First, not all learning is memorization: much of what generative-AI models do involves generalizing from large amounts of training data, not just memorizing individual pieces of it. Second, memorization occurs when a model is trained; it is not something that happens when a model generates a regurgitated output. Regurgitation is a symptom of memorization in the model, not its cause. Third, when a model has memorized training data, the model is a "copy" of that training data in the sense used by copyright law. Fourth, a model is not like a VCR or other general-purpose copying technology; it is better at generating some types of outputs (possibly including regurgitated ones) than others. Fifth, memorization is not just a phenomenon that is caused by "adversarial" users bent on extraction; it is a capability that is latent in the model itself. Sixth, the amount of training data that a model memorizes is a consequence of choices made in the training process; different decisions about what data to train on and how to train on it can affect what the model memorizes. Seventh, system design choices also matter at generation time. Whether or not a model that has memorized training data actually regurgitates that data depends on the design of the overall system: developers can use other guardrails to prevent extraction and regurgitation. In a very real sense, memorized training data is in the model...*

Some AI companies feel that they're managing to increasingly reduce memorisation through testing. Today, exploiting memorisation requires extensive knowledge of the training data and such scenario is not representative of the typical usage.

3.3 Comparison Criteria for Reservation Measures

As will be presented in *Section 3.4*, there are a multitude of approaches to TDM rights-reservations, each of which may be used in different contexts with its respective advantages and limitations. For the purpose of comparing the measures identified through desk research and interviews (designated as "**X1A measures**" in *Section 2.5*), a set of specific characteristics have been defined, based on the analysis of the AI ecosystem, and insights from stakeholder interviews. These characteristics are presented below:

- **Typology** - is this measure **(a) location-based**, **(b) file-based**, **(c) work-based**, **(d) repertoire- based**?

  Different typologies of opt-out account for the diversity of use cases, TMD methods, and content sector characteristics, and in the context of 'appropriate means' requirement that need to be met for a valid opt-out.

- **TDM User Specificity** - does the measure allow for the expression of reservation which differentiates between different TDM users?

  User specificity is relevant as it allows for an expression of TDM reservation which might differ between specific users, which in turn might interface with licensing decisions (where specific users are granted authorisation for TDM, and a mechanism needs to be in place to allow these authorised users to access works).

- **Use-Differentiation** - can the measure be used to express TDM reservations based on differentiated use cases?

  Use differentiation is relevant because many stakeholders have suggested that their TDM reservations differ based on use cases ranging from: **(i) prohibition of all TDM**, **(ii) prohibition of all TDM for AI training**, **(iii) prohibition specifically of commercial GenAI training**, **(iv) prohibition for model real-time inference and retrieval**.

- ***Granularity*** - does the measure apply to individual works or a larger set of content based on practical organisation?

  Granularity is relevant because depending on the context in which content is distributed, a right holder may have different TDM reservations for different components of that content. Additionally, this is important given the use of repertoire-level measures by rights holders entities that manage large portfolios of content.

- ***Versatility*** - is this measure specific for some **type of content**, or can it be used for all file-types and digital assets?

  Versatility is relevant because of two competing principles affecting the adoption and deployment of a measure: (i) technical measures are affected by network effects, where the incentive to use a given measure increases with its widespread use by other rights holders, and (ii) the specificities and demands for opt-out characteristics may differ between content sectors based on the nature of those markets.

- ***Robustness*** - Is the measure **resilient against modification/removal** (intentionally by bad actors, or unintentionally through distribution processes)?

  Robustness is relevant because of the complexity of the training data value chain, in which opt-out reservations may need to be communicated to different actors engaging in different use cases.

- ***Timestamping*** - Does the measure associate a timestamp to the reservation in a **robust** way?

  Robust timestamping can be a highly valuable tool in litigation.

- ***Authentication*** - Does the measure feature a mechanism to ensure that the opt-out is declared by a **legitimate right holder** or an authorised representative?

Authentication is relevant to create trust and confidence in the reservations legitimacy. It is also relevant with regard to the legal requirement for a legally valid TDM opt-out to be expressed *'by the right holder'*.

- ***Intermediation*** - Does implementing the measure require the intermediation of a **third party** (representative organisation, third party solution provider, or distribution intermediary)?

  Intermediation is relevant as it clarifies the role of intermediaries and solution providers in enabling rights holders to exercise their opt-out, and is also relevant with regard to the legal requirement for a legally valid TDM opt-out to be expressed *'by the right holder'*. This characteristic may in turn affect other characteristics including 'openness', 'ease of implementation', and 'flexibility'.

- ***Openness*** - Is the measure proprietary or openly available for use? Is it an industry (de facto) standard?

  Openness is relevant because it may affect the extent to which the measure eventually becomes widely adopted, the balance of interest between different market players including solution providers and intermediaries, and the potential costs to rights holders associated with the adoption of a measure.

- ***Ease of Implementation*** - What level of efforts and cost are required **by rights holders** to implement the measure, **and by TDM users to detect and apply the reservation**.

  Ease of Implementation is relevant because rights holders vary in their technical and financial capacities to adopt and manage different opt-out mechanisms, with widely adopted measures that should ideally not put any particular rights holders group at a disadvantage. On the other hand, developers are more inclined to favour solutions supporting automated and cost-effective extraction of reservation information.

- ***Flexibility*** - Does the measure allow the rights holders to easily change their expression of reservation **after initial implementation**?

  Flexibility is relevant because rights holders may change their TDM reservations as the market evolves, and as their specific circumstances change.

- ***Retroactivity*** - Does this measure apply only to future acquisition, or can it manage exclusion where a **work is already included in a training dataset**.

  Retroactivity is relevant because works from rights holders may be already included in distributed datasets, and irrespective of whether this data was originally acquired legally or not, subsequent TDM users using these datasets (such as for GenAI training) may still be restricted by current opt-outs.

- ***External Effects*** - Does the measure create any **unintended effects** (external to the issue of TDM management) which might affect rights holders, TDM users, or third parties, either positively or negatively?

  External Effects are relevant because the use of a reservation measure may have peripheral effects on other aspects of a rights holders' content distribution strategy (e.g., discoverability on the open internet), or on TDM users or third parties (e.g., larger file sizes).

- ***Generative Application*** - Can the measure also be used to **identify generative output**?

  Generative Application is relevant because using a measure for both managing rights reservations on the AI input side and transparency on the output side can maximise the network effects that **support its widespread adoption**. In other words, certain measures may be beneficial as 'dual purpose' mechanisms on both the input (expressed opt-out) side, and output transparency (synthetic content disclosure) side.

- ***Offline Application*** - Can this measure be applied to content which is not directly hosted online (either **offline digital content or analogue content**)?

  Offline Application is relevant because CDSM Article 4 specifically allows for rights reservations for both content made available on the open internet, and 'other cases'.

- ***Market Maturity*** - To what extent is this measure **already used**, and has proven to be **scalable**?

  Market Maturity is relevant because already deployed measures (compared to measures under development) have proven viability, and may benefit from network effects associated with widespread adoption.

3.4 Reservation Measures

As mentioned in Chapter 2, rights reservations (TDM opt-outs) are crucial for managing rights holders' economic interests and defining AI developers' business and technical models. There is a diversity of approaches used (and proposed) for rights reservations, each with its own advantages and limitations. This section provides an overview of the solutions identified for managing copyright compliance and mitigating potential liability for infringement in TDM and GenAI training processes. It first analyses solutions used by copyright holders to reserve the right to authorise reproductions for TDM purposes (opt-out mechanism) in accordance with Article 4(3) of the CDSM (i.e., **X1A** measures).

Such measures might be categorised as **'legally-driven'** or **'technically-driven'** measures. *These terms are used for categorisation purpose only and do not have any implication on the inherent capacity of any measure to meet the legal criteria for a valid opt out (including 'machine readability').*

3.4.1 Legally-Driven Measures

'Legally driven measures' are implemented by rights holders without the intermediation of a 'Solution Provider' and often involve legal statements and **contractual provisions in natural language**, without the use of technical protocols.

3.4.1.1 Unilateral Declarations

Unilateral declarations are an example of 'appropriate means' given in the CDSM Directive for 'other cases' (non-online distribution). These are **public statements communicated by a right holder or rights holders' group**, usually a major commercial intermediary like a publisher or record label.

Unilateral declarations are also used by CMOs that manage specific exclusive rights on behalf of their members (for example, as mentioned in Chapter 2, both the German GEMA and the

French SACEM issued unilateral declarations opting-out from the use of their members' works for TDM uses).

As an example in the field of book publishing, the **Penguin Random House** has recently publicly stated that all of its **publications will include Article 4 opt-out reservation on the copyright page of books ([218]).** The text of this copyright statement is presented below.

> **Penguin Random House Opt-Out Declaration**
>
> Penguin Random House values and supports copyright. Copyright fuels creativity, encourages diverse voices, promotes freedom of expression and supports a vibrant culture. Thank you for purchasing an authorised edition of this book and for respecting intellectual property laws by not reproducing, scanning or distributing any part of it by any means without permission. You are supporting authors and enabling Penguin Random House to continue to publish books for everyone. No part of this book may be used or reproduced in any manner for the purpose of training artificial intelligence technologies or systems. In accordance with Article 4(3) of the Digital Single Market Directive 2019/790, Penguin Random House expressly reserves this work from the text and data mining exception.

In this case, the reservation is communicated directly to the party that is in possession of a copy of the book – whether this is a physical copy or an electronic copy.

At this stage the exact scope of such unilateral declaration and whether they would be considered as extending beyond a potential user, to the general public, is not entirely clear. Unilateral declarations may either be considered as **(A) a declaration which is attached to a specific copy of a work**, or **(B) a declaration which is communicated independently** of any actual copies of the work. In the case of 'independent' unilateral declarations (B), the issue may arise as to how and whether a potential **TDM user is able to have constructive knowledge of this declaration**. This is a potential limitation of such declarations.

([218]) Penguin Random House books now explicitly say 'no' to AI training, *The Verge,* 19 October 2024, (accessed 15 December 2024).

Another form of unilateral declarations used by several major rights holders are declarations posted on their websites. As an example in the field of music publishing, Sony Music published a 'Declaration of AI Training Opt Out' in May 2024 ([219]). Some rights holders are also **notifying these public declarations directly to specific AI developers**. For example, in addition to the unilateral declaration of rights reservations on its website, Sony Music sent letters to 700 AI and music streaming companies to inform them that it is opting-out of AI training ([220]).

Another example below shows the unilateral declaration on the website of Concord Music, a USA-based music publishing company and group of independent record labels. Several observations can be made: Firstly, **the reservation explicitly refers to content which is freely accessible on its website** (and thus possibly subject to web scraping) as well as its musical content such as lyrics, musical compositions, and sound recordings, which are not freely available on the website. This declaration could be considered both part of the website's 'terms and conditions', as well as a 'unilateral declaration'. Secondly, while Concord Music is a USA-based company, the declaration explicitly makes reference to the CDSM Directive and Article 4. Thirdly, this reservation has not been translated into the Concord Music website *robots.txt* file, that at the time of the study did not appear to prohibit web scraping generally.

Concord Music Declaration

Declaration of Rights Reservation

Last Updated: June 25, 2024

Concord Music Group, Inc. opts out of any copyright exception for text or data mining or other computational techniques.

In light of increasing development of artificial intelligence (AI) and in supplementation of Concord's Website Terms and Conditions of Use, Concord Music Group, Inc. and all of its affiliates (collectively, "Concord") declare that, except as specifically and explicitly authorized by Concord in writing, Concord expressly reserves and opts out of any copyright exception for

([219]) Declaration of AI Training Opt Out, Sony Music, 16 May 2024 (accessed 14 March 2025).

([220]) Sony Music warns AI companies against 'unauthorized use' of its content, The Verge, 17 May 2024, (accessed 14 March 2025).

text or data mining, web scraping, or similar reproductions, extractions, or uses (“TDM”) of any Concord content (including, but not limited to, musical compositions, lyrics, audio recordings, audiovisual recordings, artwork, images, data, etc.). This reservation applies to any purposes, including the training, development, or commercialization of any AI system, and by any means, including bots, scrapers, or other automated processes, to the fullest extent permitted by applicable law in all relevant jurisdictions, including for the purposes of Article 4(3) of Directive ((EU) 2019/790) and all national laws having transposed the same. Concord’s rights reservations apply to all existing and future Concord content, including musical works that may be identified through publicly available means or listed in databases maintained by organizations such as the International Confederation of Music Publishers (ICMP). Concord’s reservation of rights set forth herein is without prejudice to any and all prior reservations of rights and legal rights and remedies, all of which are expressly reserved. Inquiries regarding AI training or text and data mining permissions can be sent to:…
https://concord.com/robots.txt (as of 28-04-25) User-agent: * Disallow: /wp-admin/ Allow: /wp-admin/admin-ajax.php Crawl-delay: 10 Sitemap: https://concord.com/sitemap_index.xml

3.4.1.2 Databases Listing Unilateral Declarations

As noted above, a challenge with unilateral declarations is that it requires potential TDM users to be aware of them. In that respect, some major rights holders are also **consolidating their unilateral declaration**s in a single place. This increases their visibility and supports efforts to identify expressed unilateral declaration.

The most notable example of such a consolidation approach is the platform RightsAndAI ([221]), which was launched by the International Confederation of Music Publishers (ICMP). This platform is open to music companies and music CMOs to make rights reservation declarations, and have their names added to a consolidated list of rights holders. The 'metadata on reserving rights holder' has over 1,300 entries of music industry companies that have used this platform to declare rights reservations on their repertoires. According to ICMP, over 80% of the global music publishing market (by share) has already united in this common platform to reserve their rights for AI training. Participating rights holders are identified by their IPI numbers (a standardised unique identifier used in the music industry to identify songwriters, composers, and publishers). It is important to note that the right reservation declarations listed are about 'all rights' of a given right holder (or repertoire), without providing information on the individual works covered by such reservation. The RightsAndAI declaration is reproduced below.

RIGHTSANDAI DECLARATION

As rightsholders and on behalf of our songwriter and composer partners worldwide, we affirm the critical importance for Artificial Intelligence (AI) development to be ethical, responsible and licensed and reserve all our rights under copyright without prejudice to any reservation of rights we may have made elsewhere.

Any access, use, reproduction, distribution or exploitation of any copyright-protected musical works and associated literary works or data that we own, control or represent (together "Works"), including by web crawling technologies, data scraping or any and all present or future forms of Text and Data Mining (TDM), for AI training without a valid licence breaches the copyright in the Works and denies our songwriters and composers' appropriate consideration for their work.

Without lawful access, any unlicensed commercial use or output built on the Works in data or audio or audiovisual form – for example training Large Language Models (LLMs) or Generative AI - also breaches the copyright in the Works. Scraping or accessing music online without a license is prohibited.

([221]) See RightsAndAI website (accessed 14 March 2025).

> Without limiting the above, all rights in the Works are expressly reserved in relation to relevant Text & Data Mining (TDM) provisions and copyright exceptions, including but not limited to Article 4 of the 2019 EU Copyright Directive, The United States Code (USC) 'Fair Use' Exemptions, 2020 Copyright Act of Japan, The 2021 Copyright Act of Singapore, the 2007 Copyright Act of Israel.
>
> There is no legal or moral excuse for the unlicensed access and exploitation of creators' works for AI training, use or output.
>
> Respect for laws and rights will ensure rapid but responsible innovation, continued investment and a sustainable future for AI tech and creative sectors worldwide.

3.4.1.3 Licencing Constraints

'Licencing Constraints' refers to measures in licensing agreements (concerning non-TDM uses) which contain clauses opting-out licensed works from TDM. Such reservations are only binding on the parties to the contract, but they allow rights holders to **separate TDM rights from the broader right of reproduction**.

As discussed in the previous analysis of the *'by the right holder'* requirement for a valid opt-out (*Section 2.2.1.8.2*), there may be a presumption that the capacity to authorise or opt-out of TDM is attributable to a licensee who has agency over a work, through the delegation of the right of reproduction. Licensing constraints on TDM may therefore represent **a contractual modification of this implied agency**.

As an example, the Authors Guild, a professional organisation for USA literary authors, has recommended the inclusion of a model 'AI Training clause' in contracts between authors and publishers when negotiating publishing and distribution agreements ([222]). This clause does not explicitly prohibit TDM, but all uses for GenAI training purposes, which *may* involve TDM. The clause also prohibits the party from sub-licencing (i.e., authorising) for AI training purposes, but it does not place an affirmative obligation on the other party to prohibit such potential uses,

([222]) AG Recommends Clause in Publishing and Distribution Agreements Prohibiting AI Training Uses, The Authors Guild, 1 March 2025 (accessed 14 March 2025).

like through appropriate opt-out means, as this clause was likely drafted with American publishing agreements in mind.

Authors Guild Model Contractual Clause

No Generative AI Training Use.

For avoidance of doubt, Author reserves the rights, and [Publisher/Platform] has no rights to, reproduce and/or otherwise use the Work in any manner for purposes of training artificial intelligence technologies to generate text, including without limitation, technologies that are capable of generating works in the same style or genre as the Work, unless [Publisher/Platform] obtains Author's specific and express permission to do so. Nor does [Publisher/Platform] have the right to sublicense others to reproduce and/or otherwise use the Work in any manner for purposes of training artificial intelligence technologies to generate text without Author's specific and express permission.

3.4.1.4 Website Terms and Conditions

As discussed in the analysis of the *'appropriate means'* requirement (*Section 2.2.1.8.3*), **website terms of services** are common measures used to opt-out of TDM. However, this particular measure is limited by the fact that it only relates to specific copies of a work hosted on a specific website (i.e., it is 'location-specific').

Also, as previously explained there is still a debate on whether website terms and conditions meet the **'machine-readable' criterion** for cases of online content. However, the **location and positioning** of terms and conditions can raise issues for practical implementation as an opt-out mechanism, such as when these terms are contained on a separate page of a website or a separately hosted file (as discussed in *Section 2.2.1.8*). Furthermore, website terms and conditions are typically **expressed in the natural language of the target audience** of a website, with linguistic diversity adding to the practical complexity of machine-readability.

Some rights holders groups have proposed **standardised models for website terms and conditions as an opt-out mechanism**. For example, the French (book) Publishers

Association (*Syndical national de l'édition* – SNE) has developed a 'standard clause opposing TMD by AI' which it recommends publishers to include in their website terms and conditions ([223]). Interestingly, the SNE suggests that this language might also be included in 'legal notices' which might take the form of direct notices to potential TDM users, or even general unilateral declarations.

SNE Standard Clause on TDM and AI

Clause type

Fouille de textes et de données

PROPRIETE INTELLECTUELLE

La structure du Site ainsi que l'ensemble des contenus diffusés sur le Site sont protégés par la législation relative à la propriété intellectuelle.

Les photographies, illustrations, dessins ou tout autre graphique, documents, les signes, signaux, écrits, images, sons ou messages de toute nature figurant sur le Site (ci-après « les Contenus ») ne peuvent faire l'objet d'aucune reproduction ou représentation sans l'autorisation préalable expresse et écrite de […].

POLITIQUE FOUILLE DE TEXTES ET DE DONNEES < TDM-RESERVATION: 1>

[…] s'oppose à toutes opérations de moissonnage et de fouille de textes et de données au sens de l'article L. 122-5-3 du code de la propriété intellectuelle. Cette opposition couvre l'ensemble du Site et des Contenus auxquels il donne accès. Toutes opérations de moissonnage et de fouille de textes et de données visant le Site et ses Contenus, y compris par des dispositifs de collecte automatisée de données constituent donc des actes de contrefaçon sauf obtention d'un accord spécifique formellement exprimé de […].

L'article R. 122-28 du code de la propriété intellectuelle précisant que l'opposition mentionnée au III de l'article L. 122-5-3 peut être exprimée par tout moyen, y compris par le recours à des conditions générales d'utilisation d'un site internet ou d'un service, l'absence de metadonnées associées au

([223]) Une clause-type pour s'opposer à la fouille de textes et de données par les intelligences artificielles (in French), SNE (accessed 14 March 2025).

> Site, répertoires du Site, Contenus du Site est sans incidence sur l'exercice du droit d'opposition exprimé par les présentes conditions générales d'utilisation.
>
> Pour faciliter la lecture de ce droit d'opposition par tout dispositif de collecte automatisée de données, cette opposition est également exprimée ainsi < TDM-RESERVATION: 1>.

Several observations can be drawn from the analysis of this model clause. First, the clause not only makes explicit reference to the enabling legal provision under French Law, but also explicitly cites that law permits the use of **website terms and conditions as a valid opt-out mechanism**. The clause also stresses that *'**the absence of metadata** associated with the Site, directories of the Site, or the Content of the Site **does not affect the exercise of the right** to oppose as expressed in these general terms of use'.* This may constitute a pre-emptive rebuttal of any argument that metadata reservations and terms and conditions are cumulative conditions for a valid opt-out. The clause also contains the following text:

> *"To facilitate recognition of this right to oppose by any automated data collection tools, this opposition is also expressed as follows: < TDM-RESERVATION: 1>"*

This may be a way to address the 'machine-readability' of the clause, that can be detected without the use of natural language processing capabilities by a TDM user. The Boolean operator "**< TDM-RESERVATION: 1>**" partially incorporates the mechanism of a 'technically-driven measure' into the 'legally-driven measure' of terms and conditions.

3.4.2 Technical Reservation Measures

Technical reservation measures are developed on the basis of internet-related languages and protocols (HTML, HTTP, ODRL, RightsML), as well as technical instruments (blockchain or federated registries). Detailed explanations on these different protocols and technical instruments are provided in *Annex XI.*

As for technical opt-out mechanisms used by rights holders **[X1A Measures]**, a general typology would distinguish between two types of measures: **(A) Location-based (or 'web-**

based') measures, and **(B) Asset-based (or 'content-based')** measures. The differences between the two are detailed in *Section 3.5.1*.

3.4.2.1 Robots.txt (REP Protocol)

Initially designed in the mid-1990s to manage web server load by controlling bot traffic, robots.txt has evolved into a mechanism primarily used to express preferences for content indexing ([224]), more recently for AI related web-scraping.

Robots.txt is the implementation of the **Robots Exclusion Protocol (REP)**. It provides crawlers with information on which **URLs** they are allowed or disallowed to access on a website.

It consists of a **file** stored in the website root directory where "Allow"/ "Disallow" directives are listed together with the relative URLs and often with the indication of the specific crawler user-agent to which such directives are directed at. Below is an example in which the permission to access a specific website's directory is denied:

```
User-agent: Googlebot

Disallow: /test/
```

The robots.txt file is publicly available as default and can be accessed by appending the string *"/robots.txt"* after each website's base URL. For example, *Figure 3.4.2-1* shows the content of the file on the site of The New York Times as of January 2025, which address some disallow directives to all user-agents:

([224]) For example, robot.txt can be used to prevent specific web content from appearing in the SERP (Search Engine Results Page) of Google or other search engines.

```
# New York Times content is made available for your personal, non-commercial
# use subject to our Terms of Service here:
# https://help.nytimes.com/hc/en-us/articles/115014893428-Terms-of-Service.
# Use of any device, tool, or process designed to data mine or scrape the content
# using automated means is prohibited without prior written permission from
# The New York Times Company.  Prohibited uses include but are not limited to:
# (1) text and data mining activities under Art. 4 of the EU Directive on Copyright in
# the Digital Single Market;
# (2) the development of any software, machine learning, artificial intelligence (AI),
# and/or large language models (LLMs);
# (3) creating or providing archived or cached data sets containing our content to others; and/or
# (4) any commercial purposes.
# Contact https://nytlicensing.com/contact/ for assistance.

User-agent: *
User-agent: Googlebot
Disallow: /ads/
Disallow: /adx/bin/
Disallow: /athletic/wp/wp-admin/
Allow: /athletic/wp/wp-admin/admin-ajax.php
Disallow: /athletic/async-*
Disallow: /athletic/search/*
Allow: /athletic/search/$
Disallow: /athletic/checkout/
Disallow: /athletic/checkout?plan_id*
Allow: /athletic/checkout/$
```

Figure 3.4.2-1: A piece of the robots.txt file hosted on the site of The New York Times ([225]).

With the rise of GenAI, a clear trend consists of extending the use of robots.txt to enable website owners to declare if they wish to 'opt-out' their site's content from being crawled by AI **web-scrapers**. This is achieved by adding **"Disallow" directives** to the file, specifying the relative URL of the page containing the works that must not be used for training (note that **the entire page won't be crawled**), and assigning the **AI crawler's name** to the "user-agent" property. To block multiple crawlers, each must be listed with its exact name (see *Section 3.1.2.2* on Web Scraping for some examples). However, a certain level of flexibility can be achieved through the use of wildcards in the URLs.

3.4.2.1.1 Market Dynamics and Robot Exclusion Protocols

Stakeholder interviews, industry discourse, and academic literature all suggest that REP is commonly used as a tool or benchmark for managing the relationship between rights holders and TDM users, including AI developers.

([225]) See New York Times website (accessed 31 January 2025).

According to **Originality AI's** ([226]) data on the percentage of the top 1000 sites disallowing AI crawling through REP, as of February 2025, nearly 25% of those sites were using the protocol for blocking the GPTbot (the crawler from OpenAI) and the CCBot (gathering data for the Common Crawl dataset, see *Section 3.1.2.1.2* on Common Crawl). *Figure 3.4.2-2* reflects on the evolution of this data over time.

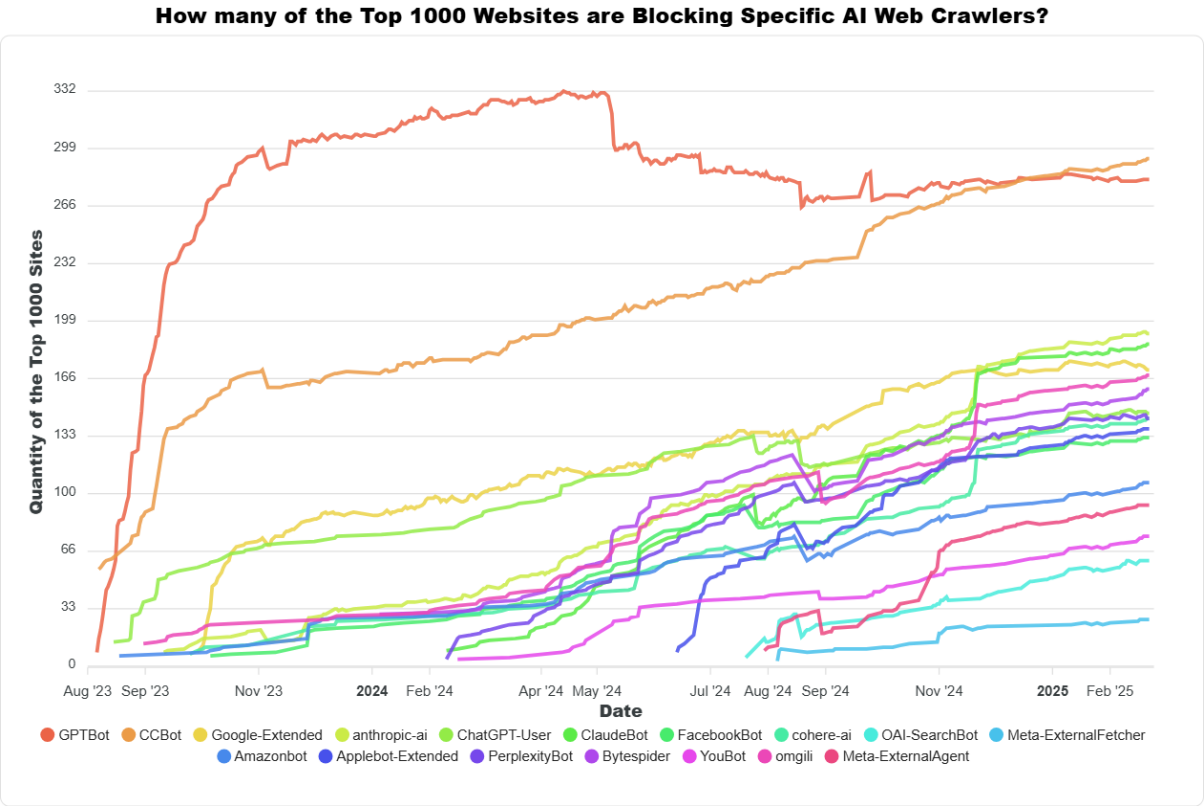


Figure 3.4.2-2: Trend of the op global 1000 sites' usage of robots.txt to block AI crawlers ([227]).

As such, this analysis proceeds by framing **REP as a benchmark and point of reference for TDM opt-out measures generally.**

([226]) Block AI Bots from Crawling Websites Using Robots.Txt, Originality.ai, 22 August 2024 (accessed 14 March 2025).

([227]) Ibid.

3.4.2.1.2 REP as a Standard

The development of REP long-preceded the current AI boom and the issue of TDM opt-out. The roots of REP are in the 1994 document *'A Standard for Robot Exclusion'* ([228]), which was initially submitted as proposed as an Internet Engineering Task Force (IETF) ([229]) standard in 1996 (Internet Draft *draft-koster-robots-00.txt*) ([230]). If REP emerged as a ***de facto* standard** and an important part of internet architecture, it is only being considered as a **Proposed Standard (RFC 9309)** since September 2022, though it is not yet elevated to an Internet Standard (reserved for the most mature stable standards) ([231]).

This RCF 9309 Proposed Standard developed on the previous *de facto* REP standard by adding that *"The identification string SHOULD describe the purpose of the crawler."* This document established the best practice of **purpose-specific identification strings for crawlers** (but does not codify it, as this identification practice is still not strictly mandatory as part of PCF 9309). Currently, there are several active proposals (Active Internet Drafts) for further adapting REP as an IETF standard and dealing with the issue of use-specificity, specifically in the content of crawling for AI training (See *Annex XII*).

While there is active debate on further adapting REP, it should be stressed that cementing REP as a standard for internet architecture is a distinct issue (but a possible influencing factor) from establishing REP as a possible standard for TDM opt-out. Stakeholder interviews reveal a sensitive and nuanced *'political economy'* around the question of developing TDM opt-out standards, and the role of REP. A common theme emerging from rights holders and solution providers interviews is the perception that **AI developers wish to push for REP to be recognised as a standard TDM opt-out measure**, with some suggesting that this coincides with **resistance to adopt any other standard** approaches which may require **high implementation costs for developers**.

Some interviewed stakeholders from the category of 'content providers' reported on the challenges they face regarding inherent limitations of REP, including the fact that:

([228]) See Robots.txt website (accessed 14 March 2025).

([229]) The IETF is a standard setting organisation that develops voluntary standards for the internet.

([230]) See Robots.txt website (accessed 14 March 2025).

([231]) See IETF Datatracker website (accessed 14 March 2025).

- It can only express a reservation, but **cannot enforce its compliance**: it is the crawler itself which has to be programmed for skipping some content if its user-agent identifier is indicated in some of the “Disallow” file’s directives.
- It is **not possible** to indicate a group of **agents under the same company or under the same category**.
- It is not designed for complex policy expressions, or even expressing policy based on actual **material use** of the gathered data.
- It lacks **granularity**, as entire files (such as HTML pages or other file formats like images) can be marked for exclusion, but not specific content within them (e.g., a particular text inside an HTML page) ([232]).
- Any **policy changes after crawling are not easily taken into account**. This is particularly significant as large AI models take a lot of time and effort to create and tend to be rather long-lived, whereas policies may change quite rapidly (Jiménez, J. & Arkko, J., 2024);
- It is under the **website’s administrators’ control**, which can or cannot be the direct right holder of the website's content.

The following subsections explore in detail some of these challenges related to the **practical implementation of REP as an opt-out mechanism,** as well as some of the solutions explored to address them and the role of REP in the broader AI ecosystem.

3.4.2.1.3 Enforceability of REP

REP is a **voluntary mechanism** which relies on the good faith of crawler deployers. It has **historically been a voluntary standard** based on good faith and mutual trust between webmasters implementing the protocol and entities that use crawlers to perform web scraping. This approach is reflected on the **original REP website,** that has not been updated since

([232]) However, it is possible to apply the reservation on an image (or other type of content) referenced by a webpage, because the image file is external to the HTML file and has its own URL.

2007, which states that *"There is no law stating that /robots.txt must be obeyed, nor does it constitute a binding contract between site owner and user, but* ***having a /robots.txt can be relevant in legal cases****."*([233])

As previously discussed, web scraping in violation of website terms and conditions may give rise to breach of contract claims, depending on the context. As for REP, the general sentiment in the industry seems to be that due to its voluntary nature, it is **unlikely to constitute a legally binding contract.**

However, *if* REP is deployed as a TDM opt-out mechanism, the cause of action for violating the *robots.txt* instructions and engaging in acts of reproduction or extraction of data for the purpose of TDM may not be a breach of contract, but an infringement of copyright and related rights. It is therefore important to **distinguish between** the issues of '**REP compliance in general**' and '**REP compliance in the context of a TDM opt-out**'. The issue of compliance with REP is not new and is closely linked to its development as a *de facto* standard to express preference for content indexing.

3.4.2.1.4 REP and Use-Differentiation

The various proposals to adapt REP listed in *Annex XII* can be seen as a response to demands for an opt-out solution that is based on **existing REP principles** but allows for the **disaggregation of different uses**. However, while these proposals differentiate between crawling for general purposes (like search engine indexing) and crawling for AI uses, they do not further distinguish between different types of AI uses such as (i) general AI model training, (ii) GenAI model training, and (iii) retrieval and inference for RAG.

The issue of use-differentiation is driven by two considerations and is critical for both rights holders and wider civil society. First, some rights holders indicate a willingness to allow some forms of TDM, but not TDM for AI training purposes. Others may wish to allow their content to be used for AI model training, but not specifically for commercial GenAI purposes. Finally, others may wish to specifically prohibit scraping when used for inference and contextualisation

([233]) See Robots.txt website - FAQ (accessed 14 March 2025).

within RAG applications. The complexity of the AI ecosystem thus suggests that an opt-out mechanism should **reflect different use cases throughout the data value chain**.

Second, REP has been traditionally used for indexing and archival purposes, which is a very distinct use from data acquisition for AI training. There is a persistent concern amongst many stakeholders that using REP as a TDM opt-out mechanism will not only block unwanted bots which scrape data for AI training, but also **bots which are used to index content on search engines** and ensure that a right holder's content can be discovered and found on the open web. This problem is made more acute with regard to large technological companies operating in both the search engine and AI development fields and use web scraping for both purposes. A number of stakeholders demand that such companies use different identifiers for bots based on their specific purposes. The REP revision proposals (listed in *Annex XII*) all aim at this, though through different mechanisms.

A September 2024 Policy Brief from the European Commission notes that *"large players offering generative AI foundation models may use their market power to limit choice or distort competition in downstream markets, when distributing and commercialising AI applications"*([234]), giving the example of possible tied-selling of an AI Model with Search Engine Services (Kowalski, Volpin, & Zombori, 2024). However, at this stage there does not appear to be any public discourse on disaggregating 'web scraping for search engine indexation' from 'web scraping for AI training data acquisition' as a **potential competition law issue**.

3.4.2.1.5 User-Agent Information Asymmetry

A key limitation of REP for rights holders is that depending on the website's specific strategy, it may need to constantly monitor for market developments as new AI companies and dataset developers release new crawlers. Furthermore, there is **no affirmative obligation on a company to publicly announce the identifier for its crawler**. This information asymmetry may affect a website's decision as to whether it wishes to use a **blacklist approach** (i.e., disallow all bots unless otherwise specified), or a **whitelist approach** (i.e., allow all bots unless otherwise specified). Furthermore, there may be an incentive in delaying the

([234]) Competition Policy Brief, Issue 3, European Commission, September 2024 (accessed 14 March 2025).

announcement of a crawler identifier only after it has already engaged in scraping. This has led to the development of a number of **solutions and resources to list crawler identifiers** used by AI companies that websites may wish to block ([235]). Services offering APIs that can be used to automatically update a website's *robots.txt* file as new AI bots and scrapers are announced have also emerged ([236]). Such services aim at lessening the burden on webmasters to monitor the market for developments in new AI crawler user-agents and manually modify their websites' *robots.txt* files.

3.4.2.1.6 Pre-emptive User-Agents

As discussed above, REP (i) lacks a broad AI-use-specific reservation option, and (ii) it sometimes requires proactive monitoring for new AI user-agents. In this context, some stakeholders have recommended using REP to **pre-emptively disallow user-agent identities** based on their purposes (even though REP does not currently have use-differentiation in-built into its protocol). Such an approach would allow rights holders to pre-emptively declare their opt-out reservations, before the REP standard is revised, and a new crawler identifier is accounted for (even if at the technical level such declaration is ineffective). This recommendation has been made by the Czech Association for Internet Development (SPIR) to disallow the non-existent *'MachineLearning'* user-agent in the *robots.txt* file ([237]).

SPIR described this as 'a proposal for standardised communication', and effectively using REP as an **existing platform to communicate new instructions** which are actually outside of the established REP protocol. Crawlers would thus agree to read *"User-agent: MachineLearning"* not as an instruction to a bot identified as *'MachineLearning'*, but rather as a broad indication of use-based TDM opt-out. This approach may possibly only meet the relevant legal requirements for a valid TDM opt-out under CDSM Article 4, where the parties pre-emptively agree to recognise the instruction as such, and where the crawler incorporates this into their technical interpretations of *robots.txt* files.

([235]) For example, see Github website (accessed 14 March 2025).

([236]) For example, see Dark Visitors website (accessed 14 March 2025).

([237]) Online Publishers Artificial Intelligence Datalines, SPIR, 7 July 2023 (accessed 14 March 2025).

This approach might be seen as using the *robots.txt* file as a **form to make a unilateral declaration**, with instructions that do not map to an actual technical effect within the confines of the REP protocol, but takes advantage of the fact that the file is known to be read by crawlers.

3.4.2.1.7 Referencing Terms and Conditions within REP

As discussed above, there are various limitations inherent to REP in terms of its ability to define permissions on a granular and use-differentiated level. At the same time, a key benefit of REP is that it is **well-established as a technical measure** and is understood to be **inherently machine-readable**. An advantage of legally-driven solutions such as website terms and conditions (discussed in *Section 3.4.1.4* above) is that **natural language can be crafted to be as specific** and granular as a right holder desires. However, such natural-language measures have a drawback of unguaranteed visibility, and even contentions about meeting the machine-readability criterion.

In this regard, a new approach is developing that consists of cross-referencing website terms and conditions within the *robots.tx*t file. Introduction of such natural language instructions into *robots.tx*t does not alter REP instructions, particularly as natural language comments are explicitly ignored by crawlers under the REP framework. However, these comments may be useful to **facilitate further awareness of the existence of an opt-out position** (i.e., increase the effectiveness of the opt-out communication and increase the probability that it is successfully detected). This may be the case if a TDM user goes beyond REP and incorporates 'state-of-the-art technologies' for identifying rights reservations (in the sense of AI Act Article 53(1)(c)), particularly as consulting *robots.txt* files is a standard practice during web crawling. Furthermore, natural language instructions into *robots.txt* might also provide **a direct link to the website's terms and conditions page** bringing further attention of these conditions to TDM users.

In this regard, the practice of cross-referencing natural-language website terms and conditions within machine-readable REP, is a corollary measure to incorporating machine-readable language (e.g., "<TDM-RESERVATION: 1>") in natural-language website terms (discussed in

Section 3.4.1.4 above). The box below is an example of an excerpt from Meta's REP instructions for Facebook, which illustrates this approach.

The practice of **interconnecting legally-driven measures** (website terms and conditions) **and technical measures** (REP) is interesting, as there appears to be a general disconnect between the two on the internet. As noted by the OECD, *"There is also a disconnect between the restrictions stated in website terms of service and the actual technical measures in place, as many websites do not correctly configure their robots.txt files to reflect contractual restrictions"* ([238]).

https://www.facebook.com/robots.txt

Notice: Collection of data on Facebook through automated means is

prohibited unless you have express written permission from Facebook

and may only be conducted for the limited purpose contained in said

permission.

All authorized user-agents listed on this page must comply with Meta's

Automated Data Collection Terms available at:

https://www.facebook.com/legal/automated_data_collection_terms

…

3.4.2.1.8 Beyond REP: blocking and redirection of crawlers

So far, this analysis has implied that REP directives are either implemented or not (as a possible form of TDM opt-out) and are then either respected or not (by crawlers). However, the behaviour of web scrapers is more complicated. There have been reports of some AI bots

([238]) OECD (2024), p. 33.

violating the good faith principles of REP by ignoring *robots.txt* directives, or masking their identities by not using known user-agent identifiers or attempting to mimic human traffic (commonly referred as spoofing) ([239]).

This has given rise to services that not only document bot identifiers, but **assess them based on adherence to good faith and transparency principle**s ([240]). Similarly, there is a demand for services that are able to assess and identify bad actors, particularly bots that mask their behaviour by generating automated server requests that attempt to mimic human traffic ([241]). In addition to REP, bot traffic may also be blocked at the server-level, where server management is controlled on the end of a website, rather than relying on the assumption that all crawlers will read and obey a *robots.txt file* ([242]).

Bots may also be blocked before they reach a server hosting a website. This has created **a market for bot-management services**, which can be extended to serve as a rights management system. A prime example of this is Cloudflare, a large integrated web-service company, which is used by some 19% of internet websites for network security ([243]). As discussed in *Section 3.8.2.1*, in 2024 Cloudflare has made its suite of bot management services specifically for blocking AI bots freely available ([244]). This suite activates traffic filtering rules between the protected webserver and the external internet network. Cloudflare appears to indicate that it may create a platform for direct licensing between rights holders and TDM users, through a marketplace for scraping ([245])

Another possibility to act against crawlers is to voluntarily mislead them. A number of stakeholders use measures to **act antagonistically against AI developers** and their crawling, including rights holders choosing to resort to data poisoning, as discussed in the *Section 3.8.1*. Another antagonistic approach consists of the 'honeypot approach', which uses configuration files to specifically allow AI bots and **redirect them to large files** (used for speed

([239]) Perplexity AI Is Lying about Their User Agent, Robb Knight, 15 June 2024 (accessed 14 March 2025).

([240]) See Cloudflare Radar (accessed 14 March 2025).

([241]) For example, see Cloudflare website: (accessed 14 March 2025).

([242]) For example, see mariusv/nginx-badbot-blocker, Github (accessed 14 March 2025).

([243]) See Cloudflare blog (accessed 14 March 2025).

([244]) Declare your AIndependence: block AI bots, scrapers and crawlers with a single click, Cloudflare, 3 July 2024 (accessed 15 March 2025).

([245]) Cloudflare Helps Content Creators Regain Control of their Content from AI Bots, Cloudflare, 23 September 2024 (accessed 15 March 2025).

testing) in order to waste the AI developer's resources and potentially disrupt training data acquisition ([246]).

3.4.2.2 TDM Reservation Protocol (TDMRep)

In 2020, the **Federation of European Publishers** approached **EDRLab** to create a solution aligned with **Article 4 of the CDSM Directive**. The outcome was the TDM Reservation Protocol ([247]) (TDMRep) that follows the CDSM directive closely. Although it specifies publisher policies, it does not delve into licensing, focusing solely on opt-out declarations. TDMRep provides a straightforward **boolean-based flag** for publishers to signal whether they permit TDM on their content. Such reservation is expressed via **two complementary properties** ([248]):

- ***tdm-reservation*** is a boolean value:
  - if it is set to 1 it indicates that TDM rights are reserved. If, at the same time, *tdm-policy* is informed, TDM agents may use it to get information on how they can acquire an authorisation to mine the content from the rights holders;
  - if its value is 0, then the TDM rights are not reserved;
- ***tdm-policy*** is a URL **pointing** to a TDM Policy set by the right holder.

This design prioritises simplicity, making it accessible for publishers with minimal technical resources.

TDMRep offers **both location-based and asset-based** content protection through four complementary implementation techniques ([249]):

- **TDM file on the website origin server (technique 1)**: it provides a **location-based** protection, since the TDM-protected resource is identified through the "location" property of a JSON object (See *Annex XVII*) defined in the file named **"tdmrep.json"**

([246]) Paste.Melanie.Lol website (accessed 14 March 2025).

([247]) TDM Reservation Protocol (TDMRep) - Final Community Group Report, W3C Community Group, 02 February 2024 (accessed 14 March 2025).

([248]) Ibid.

([249]) Ibid.

available in the website's repository. This file can contain one or **more JSON objects**, each expressing TDM reservations for a different resource. An example of possible content in "tdmrep.json" is:

```
[
 {
 "location": "/directory-a/",
 "tdm-reservation": 1
 },
 {
 "location": "/directory-b/html/",
 "tdm-reservation": 1,
 "tdm-policy":"https://provider.com/policies/policy.json"
 },
 {
 "location": "/directory-b/images/*.jpg",
 "tdm-reservation": 0
 }
]
```

In the example above, a web server is hosting three groups of files. The rights holders of the first group wants to express that TDM rights are reserved on these files with no possibility to acquire a TDM License. The rights holders of the second group of files (html pages) wants to express that TDM rights are reserved with a TDM Policy. TDM rights are not reserved for all JPEG images contained in the third group. Indeed, '*' is a wildcard that can be used in URLs.

- **TDM header fields in the server's HTTP responses (technique 2)**: It consists of configuring the server for adding TDM details in the **HTTP header** of the HTTP responses sent for delivering some content (HTML pages, images, and so on) to the requesting client. Since this is linked to a specific **web server configuration**, this is also a **location-based** protection. Currently this is the **preferred technique** for

implementing the protocol, as it is simple and already integrated in the **spawning.ai** API (see *Section 3.4.2.5.5*).

In the following example of an HTTP header, a TDM license may be acquired. The server returns a tdm-reservation header field with value 1 and a tdm-policy header field pointing to a TDM Policy:

```
HTTP/1.1 200 OK
Date: Wed, 14 Jul 2021 12:07:48 GMT
Content-type: text/html
tdm-reservation: 1
tdm-policy: https://provider.com/policies/policy.j
```

- **TDM metadata in website's HTML pages (technique 3)**: this way of using TDMRep is quite similar to the one described in point 2 and is again **location-based**, with the difference that the TDM reservation is **embedded in the HTML page** and covers **all the elements contained** within it. As a result, it enables only a limited level of granularity.

In the following example, an html document is associated with a TDM Policy through the **<meta> tags in its header**:

```
<!DOCTYPE html>
<html lang="en">
 <head>
 <meta charset="utf-8">
 <meta name="tdm-reservation" content="1">
 <meta name="tdm-policy" content="https://provider.com/”>
 <title>Document title</title>
 </head>
 <body>
 ...
```

```
<!-- body content -->
...
</body>
</html>
```

- **TDM metadata in EPUB files (technique 4)**: EPUB is a widely used **digital book format** that allows to encapsulate – and thus, to express TDM opt-out for – text, but also other formats. It provides an **asset-based** protection since it is embedded in the EPUB document itself, as well as for all the file types compatible with EPUB. In the following example, an EPUB file is associated with a TDM policy through a pair of **<meta> tags** contained in the metadata section:

```
<package prefix="tdm: http://www.w3.org/ns/tdmrep#" ...>
 <metadata ...>
 <dc:title>Document title</dc:title>
 <meta property="tdm:reservation">1</meta>
 <meta property="tdm:policy">https://provider.com/policy</meta>
 </metadata>
</package>
```

EPUB file format

EPUB (Electronic Publication) is a widely used **digital book format** that supports reflowable text, multimedia, and interactivity, making it compatible with most e-readers and devices.

The EPUB format provides a structured method for representing, packaging, and encoding web content—including HTML, CSS, SVG, and other resources—into a single-file container. This container is based on the ZIP format and houses all necessary resources for rendering an EPUB publication. The key component within this structure is the Package Document, an XML file that

centralises metadata, defines individual resources comprising the package, and establishes the reading order ([250]).

3.4.2.2.1 TDMRep's Limitations and Further Developments

Similar to REP, if TDMRep can be used to express TDM opt-out, such reservations cannot be enforced, and it is up to TDM users to find a way to automate the parsing of the policies pointed by the 'tdm-policy' variable. TDMRep developers expressed concern that AI and TDM actors currently show little interest in opt-out systems like TDMRep, which serves as a "no trespassing" indicator rather than an enforcement tool.

As for rights holders implementing the protocol, in case of use of technique 1, 3 or 4, some manual work is needed, for each asset to protect, to write down the protocol directives: this may cause accidental errors and conflicts in case more than one technique is implemented regarding the same content. In that respect, the protocol also defines a priority between the four different types of implementations, to interpret it correctly and address possible conflicts.

In case of the implementation technique 1, 2 or 3, the protection is location-based: this means that, in case the copyright holder and the website owner are two different entities, they must agree on TDM policies.

In case of the implementation of technique 4, **embedded metadata can be subject to tampering**, some signature-based verification methods to address this issue and detect such tampering are being discussed. They would support flagging instances where critical TDM opt-out metadata have been removed. In principle, such robustness checks can be integrated into standards themselves.

Some publishers also expressed concerns about their **content being stripped of metadata** before being processed by AI. If TDMRep developers attempt to mitigate this risk by supporting **metadata embedding**, they recognise that further protections are needed. Integration with **blockchain** is also under evaluation. This technology would primarily be used to handle policy-level data, to facilitate transparency and accountability.

([250]) EPUB (Electronic Publication) File Format Family, Library of Congress, 6 May 2024 (accessed 14 March 2025).

Regarding the level of granularity of opt-out reservations, TDMRep has had some coordination (with reference to **PDF** files in particular) with the developers of the **C2PA protocol** (see *Section 4.3.1.1*) regarding the granularity of opt-out reservations, i.e., what different permissions can be flagged based on different **TDM uses**: both protocols aim to allow flags for **TDM, AI training, and GenAI training**.

As for adoption of the protocol beyond the book publishing sector, specific support for HTML and EPUB files integration has been provided as they are commonly used file formats in the **news publishing industry**. EDRLab's members also stated interest in expanding TDMRep support across more formats and simplifying implementations through HTTP headers, which are compatible with cross-media applications. However, if the news publishing sector is more opened to technical solutions like TDMRep for item-level control, adoption of such a protocol may prove more challenging with other content sectors, like the music industry that seems to be more interested by a legally driven and catalogue-based approach to TDM reservation.

3.4.2.2.2 TDMRep: Market Maturity

Interviews with stakeholders revealed that adoption rates of the protocol vary significantly between countries, with estimated peak adoption rates of **50%-80% among trade publishers and 70% among publishers of learning materials in Finland**. Overall, TDMRep has been predominantly adopted in Europe, particularly in text-based content sectors such as **trade publishing**, **Scientific, Technical, and Medical (STM) publishing**, **digital content distribution platforms**, and **newspapers**.

However, the available statistics are not exhaustive. As the full technical specification is freely accessible on the W3C CG website, many rights holders have implemented TDMRep independently without notifying the working group. A publicly available list of adopters is regularly updated ([251]).

In Italy, some major trade publishers and digital platforms, including **Mondadori**, **Casalini**, and **Edigita,** have incorporated TDMRep into their EPUB files, ONIX metadata, and websites ([252]).

([251]) w3c/Tdm-Reservation-Protocol, Github (accessed 30 January 2025).

([252]) ONIX metadata is a standard used in the publishing industry to describe and exchange book information, including title, author, ISBN, pricing, availability and rights.

In Germany there are some major companies also adopting TDMRep including **Bookwire** that distributes e-books from 3,000 publishers worldwide.

In France, there has been a growing adoption of TDMRep, with notable adopters including **Le Parisien**, **Le Télégramme, Radio France, France Bleu**. As of March 2025, out of 442 websites belonging to media, books and news publishers, over 24%, **integrate TDMRep in their websites** ([253]).

Beyond the EU, the protocol has been adopted by global publishing entities, including **Penguin Random House, American Chemical Society, Springer Nature, American Psychological Association, and IEEE**. As highlighted in stakeholders' interviews, STM publishers in particular have demonstrated a high level of interest for the protocol.

3.4.2.3 C2PA Training and Data Mining Assertions

This protocol, fully detailed in *Section 4.3.1.1* on Provenance Tracking Solutions, was initially developed to address the prevalence of misleading information online through the development of media standards for **certifying** the **provenance** of media content.

Moreover, the protocol also includes the possibility to bind machine-readable 'Training and Data Mining Assertions' to media files. These assertions are stored within a C2PA manifest, which is incorporated into the file's metadata. Their integrity and authenticity are preserved using cryptographic techniques.

The Training and Data Mining Assertions enable differentiation between various TDM processes, **including Data Mining, AI Training, GenAI Training, and AI Inference**. These assertions specify whether a particular TDM **action is permitted, prohibited, or subject to conditions**. The syntax of these TDM assertions is detailed in *Annex XIII*.

([253]) La Liste Des Sites Qui Ont Adopté Le Protocole TDMRep (in French), Datawrapper (accessed 30 March 2025).

3.4.2.4 JPEG Trust: Rights Declaration Solution (Under Development)

As detailed in *Section 4.3.1.3* below, Joint Photographic Experts Group (JPEG) is working on expanding JPEG Trust Core Foundation to also include rights and ownership declarations, embedding that information into media's metadata (providing asset-based protection). The company is adopting the W3C recommendation Open Digital Rights Language (ODRL) and the C2PA Training and Data Mining Assertions as a reference for data formats.

3.4.2.5 The solutions provided by Spawning.ai

Spawning ([254]) is building a set of ecosystem-wide solutions aiming at addressing the needs of both rights holders and AI developers on TDM reservations.

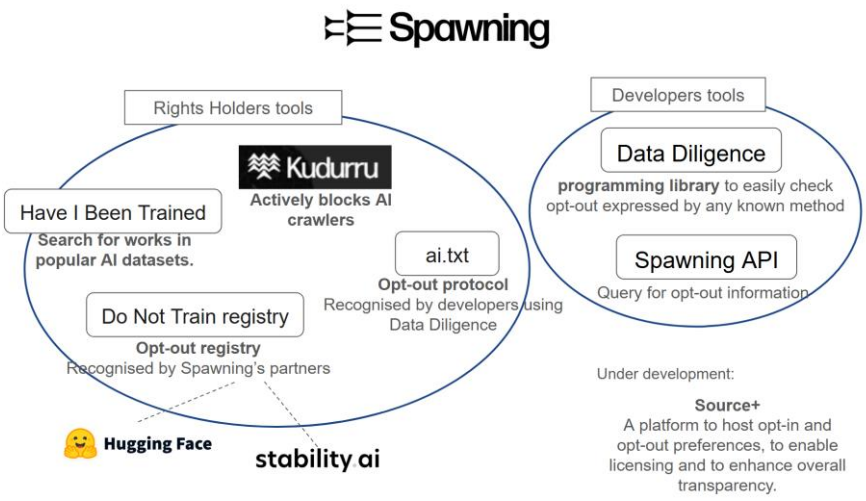


Figure 3.4.2-3: Overview of the solutions deployed by Spawning.ai.

Spawning.ai's **Do Not Train Tool Suite** consolidates machine-readable opt-out methods around a Do Not Train Registry, providing a set of tools for rights holders and AI developers to support the expression of and compliance with TDM reservations.

([254]) See SpawningAI website (accessed 14 March 2025).

3.4.2.5.1 Do Not Train Registry

The Do Not Train Registry (DNTR) lists the TDM opt-out expressed (data-use reservations). Rights holders can request inclusion in the registry via Spawning.ai's **APIs** to opt-out their entire domain or specific works. For example, Shutterstock ([255]), by registering its domain in the DNTR, automatically opted-out more than 400 million media URLs, and 200,000 new images that are uploaded to its site daily ([256]).

The **verification process** for opt-out requests remains informal and resource-intensive, requiring email correspondence (for individual creators), cross-referencing submitted works, and formal agreements (for rights holders organisations and CMOs).

The rights reservations established in the DNTR are designed to align with the specifications outlined in Article 4(3) of the CDSM Directive. These reservations are machine-readable and structured to assist AI developers in seamlessly integrating them into their data workflows (see below) ([257]).

3.4.2.5.2 Ai.txt Protocol

Complementing the APIs, Spawning.ai has introduced the **ai.txt** protocol, a machine-readable file designed to be placed in the root directory of a website. ai.txt files can be created directly from Spawning.ai's website (see *Figure 3.4.2-4*) and enable website owners to communicate data-use reservations for each type of content.

([255]) Shutterstock is a stock photography, video, and music platform that provides licensed images, footage, and audio for creative projects. It offers a vast library of royalty-free content for businesses, marketers, and creatives.

([256]) The Spawning Guide to Rights Reservations, Spawning Blog, 26 March 2024 (accessed 15 March 2025).

([257]) Ibid.

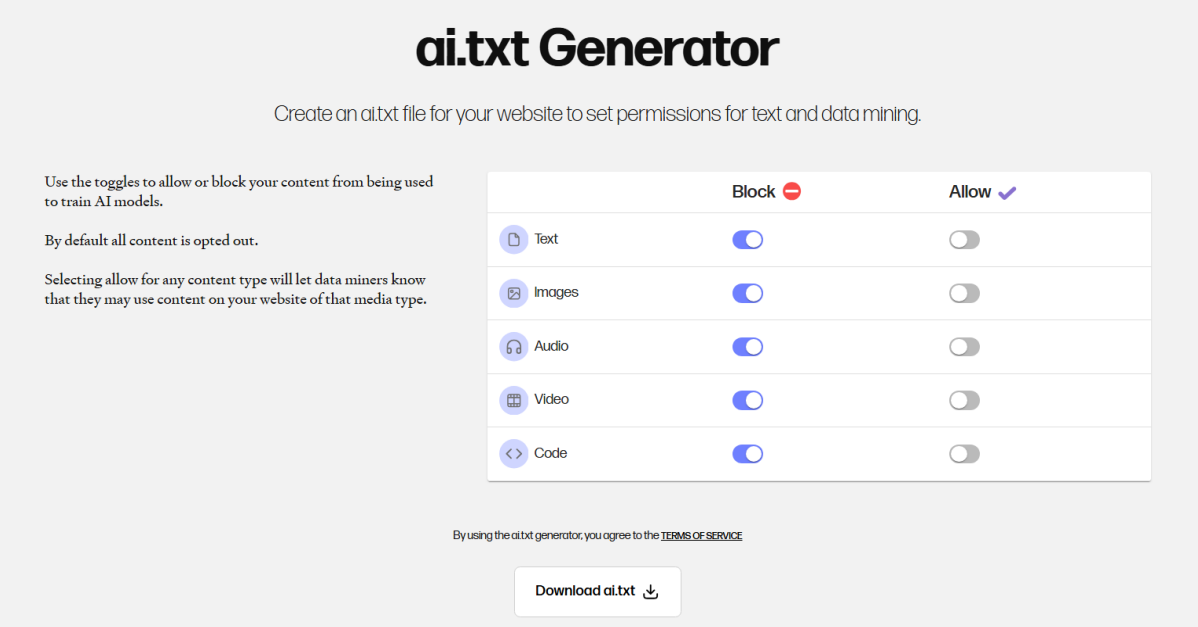


Figure 3.4.2-4: Online form allowing the download of a customised ai.txt file ([258]).

As shown by *Figure 3.4.2-5*, the resulting file's syntax is quite similar to the one of robotx.txt. The particularity of this protocol is the extensive use of the '*' wildcard, which can be used to indicate zero or more characters without limitations.

```
# Spawning AI
# Prevent datasets from using the following file types

User-Agent: *
Disallow: *.txt
Disallow: *.pdf
Disallow: *.doc
Disallow: *.docx
Disallow: *.odt
Disallow: *.rtf
Disallow: *.tex
Disallow: *.wks
```

Figure 3.4.2-5: Example of an ai.txt file content.

([258]) Create an ai.txt, Spawning AI (accessed 14 March 2025).

ai.txt supports the expression of **different reservations for each content type**. However, as Matt Rogerson (Guardian News & Media) pointed out ([259]), this change may be of limited use "*because there is no reason to think all the content of a particular file format on a website is either suitable or unsuitable for consumption by AI.*"

Another difference between REP and ai.txt is that REP is usually consulted before the site is scraped, whereas ai.txt intervenes before the files are actually downloaded. This brings some advantages with regard to a number of AI datasets that do not contain the actual content to be used for AI training, but instead provide a link to the content ([260]).

In the case of a right holder expressing an opt-out for an image linked from the LAION-5B dataset (see *Section 3.1.2.1.3*), REP would be ineffective as LAION-5B provides the URL from a website that has already been scraped and where the image can be retrieved. In this case, the crawler may skip the "discovery phase" where robots.txt is typically checked. This applies to AI training datasets beyond LAION-5B, for instance, ImageNet.

In contrast, the ai.txt protocol functions at the point of image retrieval, thereby preventing the download of the image even if the website scraping has already occurred. In addition, if questions remain on the optional nature of TDM opt-out expressed through REP, in contrast, ai.txt **takes direct aim at the EU TDM Article 4 exception** by explicitly providing a machine-readable opt-out.

Spawning.ai provides tutorials on how to deploy ai.txt on common website builders such as WordPress, Squarespace and Shopify. The protocol is also compatible with the Data Diligence developer suite (see below), enabling AI developers to easily parse it.

3.4.2.5.3 Have I Been Trained?

"**Have I Been Trained?**" is an online tool allowing rights holders to search for their works in **LAION-5B,** one of the most used AI training datasets.

([259]) Guardian news & media Draft paper on an ai.txt protocol, Guardian News & Med, IETF, 9 August 2024 (accessed 14 March 2025).

([260]) Ai.Txt: A New Way for Websites to Set Permissions for AI, Spawning AI, 30 May 2023 (accessed 1 February 2025).

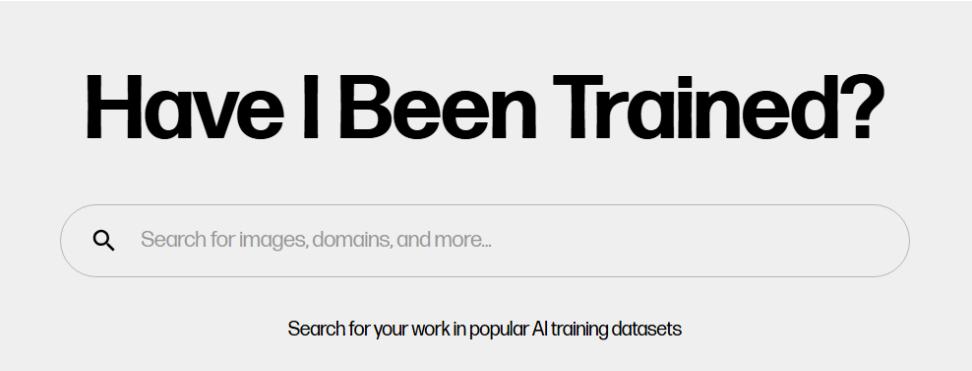


Figure 3.4.2-6: Main page of the 'Have I Been Trained' online tool ([261]).

"**Spawning.ai Browser Extension**" highlights scraped content while surfing the web by checking if the media appears in the LAION-5B training dataset ([262]). It can be integrated into the web browser to support the inspection on any webpage consulted.

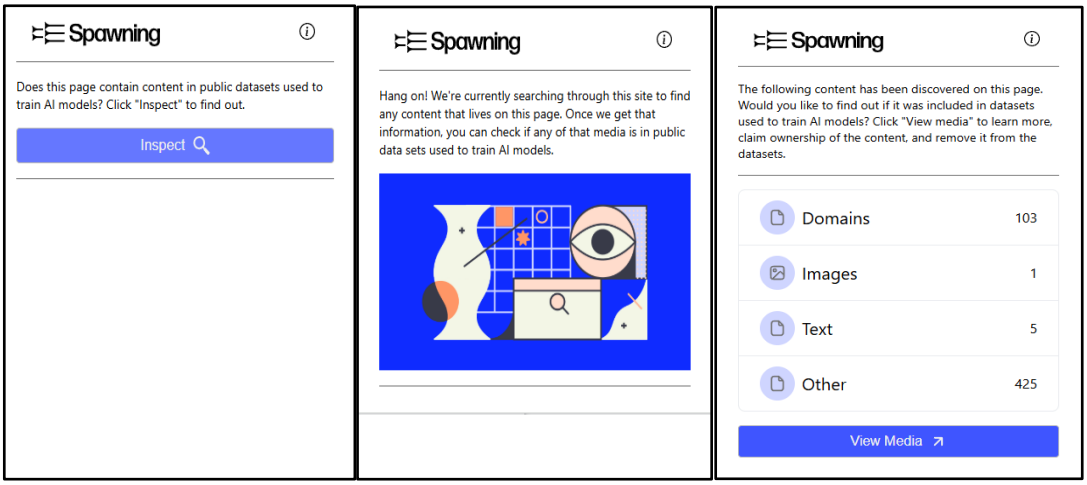


Figure 3.4.2-7: Screenshot of the "Have I Been Trained" browser extension, allowing the inspection of a given webpage to detect if it hosts content which is also present in the LAION-5B dataset.

([261]) Have I been Trained?, Spawning AI (accessed 24 February 2025).

([262]) Spawning AI website (accessed 8 November 2024).

Interviews revealed that work is in progress to expand the search beyond the LAION-5B, into more training datasets.

3.4.2.5.4 Kudurru

"Kudurru" is a software that **actively blocks AI scrapers**. It monitors popular AI datasets and dataset providers for scraping behaviour and coordinates amongst the network to quickly identify scrapers. When a scraper is identified, its identity is broadcasted to all Kudurru-protected sites that can collectively block the scraper from downloading content from their respective host. According to Spawning, it is **more efficient than a simple opt-out because it cannot be ignored** ([263])**.** Moreover, it can generate logs and evidence that rights holders might use in legal actions against unauthorised data usage.

However, Spawning.ai may be **reconsidering the long-term support for Kudurru**, as concerns have emerged regarding the possibility it could undermine the principles of an open and accessible internet, its potential to intensify conflicts between rights holders and AI developers, and ethical and legal issues related to the methods it employs to block or misdirect scrapers.

3.4.2.5.5 Spawning.ai Developer Tools: Data Diligence and API

The DNTR is used by Spawning's partners such as Stability AI (media in the DNTR were excluded from the training of Stable Diffusion V3 ([264])) and Hugging Face, as well as other AI developers. Spawning provides an API for AI developers that allows them to check the datasets they use or develop against its DNTR.

In addition, Spawning is providing "**Data Diligence**" that is a programming library (written in Python language) that aims to make it simple for TDM users to respect opt-outs by providing

([263]) Spawning AI website (accessed 8 November 2024).

([264]) The Spawning Guide to Rights Reservations, Spawning Blog, 26 March 2024 (accessed 15 March 2025).

a consistent interface to check if a given work in the training dataset is opted-out using any known method. This means that the library not only integrates with the APIs of the DNTR, but it is also able to **check if the inspected media or its location contain any form of machine-readable information**, for example, by parsing the HTTP header ([265]). The aim is to **make respecting opt-outs as easy as possible, while being flexible enough to support new opt-out methods** as they are developed ([266]).

In that respect, Spawning.ai is already integrating multiple opt-out methodologies, including TDMRep (discussed in *Section 3.4.2.2*). This makes it an aggregator of opt-out information, which can be made available to AI companies through Spawning's APIs. Those services allow querying for opt-out information related to a specific URL or domain.

The company is also exploring techniques to allow developers to exclude opted-out data without identifying the specific content, preserving privacy and complying with data protection regulations.

3.4.2.5.6 Spawning.ai's Limitations and Further Developments

Spawning.ai's solutions support rights holders to express their TDM reservations, to verify if their content is used in AI training datasets (with its "Have I been trained") and to enforce such reservations (with its "Kudurru"). At the same time, they support AI developers in complying with TDM opt-out expressed (with its DNTR API and "Data Diligence" tools for AI developers).

The combination of these different solutions and the aggregation of opt-out information could be seen as an attempt to address the respective limitations of individual technical reservation measures, or perhaps it is a technology-oriented approach to reconcile the interests of both rights holders and AI developers. However, the company is still facing challenges with:

- The need for a **scalable and reliable solution** to handle a growing number of TDM opt-out requests to be added to its DNTR without compromising its accuracy. This

([265]) Ibid.

([266]) Spawning AI website (accessed 8 November 2024).

includes the need for a scalable **verification process** on the rights of individual rights holders or rights holders organisations submitting content for opt-out.

- The need for a viable business model, as Spawning.ai currently does not generate revenue from its opt-out service.

In that respect, Spawning.ai is developing Source+ as a comprehensive initiative aimed at establishing an ethical and **transparent** framework for the inclusion of creative works in **AI training datasets**. It is still **under development**, and Spawning.ai declares it is planning to further work on it in 2025 and beyond to enable fair compensation towards creators' opt-in. Through this initiative, Spawning.ai also aims to create an **open data market**, helping to set pricing standards for data used in AI training, recognising the challenge in establishing the data's value (See *Section 2.4*).

The Source+ platform is to be built around a dual mechanism. At its core, Source+ offers artists and rights holders the ability to **opt-in** or **opt-out** of having their works utilised for AI training, providing a structured mechanism through compensatory licensing agreements. On the other side, it facilitates developers' incumbency of excluding a specific work from training processes through the **machine-readable** "Do Not Train" registry.

One of the key features of Source+ is the emphasis on responsibly curated datasets. Spawning.ai has developed practices that ensure only ethically sourced data, such as public domain or CC0-licensed content, is used. This includes the **validation of licensing** information to exclude content with questionable status, thereby reducing legal risks and supporting the ethical standards of AI training processes. The company is also planning to introduce a **single training licence option** in the first quarter of 2025. This will allow AI developers to license data for a one-time training purpose, with clear terms and compensation structures, aiming to **simplify the licensing process**, making it more accessible for both **small and large AI developers**.

Source+ has facilitated partnerships with major AI platforms such as **Stability AI** and **Hugging Face**, ensuring that the reservations of rights holders are respected throughout various development environments. These collaborations exemplify a commitment to integrating consent into AI practices across the industry, setting a precedent for ethical data use and a shift towards **transparency**.

3.4.2.6 The Use of Liccium Infrastructure for TDM Opt-out Management

Liccium is an organisation that provides rights holders with a platform to digitally sign and protect their original works, building trust in ownership, attribution, and authenticity of digital media content ([267]). The platform allows content creators to sign their original works, and publicly declare ownership and metadata associated with their works using **International Standard Content Code (ISCC)** based content fingerprinting (see below) and soft-binding technology. This ensures that works remain verifiably linked to their claims, even if the content is modified or metadata is removed.

Metadata and rights information are stored in a federated registry system, underpinned by Liccium's **Trust Engine**, which ensures cryptographic integrity and prevents tampering (see *Annex XI.6* for details on Federated Registries). The use of the ISCC, detailed below, makes this technology a prominent example of a **fingerprinting-based reservation solution** ([268]).

Liccium's Trust Engine allows different **sectors** (e.g., publishing, music, and news) **to maintain separate registries** that can interact seamlessly. It achieves this by leveraging a **decentralised network of registries**, designed to be scalable, where each node can operate autonomously and periodically sync with the other registries, enabling consistent data integrity. This multi-registry setup enables more **tailored management of rights across different types of content**, supporting flexibility and scalability.

To ensure that the data stored in the federated registries is consistent, Liccium leverages **digital signatures and identity verification**.

Liccium uses **W3C Verifiable Credentials**, preventing unauthorised parties from creating fraudulent declarations. The standard provides cryptographic guarantees of the right holder identity. Liccium's registries also support **C2PA** manifests (see *Section 4.3.1.1* on C2PA) for secure documentation and origin verification of digital media assets. Liccium's infrastructure is designed for large-scale implementation, with current users managing millions of assets.

([267]) Liccium website (accessed 10 January 2025).

([268]) Other fingerprinting-based solutions like Audible Magic (*Section 4.3.3.2.2*) could potentially also function as an opt-out measure.

The **architecture** incorporates a distributed hash table (DHT), storing only essential keys and values, that helps maintain a lightweight, scalable system while ensuring data consistency across multiple nodes. Furthermore, it is undergoing **enhancements** to improve scalability and synchronisation.

Liccium's platform and infrastructure can support different right management use cases with structured, machine-readable metadata declarations specifying if a work is publicly available, licensed under certain conditions, or explicitly restricted from specific uses. In that respect, the platform is developing as an **asset-based solution** for rights holders to declare TDM reservations and licensing terms for different AI-related uses.

On the right holder side, TDM opt-out declaration (and potential licensing terms) can be added to the metadata and digitally signed, ensuring immutability and authenticity.

On the AI developer side, by querying the registry for a specific ISCC code, they can validate content status and associated rights before ingesting (or not) the related content into their training processes.

The system facilitates **multi-tiered transparency**—providing public access to records of data used for GenAI training while maintaining restricted access for non-generative training (or general TDM), which remains available only to regulatory authorities. This tiered access model aligns with regulatory requirements, such as Article 4 of the CDSM Directive and Article 52 of the AI Act, ensuring compliance while balancing transparency with confidentiality.

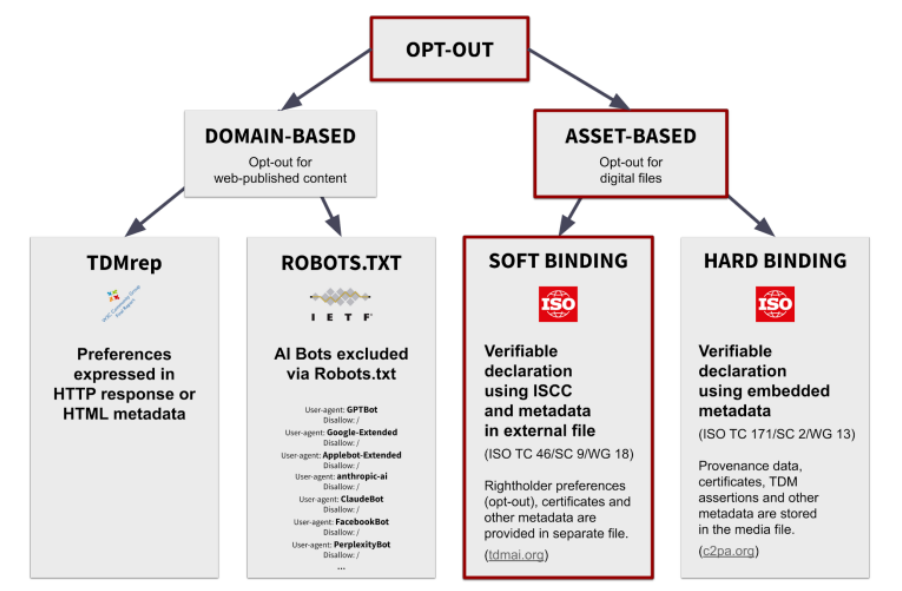
The **obligations of AI developers** implementing the system can be summarised as:

- **Generating ISCC** codes from the digital assets already in their systems;
- **Localising access to the federated database** for content validation (as API queries are not viable at this scale);
- **Conducting neighbouring similarity checks** to detect content that may be derivatives or near matches.

International Standard Content Code (ISCC)

ISCC serves as a **decentralised identifier** for digital content. Its development began in 2016 driven by Liccium with support from the **European Commission** and it is now an **open source** and public **ISO standard**. The ISCC Foundation provides **free, open-source tools for generating ISCC** codes, providing an open framework fostering competition by allowing different platforms to implement ISCC, facilitating diverse solutions that cater to specific types of rights holders.

ISCC codes are used to maintain the reference to metadata and rights information throughout the content lifecycle. The ISCC's **soft-binding ([269])** method links metadata and opt-out declarations externally to the content file, preserving this information **even if embedded metadata is removed** during online sharing. This mechanism ensures metadata integrity even in environments where media files are shared on social platforms or undergo transformations, such as format changes or compression. *Figure 3.4.2-9* summarises the different methods for associating TDM reservations with the related content.



([269]) "Soft-binding" is used in contrast to "hard-binding", which is about using cryptographic techniques to link the asset to the related metadata (In contrast, C2PA is an example of use of hard-binding).

Figure 3.4.2-9: Different approaches for binding the opt-out information to the relative asset. 'Domain-based' is used as a synonym of 'location-based'. This schema highlights the approach enabled by the ISCC ([270]).

ISCC codes are created using a mix of cryptographic and similarity-preserving hashes. Unlike other standardised identifiers (e.g., ISBN, DOI, ISRC), ISCC codes are **derived directly from the content file**, enabling anyone with access to the file to independently generate the same or a similar identifier. This enables **unambiguous identification of identical content** or **probabilistic matching of similar content**. The codes can be calculated from **all file formats**. For different versions or formats of the same content, the identifiers differ but still align based on the degree of modification. Scalable technology for matching highly similar content, such as nearest neighbour search, supports this process. The soft-binding has the significant limit as it is **ineffective in tracking heavily modified images** (such as cropped or rotated images). To address this limit, Liccium encourages rights holders to register multiple versions of content to enhance match reliability. For text-based media, ISCC can tolerate **up to 20% text alteration without compromising match accuracy**. This reduces the risk of false positives and provide flexibility for minor text adjustments.

According to the ISCC white paper ([271]), these are examples of ISCC uses:

- Distinguishing different versions of the same content.
- Clustering similar content.
- De-duplicating and disambiguating content across repositories.
- Assigning identifiers to granular content chunks.
- Verifying data integrity or detecting data manipulation.

The Liccium platform implements **TDM.ai**, a protocol which is building on ISCC to bind robustly machine-readable reservations for TDM to digital media assets. It is specifically tailored for

([270]) Metadata Binding, TDM AI (accessed 18 December 2024).

([271]) ISCC Content Fingerprinting, Liccium (accessed 15 March 2025).

training models and applications of GenAI, and addresses the challenge of **metadata binding**, by leveraging the **ISCC code's soft-binding** method (see the previous section) ([272]).

TDM.ai integrates the ISCC code, **federated opt-out registries** and the **W3C recommendation for cryptographically verifiable credentials** as illustrated in *Figure 3.4.2-10.*

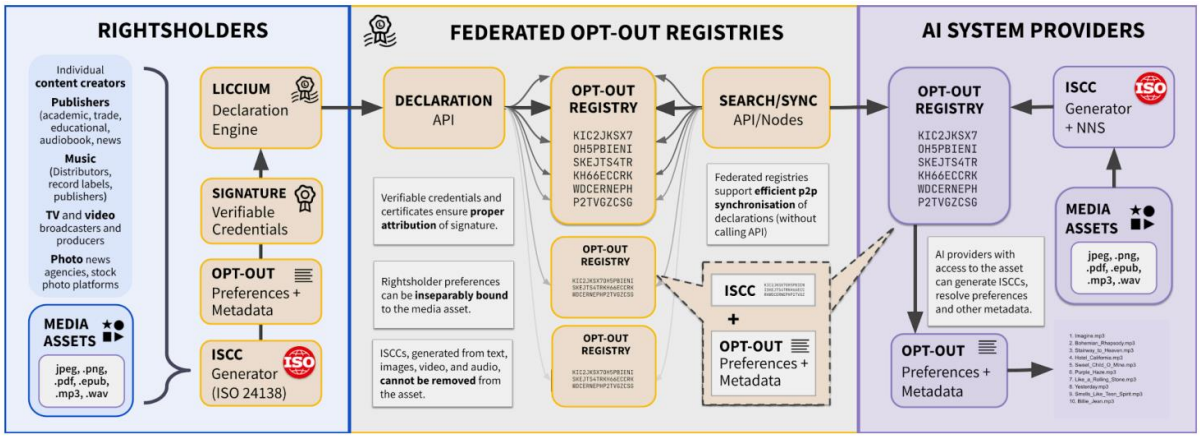


Figure 3.4.2-10: Infrastructure schema of the TDM.ai protocol ([273]).

The rights declaration can be resolved directly from the content-derived identifier, the **ISCC code**. Thanks to it, the protocol ensures a reliable method of identifying content that is robust to common problems such as metadata stripping and watermark removal ([274]).

The use of **verifiable credentials** enhances trust and verifiability, ensuring that the declarations are genuine and can be traced back to the original rights holders, depending on their privacy needs and preferences.

To enhance trustworthiness, it is recommended that declaration metadata include publicly accessible **Verifiable Credentials based on W3C standards**. All declarations are digitally

([272]) What is the TDM·AI Protocol, TDM AI (accessed 15 March 2025).

([273]) Ibid.

([274]) Ibid.

signed and provided with **timestamp**, an aspect that is often overlooked in opt-out discussions, but especially relevant for infringement cases ([275]).

ISCC codes and selected preferences can be publicly declared in open, centralised, or federated metadata directories. These directories link rights holders instructions to the unique ISCC code of the media asset. Directories must be publicly accessible to facilitate ISCC code discovery ([276]).

TDM supports valuable applications, and opting out may negatively impact rights holders by restricting the use of their works. TDM·AI seeks to establish a communication protocol to clarify rights holders' reservations, distinguishing between **general TDM, TDM for AI, and TDM for GenAI purposes**. The system also opens the possibility for copyright holders to **licence** their content. In addition, this solution can be used to **mark artificially generated content**, as required by Article 50(2) of the AI Act, without the use of watermarking ([277]).

The architecture includes, along with the Opt-out Registry, an Individual **Opt-out Confirmation Registry**, allowing AI providers to confirm that rights holders' reservations have been acknowledged and respected. This registry can be publicly accessible or permissioned, based on the provider's preferences and regulatory requirements. ([278])

3.4.2.7 Valunode Open Rights Data Exchange (ORDE)

Valunode ([279]) is an open infrastructure project that is under development and at the stage of a pilot project. The initiative is not strictly focused on TDM opt-out, but forms part of a broader effort to develop a copyright infrastructure facilitating copyright protection and content monetisation through a **marketplace for verifiable rights data**. In this context, the expression of TDM opt-out is just one use-case that such a pilot project could help address.

([275]) Ibid.

([276]) Ibid.

([277]) Ibid.

([278]) Ibid.

([279]) Valunode website (accessed 10 February 2025).

The project emerged from research led by the Copyright Infrastructure Task Force ([280]), with Valunode working on a secure and scalable rights data exchange, actively participating in EU programs to drive innovation. The company also collaborates with **TRACE4EU** (see *Section 4.3.1.4*), a consortium co-funded by the "Digital Europe programme", to build a service infrastructure ensuring traceability of digital rights.

Central to this effort is the development of a distributed marketplace, known as the **Open Rights Data Exchange ([281]).** The marketplace empowers rights holders to declare creative works, establish their rights across sectors, and obtain machine-readable registration certificates. Moreover, Valunode enables online platforms and rights users to access interoperable data necessary for licensing, distribution, and remuneration, which is aligned with the EU Data Governance Act. It aims at leveraging open standards such as C2PA, JPEG Trust, Dublin Core™ and W3C ODRL, to improve interoperability in the following fields:

- **Verifiable credentials** (to identify rights holders);
- Asset declaration, using lightweight and similarity-preserving **fingerprinting technology** to identify **all kinds of digital content** (text, image, audio, video) for **cross-sector applicability** (journalism, books, music, film, etc.).
- Tokenisation of media ([282]), to immutably bind authors' identities with content identification.
- Blockchain-based infrastructure for secure, resilient, and distributed verification of rights credentials, enabling decentralised provenance tracking and identity authentication.
- **Digital wallets** for managing, exchanging, and securely storing self-sovereign identity (SSI) credentials and rights-related metadata.

The **European Blockchain Services Infrastructure (EBSI) ([283])** is used in the context of the pilot project to store and facilitate the exchange of trusted rights data in a **distributed infrastructure**.

([280]) The Copyright Infrastructure Task Force aims to create a cohesive system that allows digital content to carry essential information about its origin, rights, and permissible uses. Acting as a standardisation forum rather than a standard development organization, the task force facilitates collaboration among EU member states and affiliates to address challenges posed by AI and digital content.

([281]) Open Rights Data Exchange, EBSI, European Commission (accessed 10 February 2025).

([282]) For the definition of 'Media Tokenisation', see the *Glossary*.

([283]) EBSI website, European Commission (accessed 14 March 2025).

As part of the project, rights holders will register rights and receive registration tokens. Rights users will query rights management information and receive trusted rights data. The following *Figure* schematises the designed process for asset registration:

- A songwriter can request (1) their **creator ID** (verifiable credentials);
- The creator ID is stored (2) in the artist's **digital wallet**;
- Each creator holding a valid creator ID can register (4) themselves on ORDE;
- To confirm the registration, ORDE must verify (5) the authenticity of the credentials against the cryptographically recorded data on the EBSI ledger;
- When a creator uploads a work, ORDE fingerprints it and immutably **binds this content fingerprint with the creator ID**. This binding is stored and **timestamped** (7) on EBSI;
- Also, ORDE will allow users to associate further information to a creation, such as the roles and creator IDs of other contributors and machine-readable terms and conditions for the use of the work. This information can either be public or stored in a permissioned database. ORDE will generate a **registration token** (8) pointing to several verifiable credential information.

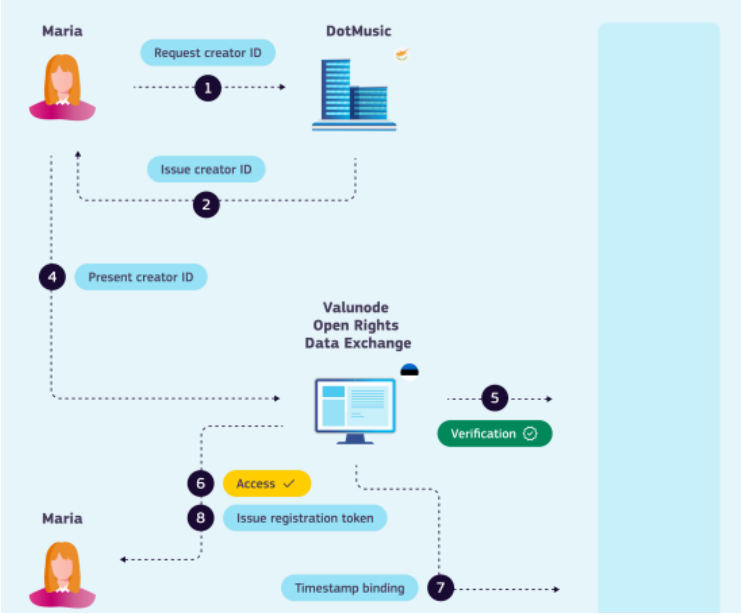


Figure 3.4.2-11: Open Rights Data Exchange use case ([284]).

([284]) Open Rights Data Exchange, EBSI, European Commission (accessed 10 February 2025).

ORDE will also enable the use of natural language processing, rights languages, and other tools to **convert narrated terms and conditions into machine-readable clauses**. This includes, among other functionalities, the ability to express TDM opt-out reservations. As per the latest publicly available details on EBSI's website, the ORDE project is expected to be completed by April 2025.

3.5 Comparison of Reservation Solutions

This section provides a comparison of the measures presented in *Section 3.4*, in relation to the criteria outlined in *Section 3.3*. The comparison is summarised into two tables: *Table 3.5-1* and *Table 3.5-2*, which focus on **legal** and **technical approaches** respectively. Each column represents a distinct technique, while each row is dedicated to a specific criterion. By providing a brief description in each cell, these tables are designed to encapsulate the key details elaborated upon in the corresponding chapters dedicated to each technology.

In general, all the solutions presented below allow rights holders (or their representatives) to express rights reservations **without providing any technical means to enforce such reservations.** TDM users are generally responsible for properly configuring their data collection policies, scraping tools, and data cleaning procedures, to ensure compliance with rights holders reservations.

Generally, as **legally-driven** measures are not restricted to applying to **content made publicly available online**, their typology might be described as applying at a **'work-based'** level (i.e., the intellectual property right in the intangible content, irrespective of the material copies in which the work is embedded), or even a **'repertoire-based'** level (i.e., a collection of protected works owned or managed by the same stakeholder). On the other hand, **technically-driven** measures are generally categorised as either **'location-based' and 'asset-based'**. Both approaches imply respective advantages and limitations, which are discussed in *Section 3.5.1* below. In certain cases such reservation measures may be complemented by certain non-reservation measures (outlined in *Section 3.8*) which might be used to restrict access to content or diminish its usefulness for AI training purposes.

	Unilateral Declarations	Licensing Constraints	Website Terms and Conditions
Typology	Work-based or **repertoire- based.**	**Work-based.**	**Location-based.**
TDM User Specificity	In principle, natural language permissions allow for a right holder to **differentiate permissions based on user type** (or even specifically identified users). Such differentiation however does not appear to be frequently used.	In principle, natural language permissions in licensing terms allow for a right holder to **differentiate permissions for specifically identified users or user types**.	In principle, natural language permissions allow for a right holder to **differentiate permissions based on user type** (or even specifically identified users). Such differentiation however does not appear to be frequently used.
Use-differentiation	Natural language permissions allow for **differentiation between permitted uses.**	Natural language permissions allow for **differentiation between permitted uses.**	Natural language permissions allow for **differentiation between permitted uses**.
Granularity	In principle, natural language permissions allow for granularity in permission. However, unilateral declarations are typically made at the repertoire-level.	In principle, natural language permissions allow for granularity in permission. Granularity follows from the **objects that are the subject of licensing** agreements.	In principle, natural language permissions allow for granularity in permission. However, Terms and Conditions are **generally meant to apply to the entire content** of a website (or page) and thus typically lack granularity.

	Unilateral Declarations	Licensing Constraints	Website Terms and Conditions
Versatility	Unilateral Declarations **may be used for any type of protected content** in any market.	Licencing Constraints **may be used for any type of protected content**, but specifically apply to the content exploited through a licensing agreement.	Terms and Conditions **can be applied to any type of protected content** but are specifically relevant for content distributed online.
Robustness	Unilateral Declarations are **highly robust** as they are controlled solely by the party making the declaration. However, this may be constrained by the difficulty of **locating** them.	Licencing Constraints are **highly robust** as they are fixed in licencing agreements.	Website Terms and Conditions offer moderate robustness as they are **controlled by the hosting website**. However, this may be constrained by the difficulty of **locating** them within the website's structure.
Timestamping	Unilateral Declarations are generally **de facto timestamped** in that they are often (but not strictly) made through public press releases or news items.	Licencing Constraints are timestamped as they are contained within a larger licensing agreement with an applicable **contractual Effective Date**.	Publicly available Website Terms and Conditions frequently include **timestamps indicating their publication and applicability**.

	Unilateral Declarations	Licensing Constraints	Website Terms and Conditions
Authentication	Unilateral Declarations are de facto made by rights holders themselves, and **so authentication is not an issue**. False Unilateral Declarations are likely to be rare and would amount to fraudulent assertions.	Licencing Constraints are contained in licensing agreements made by rights holders themselves, and thus **authentication is not an issue**. False licensing agreements made by parties who do not have legitimate rights over a protected work amount to fraudulent and infringing contracts.	Website Terms and Conditions are implemented by websites and online platforms that may host or link to copyright protected works. Website owners may or may not be the rights holders in hosted content. There is generally **no standard mechanism to validate** whether Terms and Conditions align with the exclusive rights of rights holders whose works may be hosted on such websites.
Intermediation	There is **no intermediation**, as a Unilateral Declaration is made directly by a right holder.	There is **no intermediation,** as Contractual Constraints are stipulated directly by a right holder.	Website owners may or may not be the rights holders in the content they host. For rights holders to express TDM reservations through Website Terms and Conditions, they must **coordinate with websites** that host copies of their works.
Openness	Legally-driven measures using natural language are available for **any right holder** to use.	Legally-driven measures using natural language are available for **any right holder** to use.	Legally-driven measures using natural language are available for **any right holder** to use.

	Unilateral Declarations	Licensing Constraints	Website Terms and Conditions
Ease of implementation	Unilateral Declarations are **easy and cheap to implement** on the right holder end, as they require mere public press releases or news items. For AI Developers, there are however significant **transaction costs** in searching out to determine whether a right holder has made a Unilateral Declaration, identifying the works actually covered by such a Declaration, and then incorporating this opt-out information into data curation methodologies.	Licencing Restrictions Declarations are **easy and cheap to implement** on the right holder end, as they require mere inclusion of TDM opt-out reservations in standard licensing contracts. For AI Developers, there are likely minimal costs in complying with these measures as the existence of the measure and the works to which it applies are explicitly known. The AI developer may incur cost in incorporating this information into data curation methodologies.	Website Terms and Conditions are **easy and cheap to implement** for users who make their protected works available via their own websites. They can however be costly if a right holder attempts to coordinate with third party websites on which their content is linked or hosted. For AI Developers, Website Terms and Conditions may be relatively-low cost to comply with in terms of **incorporating natural language processing into web scraping practices**, and the emerging trend of including explicit opt-out tags in Website Terms.
Flexibility	Unilateral Declarations are **highly flexible**, as a right holder can, in principle, make new Declarations at any point in time. However, with the difficulty of **locating** them, successive declarations may have a further negative impact on their robustness.	Licensing conditions are **not generally flexible,** as they require renegotiation of licensing terms or amendment of existing contracts.	Website Terms and Conditions are **highly flexible,** as website managers can, in principle, make new Declarations at any point in time.

	Unilateral Declarations	Licensing Constraints	Website Terms and Conditions
Retroactivity	Unilateral Declarations are strictly speaking **not retroactive**, though initial declarations can be used to clarify pre-declaration positions.	Licencing Restrictions are generally **not retroactive**.	Website Terms and Conditions are **not retroactive**.
External Effects	Unilateral Declarations are usually targeted specifically for AI training and TDM and have **minimal external effects**. However, they may or may not be interpreted as signalling a right holder's willingness to enter licensing negotiations.	Licencing Restrictions are typically highly specific and **unlikely to have external effects**.	**Overly Restrictive Website Terms** and Conditions can possibly hamper non-AI activities such as search engine indexing, or even certain privileged uses under copyright exceptions and limitations.
Generative Application	Legally-driven measures are not relevant to output transparency issues.	Legally-driven measures are not relevant to output transparency issues.	Legally-driven measures are not relevant to output transparency issues.
Offline Application	Unilateral Declarations are **explicitly applicable** to content not represented in digital copies (offline) just as much as digitised works.	Licencing Restrictions can be **applicable** to content not represented in digital copies.	Website Terms and Conditions are by definition **not applicable to** non-digitised works.

	Unilateral Declarations	Licensing Constraints	Website Terms and Conditions
Market Maturity	Unilateral Declarations are '**mature'** in that public statements regarding permissible activities have long been used by rights holders. However, full market maturity requires established practices for incorporating unilateral declarations into data curation practices.	Licencing Restrictions are **mature** in that they are the standard reference points for what licensees can and cannot do with content. However, at this stage it is unclear if and how TDM specific provisions are effectively included in licensing agreements.	Website Terms and Conditions are **mature** in that they are standard practice in web design. However, not all websites include AI or TDM specific provisions in their Terms and Conditions. Full market maturity requires established practices for web crawling to incorporate natural language processing to account for TDM restrictions in Website Terms and Conditions.

Table 3.5-1: Comparison between Legally-driven Reservation Solutions.

	Robots Exclusion Protocol	TDM Reservation Protocol (TDMRep)	C2PA TDM Assertions	ai.txt (Spawning)	Do Not Train Registry (Spawning)	JPEG Trust Core Foundation v2	TDM.ai protocol (Liccium)	Open Rights Data Exchange (Valunode)
Typology	Location-based.	Both **location** and **asset-based (hard-binding)**, depending on the adopted implementation.	Asset-based (hard-binding).	Location-based.	Location-based.	Asset-based (hard-binding).	Asset-based (soft-binding).	Asset-based (digital items are identified by their fingerprint).
TDM Users Specificity	The names of specific **bots must be listed** individually.	Can be specified within the 'tdm-policy' field, though the protocol **does not mandate a standardised format** for expressing user-specific restrictions.	Can be specified within the 'constraint_info' field, though the protocol **does not mandate a standardised format** for expressing user-specific restrictions.	No.	No.	Yes	No.	User-specific permissions could be **encoded within the machine-readable terms and conditions** associated with the asset.
Use-differentiation	No.	Could be specified via the 'tdm-policy' field. However, the protocol **does not define how to express this information in a standardised way**.	Yes. Differentiation between: Data Mining, AI training, GenAI training and AI-inference.	No.	No.	Yes. Differentiation between: Data Mining, AI training, GenAI training and AI-inference.	Yes. Differentiation between: TDM, AI training, GenAI training.	Use-differentiation could be **specified in the machine-readable terms and conditions** associated with the asset.

	Robots Exclusion Protocol	TDM Reservation Protocol (TDMRep)	C2PA TDM Assertions	ai.txt (Spawning)	Do Not Train Registry (Spawning)	JPEG Trust Core Foundation v2	TDM.ai protocol (Liccium)	Open Rights Data Exchange (Valunode)
Granularity	**File-level**: Reservations can be specified for each piece of content contained within a single file on the server, such as HTML files, images**,** and other digital assets. However, there may be a limit of the robots.txt file size taken into consideration by crawlers. The '*' character can be used as a wildcard to indicate more than one file in a row, for example all files with specific extensions such as*.jpg or *.pdf.	Supports opt-out declarations at **file, webpage and web server level**, depending on the adopted implementation.	**File-level**.	At the website level, different reservations can be specified for **each file extension**.	Both domain and asset-level (asset URLs).	Asset-level.	Asset-level.	Asset-level.

	Robots Exclusion Protocol	TDM Reservation Protocol (TDMRep)	C2PA TDM Assertions	ai.txt (Spawning)	Do Not Train Registry (Spawning)	JPEG Trust Core Foundation v2	TDM.ai protocol (Liccium)	Open Rights Data Exchange (Valunode)
Versatility	Inside robots.txt, the files' URLs on the server can be inserted **without any limitation regarding the format of the underlying content or its file extension**.	The location-based approaches **don't limit the file's format**, but the asset-based version currently is compatible only with **EPUB** archives, which can **contain files respecting a wide range of different formats**.	It supports **image**, **video**, and **audio** formats, while also providing **partial coverage for text** file formats such as PDF and HTML.	It works with all file extensions.	It works with each file having a URL **without any limitation regarding the format of the underlying content or its file extension**.	Primarily designed for **JPEG** files; future developments may extend support to additional media formats, such as video and audio ([285]).	It is compatible with any **file format that supports ISCC computation**. The list of supported formats, which includes many widely used extensions, can be found on Liccium's website ([286]).	Designed for use across **different sectors and file formats**.

([285]) JPEG Trust White Paper, ISO/IEC, 2024.

([286]) Ibid.

Robustness	The robots.txt file is managed solely by the **website administrator**, who is responsible for implementing protective measures against unauthorised modifications. Its robustness depends on proper **access control and server security configurations**. **Adherence by AI crawlers is voluntary** and cannot be enforced. With the **rapidly evolving trends of web crawlers,** if instructions are not updated with regard to new web crawlers, they are not taken into consideration.	The **location-based** implementations offer more robustness than the **asset-based** one. The latter can be vulnerable to **EPUB metadata tampering**, whereas location-based approaches require direct access to the web server, which is secured by admin credentials.	The **specifications were designed** for potential threat scenarios, with the risk of metadata tampering mitigated through the use of **cryptographic algorithms**. However, the protocol does not include mitigations against **metadata removal**.	The ai.txt file is managed solely by the **website administrator**, who is responsible for implementing protective measures against unauthorised modifications. Its robustness depends on proper **access control and server security configurations**.	Spawning integrates **cryptographic technologies** into the registry to ensure robustness and privacy.	It is susceptible to **metadata tampering**, but unauthorised modifications are detectable using cryptographic signatures and embedded metadata consistency checks. Thus, the **trust indicators** reflect the results of the tampering detection procedures.	Robustness is linked to the reliability of ISCC computation, that is supporting image or identification **even if slightly modified**. Soft-binding does not suffer from metadata tampering.	Designed for a **robust and privacy-preserving** process.

	Robots Exclusion Protocol	TDM Reservation Protocol (TDMRep)	C2PA TDM Assertions	ai.txt (Spawning)	Do Not Train Registry (Spawning)	JPEG Trust Core Foundation v2	TDM.ai protocol (Liccium)	Open Rights Data Exchange (Valunode)
Timestamping	Timestamping in REP depends on version control. If properly implemented, a **version control system** can log changes to the robots.txt file, preserving historical records of declared reservations.	In case of implementation of techniques **(1), (3) and (4)**, timestamping is not automatically included. When adopting the implementation of technique **(2) TDM Header fields in the server's HTTP responses**, timestamping is already integrated into the HTTP header enveloping the TDMRep data.	Yes, and from version 2.0 the use of timestamping Authorities (TSA) ([287]) has been standardised for ensuring timestamp's integrity.	The ai.txt protocol does not natively support timestamping, but changes can be logged through a **version control system**.	Yes.	Yes.	Yes.	Yes.

([287]) A Timestamp Authority (TSA) is a trusted entity, often implemented through a web service, that provides cryptographically secure timestamps to verify the existence and integrity of digital data at a specific point in time.

	Robots Exclusion Protocol	TDM Reservation Protocol (TDMRep)	C2PA TDM Assertions	ai.txt (Spawning)	Do Not Train Registry (Spawning)	JPEG Trust Core Foundation v2	TDM.ai protocol (Liccium)	Open Rights Data Exchange (Valunode)
Authentication	The only authentication mechanism is restricted file access: only **administrators** can modify the robots.txt file. However, this does not authenticate whether the administrator is the actual right holder.	**Location-based** implementations: only the **web server administrator** can configure TDMRep, but this does not authenticate whether the administrator is the actual right holder. **Asset-based** implementation does not have an authentication mechanism.	C2PA manifests are **cryptographically signed** when bound to the corresponding digital content. However, this process **does not inherently verify the signer's rights over the content**. Instead, the responsibility for verifying these rights is delegated to the content user.	The only authentication mechanism is restricted file access: only **administrators** can modify the ai.txt file. However, this does not authenticate whether the administrator is the actual right holder.	Yes, but a right holder's identification is **manual**.	Yes. The actual mechanism is not public since the standard is still under development.	Yes. Implemented through **W3C recommendations for verifiable credentials**.	Yes. Implemented through **verifiable credentials**.

	Robots Exclusion Protocol	TDM Reservation Protocol (TDMRep)	C2PA TDM Assertions	ai.txt (Spawning)	Do Not Train Registry (Spawning)	JPEG Trust Core Foundation v2	TDM.ai protocol (Liccium)	Open Rights Data Exchange (Valunode)
Intermediation	If the **right holder and the website owner are two distinct entities**, then an agreement between them is needed.	If the **right holder and the website owner are two distinct entities**, then an agreement between them is needed.	The right holder can embed information within the digital assets **without strictly needing any intermediation**.	The ai.txt file is available on the website to which it refers to. If the **website owner and the rights holders are distinct entities**, then an agreement between them is needed.	**Spawning.ai** manages the **registry** acting as an intermediary between AI developers and rights holders.	JPEG Trust provides tools for extracting and interpreting trust indicators **but does not act as a direct intermediary** between parties.	The **Liccium's infrastructure** hosts all the information linked to the ISCCs.	**ORDE's distributed infrastructure** will act as an intermediation mechanism, connecting creators and AI developers.

	Robots Exclusion Protocol	TDM Reservation Protocol (TDMRep)	C2PA TDM Assertions	ai.txt (Spawning)	Do Not Train Registry (Spawning)	JPEG Trust Core Foundation v2	TDM.ai protocol (Liccium)	Open Rights Data Exchange (Valunode)
Openness	Openly available. Described in **RFC 9309** as an informational protocol.	It is a **W3C specification**, but not a W3C standard.	It is an **open standard**, and some **open-source tools** have been developed to enable C2PA manifests management.	The Data Diligence package (which serves to properly parse the ai.txt file) is **open-source**.	The APIs connected to the registry are Spawning's Property. However, the company is **open for collaboration and interoperability** and the Data Diligence package (which can be used by AI developers to check for opt-out in the DNTR) is **open-source**.	It is planned to be released as an open ISO standard in 2026.	The whole suite is **open-source** and available on GitHub ([288]).	ORDE aims to **leverage existing open standards** under the intermediation of the **Copyright Infrastructure Task Force.**

([288]) See Github website on 'Liccium/Tdmai' (accessed 14 January 2025).

	Robots Exclusion Protocol	TDM Reservation Protocol (TDMRep)	C2PA TDM Assertions	ai.txt (Spawning)	Do Not Train Registry (Spawning)	JPEG Trust Core Foundation v2	TDM.ai protocol (Liccium)	Open Rights Data Exchange (Valunode)
Ease of implementation	**Rights holders** have to properly interact with **website owners**, which in turn have to **manually** compile the robots.txt file. The file may have to be updated each time a new AI scraper agent is declared or identified. **AI developers** can find scraping permissions expressed in an easy-to-parse way directly within the sites during scraping.	**Designed to be easily implemented** by website owners and rights holders. If some rights are reserved, and the variable ***tdm-policy*** points to a detailed rights declaration that is expressed in natural language, it **adds complexity for AI Developers**, who must find a way to automatically parse a possibly highly fragmented variety of policy declarations.	Some hardware devices (e.g., cameras) **automatically include** C2PA data in the media produced. Moreover, there are **open-source** tools for manipulating C2PA manifests, allowing the implementation of the protocol in a wide range of applications. This leads to a **fragmented range of implementation difficulties**, varying according to the specific use case.	Rights holders must request website owners to enable ai.txt. Implementation may be challenging if the **same website hosts content governed by multiple expressions of reservation**, which could lead to **conflicts** due to the limited granularity of the protocol. **AI developers** can exploit the available **Data Diligence** library.	**Rights holders** can benefit from **online forms and tools** appositely designed to be user-friendly. **Developers** can exploit the available **Data Diligence** library and API to query the Registry.	**Rights holders must use specific software tools** to embed machine-readable rights declarations into metadata**.** **AI developers require dedicated tools** to efficiently extract and process these rights declarations at scale**.**	**Rights holders** must create and preserve **verifiable credentials**. **AI Developers** can use the already-available **Liccium's library for ISCC generation**, then **search** in the federated registries for the **exact ISCC or similar ones** (to match also slightly modified content). This last step could be complex and time consuming if it must be repeated for many works.	**Rights holders** would have to master the tools for managing their credentials and the digital wallet of their declared asset. Meanwhile, **AI developers** may face challenges related to the security and privacy measures of the infrastructure, which may require additional and potentially complex implementation steps.

	Robots Exclusion Protocol	TDM Reservation Protocol (TDMRep)	C2PA TDM Assertions	ai.txt (Spawning)	Do Not Train Registry (Spawning)	JPEG Trust Core Foundation v2	TDM.ai protocol (Liccium)	Open Rights Data Exchange (Valunode)
Flexibility	Adjustments to the robots.txt file can be made **at any time**. However, these changes usually only apply to future web crawls and do not affect content already scraped. If a **scraped URL is already stored in third-party repositories or training datasets**, the updated robots.txt instruction does not retroactively get it removed.	**Location-based** implementations: changes are always possible. **Asset-based** implementation: Changes are not possible for already distributed files (though it could be applied to new files for the same content).	C2PA does not permit direct modification of a C2PA manifest, as its integrity is safeguarded by digital signatures. However, **a new, updated manifest can be added alongside the existing one**, each with its respective timestamp, enabling content users to identify the latest manifest. **Nevertheless, the inability to modify already-distributed assets** remains, meaning they may continue to convey outdated information.	Adjustments can be made at any time.	Adjustments can be made at any time.	Metadata embedded into already-distributed copies of the content cannot be modified anymore.	Adjustments can be made at any time.	Flexibility may be constrained due to the **immutable nature of blockchain** storage. However, updated rights information may be appended rather than replacing prior records.

	Robots Exclusion Protocol	TDM Reservation Protocol (TDMRep)	C2PA TDM Assertions	ai.txt (Spawning)	Do Not Train Registry (Spawning)	JPEG Trust Core Foundation v2	TDM.ai protocol (Liccium)	Open Rights Data Exchange (Valunode)
Retroactivity	It does not apply to content already scraped as it would require **AI companies to check for updates of the robots.txt file in already-scraped websites**.	No.	No.	No.	Yes.	No.	Yes.	Yes.
External Effects	The indicated content is ignored by **crawlers configured to respect REP,** including non-AI-related ones.	No.	No.	If a file with a certain extension is protected, then **all the files with the same extension** are automatically protected even if unintended.	No.	No.	No.	No.

	Robots Exclusion Protocol	TDM Reservation Protocol (TDMRep)	C2PA TDM Assertions	ai.txt (Spawning)	Do Not Train Registry (Spawning)	JPEG Trust Core Foundation v2	TDM.ai protocol (Liccium)	Open Rights Data Exchange (Valunode)
Generative Application	No.	No.	Yes.	No.	No.	Yes.	Could **potentially** be adopted to tag AI-generated content (with Liccium advocating for such application).	Possible, as ORDE could facilitate tagging or provenance tracking for AI-generated content.
Offline Application	No.	**Location-based** implementations: No. **Asset-based** implementation: Yes.	Yes.	No.	No.	Yes.	No.	No.

	Robots Exclusion Protocol	TDM Reservation Protocol (TDMRep)	C2PA TDM Assertions	ai.txt (Spawning)	Do Not Train Registry (Spawning)	JPEG Trust Core Foundation v2	TDM.ai protocol (Liccium)	Open Rights Data Exchange (Valunode)
Market Maturity	In 2024, **83% of the websites in the world** host a robots.txt file ([289]). However, this percentage most likely reflect to a very large extend on the use of REP with regard to search indexing.	It has already been adopted by some of the main **stakeholders** in the publishing sector across **different European countries**.	As a transparency mechanism, C2PA is supported by widely used **camera hardware, image editing software** and **AI models.** However, there is no data on its use as a TDM opt-out mechanism.	Respected by **Stability AI** and **Hugging Face**. Adopted by large media platforms such as **Shutterstock**.	Respected by **Stability AI** and **Hugging Face**. Adopted by large media platforms such as **Shutterstock**.	Under development.	Under development.	Under development.

Table 3.5-2: Comparison between Technical Reservation Solutions.

([289]) The 2024 Web Almanac: SEO, The 2024 Web Almanac (Vol. 6, Issue 7), HTTP Archive (accessed 14 March 2025).

3.5.1 Location-based versus Asset-based

As visualised in *Figure 3.4.2-9*, technically-driven opt-out declaration solutions can be divided into location-based (or domain-based) and asset-based (or content-based). Each approach brings its respective advantages and drawbacks, which are detailed in the *Table* below:

<table>
<tr><th rowspan="2"></th><th rowspan="2">Location-based (domain-based)</th><th colspan="2">Asset-based (content-based)</th></tr>
<tr><th>Soft-binding</th><th>Hard-binding</th></tr>
<tr><td rowspan="2">Advantages</td><td rowspan="2">Mainly for TDM users that tend to process reservation information only once during data collection.</td><td colspan="2">Information is linked to the content once before distribution, eliminating the need to manage it separately for each instance where the content appears.</td></tr>
<tr><td>It is still possible to identify a specific content even if it has been slightly modified.

Each time the expressions of reservation are modified, they are automatically applied to all the distributed items.</td><td>Since the information is physically bound to the content, it is easier to retrieve it whenever it is needed.</td></tr>
<tr><td>Limitations</td><td>Not always effective in the following situations ([290]):

• when content is shared on websites, not controlled by rights holders , such as social media;</td><td>Content matching in soft-binding is technically complex and requires similarity-based interpretation, as minor modifications can obscure the original reference.</td><td>Hard-binding becomes ineffective in the following situations ([291]):

• metadata stripping;
• content is altered even to a small extent, in case</td></tr>
</table>

([290]) What is the TDM·AI Protocol?, TDM AI (accessed 14 March 2025).

([291]) Ibid.

	• when content is properly licensed to be used by licensors who do not honour the settings specified in the original rights holders' robots.txt; • when copyright protected content is illegally republished on the internet without authorisation. Not automatically applied to each copy of the content, if it appears in **different locations**.	May fail when the content undergoes substantial modifications that alter key identifying features. Dedicated registries must be maintained to store and validate reservation data, requiring continuous oversight.	the method is based on cryptographic hashing; • content is converted into a different file format, compressed or screenshotted. These are common scenarios when sharing content online or on social media, where media files are resized, compressed, or stripped of metadata for security or business purposes. Moreover, it cannot be used with all file formats and it is technically demanding. Once content is distributed, embedded reservation data becomes immutable, meaning any updates require regenerating and redistributing the modified content.

Table 3.5.1-1: Comparison between location-based and asset-based opt-out typologies.

Some interviewed rights holders reported adopting a **hybrid approach** that integrates location and asset-based solutions, with robust licensing as the core foundation.

Additionally, they emphasised the role of **licensing and contractual compliance in enforcing location-based rights declarations** for assets hosted on partner-controlled platforms.

3.6 Rights Reservation Implementation Challenges

The previous comparison of technical opt-out measures highlights the advantages and limitations of the different measures. The variety of approaches to rights reservation reflects the complexity of the AI ecosystem and the respective needs of different content sectors that have their own unique market dynamics. There appear to be **implicit trade-offs** between different characteristics, particularly between the benefits of location and asset-based mechanisms. In *Section 3.5.1* a comparison between the two approaches is reported.

Furthermore, stakeholders cautioned that **a single opt-out solution may not be suitable across all content sectors**. Each sector whether **music, audiovisual, publishing**, or others, has distinct requirements, meaning that **a one-size-fits-all approach could prove inadequate**. Moreover, some interviewees emphasised that **the responsibility for implementing opt-outs should not fall solely on rights holders**. Even when opt-outs are provided in **machine-readable form**, they remain **voluntary**, leaving room for non-compliance by AI developers. As a result, legal frameworks (e.g., website Terms & Conditions) and robust enforcement mechanisms remain essential, with central opt-out registries seen as a useful complement rather than a complete safeguard.

A broad overview of the various mechanisms and measures can help identifying potential gaps and areas of uncertainty in the system as a whole, without reference to any specific measure or solution. These challenges are summarised below.

3.6.1 Classes of Works with Overlapping Rights

Specific issues of managing opt-out permissions may arise where a certain class of copyright protected works involves overlapping rights. This is most evident for content sectors in which related rights play an important role. A number of challenges may arise in different content sectors:

- where **performers of a work wish to opt-out but must coordinate with producers** of the works in which their performances are embodied. This may arise where the

contractual agreements between performers and producers do not fully assign the performer's right of reproduction to the producer, but enter into a specific licensing agreement for permissible reproduction.

- where **the owner of a copyright protected work that is not inherently represented in digital form wishes to opt-out but must coordinate with the right holder in a specific digital embodiment of the work**. For example, a publisher of musical compositions may authorise their works to be reproduced in several different digital sound recordings, but that publisher still explicitly holds the right of reproduction in the musical work (as distinct from the protected phonogram). In such cases, there may need to be institutional coordination between publishers and phonogram producers. This may be even more complex where musical recordings are made in third countries which use compulsory licensing systems for mechanical reproductions of musical works ([292]).

- where the author of a literary work in the journalism sector must coordinate with online press publishers who enjoy specific related rights in their online press publications. **Online press publishers** that do not contractually fully acquire the rights of their author contributors **will need to coordinate TDM opt-out statuses with such authors whose positions on reservations may vary**.

- where there are **multiple co-rights holders** (or when ownership is fragmented on a territorial basis), **and co-owners have diverging TDM reservations**.

These challenges are likely resolved over time through the evolution of the contractual practices taking into consideration AI development and TDM rights reservations as explicit elements of contractual relations in different content sectors.

([292]) BMAT (a company that indexes music usage and ownership data and provides monitoring services for music collective management organisations across the world) has made interesting public comments on this matter. BMAT has suggested that if rights are inherently attached to a composition (as opposed to a sound recording), then a rightsholder should translate their list of musical works into a list of sound recordings. While this task may require some effort on the part of rights holders, it would facilitate compliance by AI companies, particularly in relation to unilateral declarations as a rights reservation measure. Such a mapping could possibly be done by associating ISWC identifiers (unique metadata identifiers for musical works) to corresponding ISRC identifiers (unique metadata identifiers for sound recordings). See BMAT website (accessed 14 March 2025).

3.6.2 Conflicts between Opt-out Mechanisms

Conflicts between different opt-out measures may conceptually arise. Some interviewed stakeholders noted that while it is unlikely that a single standard will emerge, allowing for multiple solutions to coexist, excessive fragmentation could create significant challenges for both AI developers and rights holders.

This is most likely to happen where **information in asset-based and location-based measures conflicts**. For example, a file may contain embedded metadata which does not reserve TDM rights, but is hosted on a webpage implementing technical measures to opt-out from TDM (or vice versa). TDM users faced with conflicting expressions of reservation may need to integrate some decision rules in their compliance mechanisms. This may negatively impact on the development of the TDM licensing market, as an AI developer may exclude content from its TDM on the basis of a location-based opt-out (e.g., REP opt-out of an entire website), while the underlying content may actually be available for licensing on the basis of an asset-based mechanism.

A **hierarchy of different measures** may possibly emerge with different levels of implicit authoritativeness depending on their level of granularity and direct attribution to the actual rights holders.

3.6.3 Delegation of Reservations to Platforms

Specific issues may arise when **User-Generated Content (UGC) platforms** make licensing or opt-out decisions that override individual rights holders' preferences, particularly on platforms like social media networks. Since their terms of use typically require users to grant them non-exclusive licences to reproduce and make their content available, it remains uncertain whether platforms can engage in blanket licensing or opt-outs while users retain exclusive reproduction rights. Resolving this may require updates to platform terms, potentially sparking legal disputes over consent and retroactive enforcement.

An important development in the market is thus the intersection between access restrictions and opt-in monetisation opportunities on UGC platforms. As the control of TDM permissions

on UGC platforms by users posting content on these platforms is gaining attention, it may become a new element of inter-platform competition.

For example, YouTube announced in December 2024, that it would be introducing the ability of platform users to opt-in to 'third party training', where uploaded content is used for AI training by pre-identified commercial AI partner companies. Google stresses that to be eligible for AI training, *"a video must be allowed by the creator as well as the applicable rights holders. This could include owners of content detected by Content ID."* This development does not affect the existing YouTube terms of service, which prohibits web scraping ([293]).

Another example and approach is DeviantArt, a social network for artists, that in 2022 began developing Generative AI models and introduced a new meta directive ([294]), aimed at allowing users to reserve their works from being used in model training. This measure was implemented in response to concerns from artists regarding the unauthorised use of their works in Generative AI datasets. The system operates by embedding a specific "noai" meta-tags directly into the HTML of a webpage:

- To apply the reservation to the **entire page**: <meta name="robots" content="**noai**">
- To apply the reservation **only to visual content** on the page: <meta name="robots" content="**noimageai**">

While these directives have gained attention from various online platforms, they are **not officially recognised** in Google's documentation on supported robots meta-tags ([295]). Similarly, OpenAI's crawler documentation ([296]) and Microsoft Bing's guidelines ([297]) do not list them as recognised instructions.

([293]) Third Party AI Trainability on YouTube, YouTube Help, 16 December 2024 (accessed 14 March 2025).

([294]) What's DeviantArt's New "Noai" Meta Tag and How to Install It, Foundation Web Design & Development (blog), 12 November 2022, (accessed 14 March 2025).

([295]) Robots Meta Tags Specifications, Google Search Central (accessed 14 March 2025).

([296]) OpenAI website (accessed 14 February 2025).

([297]) Robots Meta Tags, Bing (accessed 14 February 2025).

3.6.4 Expiration of Protection Terms

A fundamental premise of any TDM reservation is that it is based on a valid exclusive right, which means that the work (or other subject matter) must still be within its term of protection. Relative to the fast-changing AI technologies landscape, the term of protection for copyright is relatively long – running for the **life of an author plus seventy years** after their death, or for seventy years after an anonymous or pseudonymous work is lawfully made public ([298]). The term of protection for related rights is generally **fifty years** ([299]). The term of protection for the *sui generis* database rights is **fifteen years** after making or publication. However, since many databases are periodically modified and updated as the result of new investments, such databases may enjoy new terms of protection as new protected subject matter ([300]). The online rights of press publishers are substantially shorter, lasting for **two years** after publication ([301]).

The issue of the term of protection is important for developing TDM reservation protocols which support copyright law principles. Once the term of protection expires, a work (or other subject matter) enters into the public domain, and any associated TDM reservation ceases to be valid. A robust ecosystem should ideally thus allow TDM users (i) to verify the validity of the right including that the term of protection is still running, and/or (ii) to rely on a mechanism within the measure itself for automatic cancellation once the term of protection is over.

Given the relatively long term of protection for copyright and related rights, this issue might not arise often on the assumption that the majority of content being used is under a valid term of protection. However, problems of unintentional invalid TDM reservations i.e., those made during a valid term but then remaining as declarations after expiration, may arise in rare instances.

([298]) Directive 2006/116/EC Article (1).

([299]) Ibid. Article (3).

([300]) Ibid. Article 10.

([301]) Directive (EU) 2019/790 Article 15(4). Furthermore, the CDSM Directive's definition of 'press publication' is not restricted to online publication, but rather a collection which "is published in any media" [CDSM Article 2(4)]. It may thus be understood that the term of the press publishers; right begins from the date of first publication, irrespective of whether that first publication is online/digital or offline/physical. This is critical to understand the term of protection for the press publishers' right, and hence the time frame during which a TDM reservation would be valid.

3.7 Comparison Criteria for Non-Reservation Solutions

In addition to reservation-based measures, there exists a range of advanced technical solutions designed to protect copyrighted works by preventing their use in GenAI training. As their reservation-based counterparts, they belong to the class "**X1A**" (see *Section 2.5*). These solutions address some of the rights holders concerns about potential violations of reservation declarations.

The following criteria are considered to compare such solutions. The list closely mirrors the one used to analyse reservation measures (presented in *Section 3.3*) to support further comparisons between the two distinct approaches.

- **Typology** - is this measure **(a) location-based** or **(b) asset-based**?

  As with reservation measures, a solution can protect an item based on (a) its location (i.e., where it is stored) or on (b) a unique asset identifier, which ensures that each copy of the work is protected, regardless of the hosting platform or even if it exists as an offline copy.

- **TDM Users Specificity** - does the measure allow applying different restrictions based on the specific GenAI system?

  Sometimes this differentiation allows the solution to be more effective and to avoid situations where the application of the solution negatively affects the intended use of the impacted item.

- **Use-Differentiation** - can the measure selectively deny some type of uses of the data?

  Use differentiation is relevant because many stakeholders have suggested that their TDM reservations differ based on use cases ranging from: (i) prohibition of all TDM, (ii) prohibition of all TDM for AI training, (iii) prohibition specifically of commercial GenAI training, (iv) prohibition for model real-time inference and retrieval.

- ***Granularity*** - does the measure apply to individual works or a larger set of content based on practical organisation?

  Granularity is relevant because it impacts the level of control achievable on the affected data. Some rights holders may prefer to apply a different level of protection on different subsets of the data.

- ***Versatility*** - is this measure specific for some **type of content**, or can it be used for all file-types and digital assets?

  As for the reservation-based measures, versatility is relevant because it may affect the widespread adoption by leveraging the network effect and ensuring applicability across various sectors.

- ***Robustness*** - Is the measure **resilient against modification/removal** (intentionally by bad actors, or unintentionally through distribution processes)?

  Since these solutions primarily depend on complex technologies, this aspect is essential for evaluating the effectiveness of the proposed solution.

- ***Openness*** - Is the solution **proprietary** or **openly available** for use?

  As for the reservation-based measures, openness is relevant because this may affect both the extent to which the measure eventually becomes widely adopted and the potential costs to rights holders. However, when evaluating non-reservation solutions, the costs could become more **significant due to the increased technical complexity** involved.

- ***Ease of Implementation*** - What level of efforts and cost are required **by rights holders** to use the measure.

Ease of Implementation is relevant because rights holders vary in their technical abilities and non-reservation measures can be far more complex than their reservation counterpart.

- ***Flexibility*** - Does the solution enable the rights holder to easily make **adjustments after its initial implementation**?

  Flexibility is relevant as it allows rights holders to adapt to changing market preferences and enhances the solution's overall usability by enabling the correction of potential errors.

- ***External Effect*** - Does the measure create any **unintended effects** (external to the issue of TDM management) which might affect rights holders, TDM users, or third parties, either positively or negatively?

  External effects are important to consider, as the implementation of a measure can have unintended consequences on the content it protects. For instance, it might reduce the quality of an image or affect the usability of a protected website.

- ***Market Maturity*** - To what extent is this measure **already used**, and has proven to be **scalable**?

  As with the reservation measures, the market maturity has to be considered to evaluate the viability of the solution and the likelihood of it reaching widespread adoption.

3.8 Non-Reservation Solutions

The following subsections provide examples of measures which may fall within the category '**X1A measures'** (measures used by rights holders; see *Section 2.5*) but are not framed as mechanisms intended to meet the legal criteria for right reservations. These measures are relevant as potential alternatives to explicit reservation-based solutions, but also because some interviewed rights holders have suggested that reservation and non-reservation approaches can be combined to achieve optimal results. The use of these non-reservation solutions also reflects the perception of rights holders that existing reservation measures are insufficient to control access to their content, and subsequent use in GenAI model training.

According to some stakeholders interviewed, technical measures for enforcing protection are more advanced in industries like music and film, compared to news publishers.

3.8.1 Protective Perturbations

"Protective Perturbations" are transformations applied to data to obscure its distinctive features, making them undetectable during data mining practices (including machine learning).

One example is the technology proposed by **DataDust.ai ([302])**, which protects text content from AI-powered scrapers by using a text font that AI cannot interpret.

Moreover, a series of research endeavours have been directed toward addressing image privacy and copyright issues raised by Stable Diffusion's TDM, one of the most used diffusion models (Zhao et al., 2024).

One approach involves the addition of **imperceptible** protective adversarial perturbations to images, **preventing Stable Diffusion from learning** the features – distinctive visual elements or details, such as shapes, colours, or textures – of protected images. Here are some examples (Zhao et al., 2024):

([302]) DataDust website (accessed 18 March 2025).

- **Glaze** focuses on preventing artists' work from being used for specific style mimicry in Stable Diffusion. It optimises the distance between the original image and the target image at the **feature level**, causing Stable Diffusion to learn the wrong artistic style;
- **Anti-DreamBooth** incorporates the **DreamBooth fine-tuning process of Stable Diffusion** into its consideration. It uses bi-level min-max optimisation, where the inner step **simulates DreamBooth fine-tuning** to maximise the model's ability to learn a subject, and the outer step creates subtle tweaks to the **images that minimise this learning**. This makes the images resistant to fine-tuning while preserving their visual quality;
- **AdvMB (Adversarial Masked Blending)** works by applying targeted perturbations to **specific regions** of an image using a masking strategy. This ensures the protected areas are prioritised, maximising disruption to AI models while keeping the image visually intact.

These efforts have showcased clear results in safeguarding image data from being exploited by Stable Diffusion. After fine-tuning on images with adversarial perturbations, images generated by Stable Diffusion tend to exhibit lower quality and semantic deviations compared to results obtained from fine-tuning on clean images.

While these methods can prevent Stable Diffusion (and hypothetically other models) from deriving the benefits of training on protected images (and even negatively impact the model), it is crucial to consider their effectiveness in long term real-world scenarios: if these methods fail to adapt to various real-world usage contexts, they might give users a false sense of security. A study conducted by Zhao et al. (2024) demonstrated that **natural transformations**, such as compression and image blur, **can decrease the effectiveness of perturbation techniques** like AdvDM and Anti-DreamBooth. While these transformations may decrease the quality and the resolution of original images and their added value in AI training, such image-preprocessing methods can still be used by AI developers to **bypass the protection with acceptable costs**. In the same study, the Expectation over Transformation (EoT) ([303]) algorithm was applied to test whether it could enhance the robustness of perturbation techniques like AdvDM and Anti-DreamBooth. Despite EoT's potential to generate

([303]) The Expectation over Transformation (EoT) algorithm enhances the effectiveness of perturbation methods by ensuring the added adversarial noise remains robust under various real-world conditions. When used with algorithms that apply perturbations to images to protect them from Generative AI (GenAI) training, EoT repeatedly transforms the images (e.g., through resizing, rotation, or adding noise) and optimizes the perturbations across these variations.

transformation-tolerant adversarial examples, the algorithm failed to yield significant improvements when the applied transformations were of **moderate strength**.

3.8.1.1 Glaze: Images Protection

Glaze (Shan et al., 2023a) allows artists to add **perturbations** to their images which would **prevent diffusion model-based generators from being used to mimic their styles**. The researchers from the University of Chicago that created Glaze collaborated with 1000 artists, going to town halls and creating surveys to understand their concerns. While building Glaze, Shawn Shan et al. measured their success by how much the tool was addressing the artists' concerns (Jiang et al., 2023).

3.8.1.2 NightShade: Training Data Poisoning

In 2024, a study (Shan et al., 2024) explored the unexpected vulnerability of state-of-the-art text-to-image generative models, such as Stable Diffusion, to a novel type of data poisoning attack ([304]). These attacks were so far known to be effective only if at least 20% of the training dataset was poisoned. The study established that despite being trained on massive datasets, these models are surprisingly susceptible to what the researchers term "prompt-specific poisoning attacks".

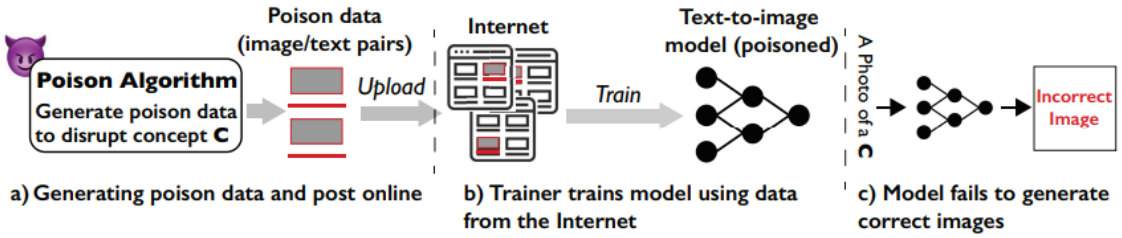


Figure 3.8.1-1: Diagram outlining the working principle of NightShade's poisoning attack (Shan et al., 2024).

([304]) A **Poison Attack** in AI refers to a malicious strategy where adversaries introduce manipulated or harmful data into a model's training process or exploit its inference phase, aiming to degrade performance, insert biases, or cause the model to behave unexpectedly.

The study identifies a key factor behind this vulnerability: "**concept sparsity**". The researchers observed that, although these models are trained on vast collections of data, the number of training samples tied to specific concepts or keywords is relatively small, leaving these concepts exposed to **targeted manipulation**.

The researchers introduced "NightShade," a poisoning attack **optimised** for this vulnerability. Unlike conventional attacks requiring extensive modifications to training data, NightShade achieves its goals with **minimal poisoned samples**, sometimes fewer than 100. These samples are **crafted to look identical to benign images**, through the introduction of **small perturbations** to evade detection.

Nightshade's effects can alter the output for specific prompts, such as making "dog" prompts generate "cats," while also affecting related prompts through a phenomenon known as "**bleed-through**". As shown in *Figure 3.8.1-2*, corrupting a concept like "dog" may also disrupt associated ideas like "puppy" or "wolf," spreading damage across the model's understanding.

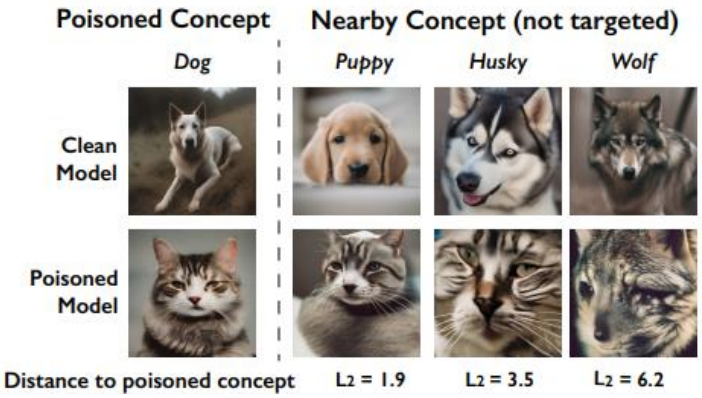


Figure 3.8.1-2: The 'bleed-through' effect of Nightshade compromises the model's generations when they are related to the poisoned concept (Shan et al., 2024).

The study highlights NightShade's potential to **destabilise a generative model entirely** when the number of poisoned concepts rises, as shown in *Figure 3.8.1-3*.

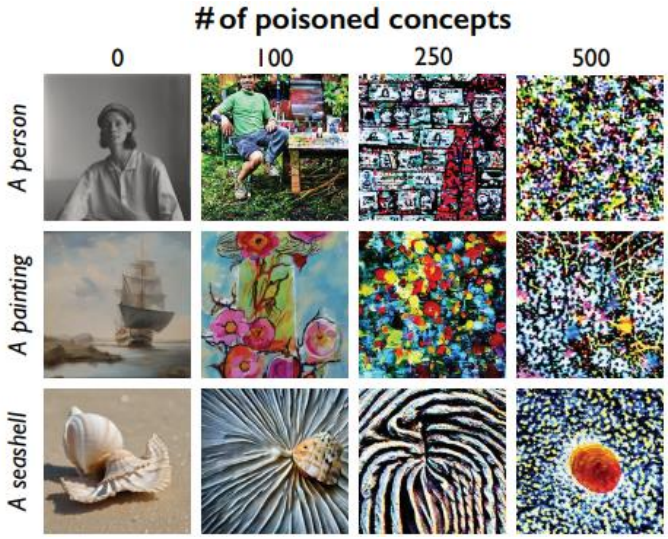


Figure 3.8.1-3: Stable Diffusion XL's outputs against a varying number of poisoned concepts (Shan et al., 2024).

Interestingly, the researchers propose an unconventional use for NightShade: as a defensive tool for content creators seeking to protect their intellectual property. By **poisoning their publicly available images**, artists and organisations could effectively deter unauthorised model training, destabilising any systems that incorporate their work without permission. Since, unlearning techniques are currently not effective enough to compensate for the poisoning, the developers of the generative model may have to restore a version of the model from before the attack. If the developers introduced mechanisms to detect and filter out NightShade's poisoned images, it would enforce opt-out as intended by the rights holders from the beginning.

3.8.2 Crawler Blockers

Industry studies and internet traffic measurements estimate that a significant percentage of traffic, 30% to 50% depending on the data provider, is due to bots. Bot traffic is impacting revenues by increasing IT costs. In addition, most of the bot traffic may be considered malicious, for example aiming at exploiting security weaknesses or performing abusive

scraping. Cybersecurity providers offer services to act against malicious bots. Such services also exist to block AI crawlers ([305]).

3.8.2.1 Cloudflare's solutions to blocking AI Crawlers

The exponential growth of AI applications has significantly increased the challenge of unauthorised data scraping for content hosts. AI developers often deploy proprietary crawlers or adapt existing web crawlers to extract data for training and inference purposes. In addition to increasing the resources (and related cost) needed by the content host to make its content available, this creates difficulties in distinguishing legitimate web traffic, such as search engine indexing, from unauthorised data collection by AI-focused bots.

Cloudflare is offering solutions to address this challenge through its **Bot Management Suite** and **AI Detection Tools.**

Cloudflare's **Bot Management Suite** provides technical solutions for detecting and managing unauthorised web crawlers in general, including those used for AI data scraping. The core features of the solution include:

- **Real-Time Monitoring and Classification**: Cloudflare analyses web traffic to identify and block unauthorised bots, particularly AI-focused crawlers attempting large-scale data ingestion. Detection relies on behavioural indicators like abnormal bounce rates, session durations, and IP-based analysis, as well as indicative of AI-specific crawlers.
- **Granular Access Controls**: Website administrators can configure policies to permit, block, or redirect bot traffic based on their requirements, offering flexibility in enforcement. **Preventing Access**: Based on the policies set, the solution provides infrastructure level enforcement to block bots that attempt to access content without authorisation, ensuring that proprietary data cannot be used for AI training processes.

([305]) See Cloudflare Radar Bot Traffic (accessed 31 March 2025), Akamai's 2024 SOTI V10 Issue 10 report "Scraping Away Your Bottom Line: How Web Scrapers Impact Ecommerce" and Imperva's 2024 Bad Bot Report.

A **real-time dashboard** provides website administrators with insights into bot behaviour, enabling them to fine-tune access policies and respond to emerging threats effectively.

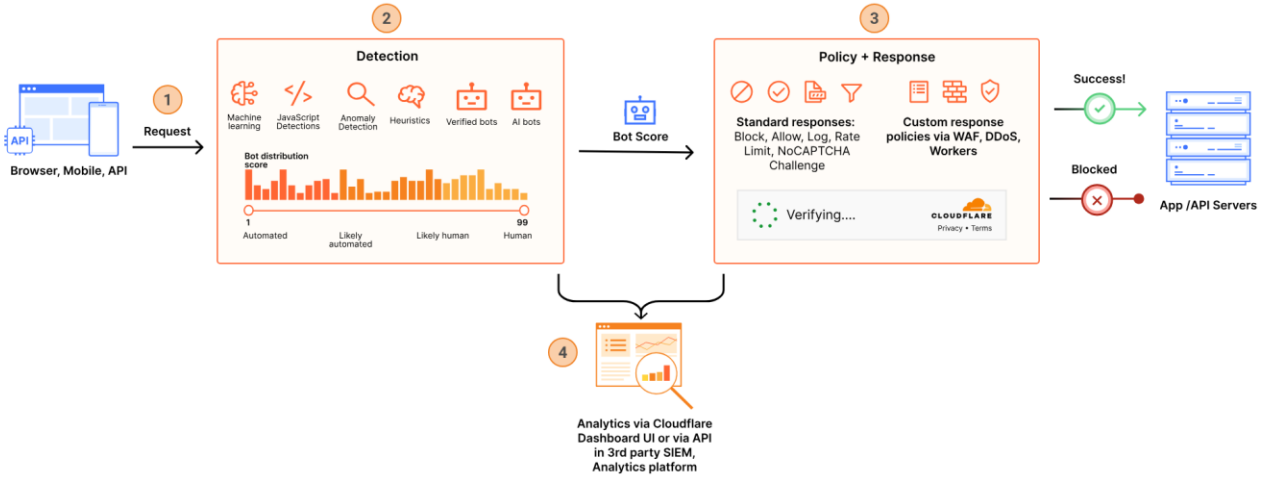


Figure 3.8.2-1: Bot Management Suite's workflow ([306]).

Cloudflare's **AI Audit and Detection Tools** provide the same solution but specifically tailored to give website owners greater control over how AI crawlers interact with their content.

Some of the most common use cases are **media organisations**, which leverage these tools to prevent unauthorised incorporation of news content into AI training datasets, as well as **educational institutions**, which also leverage these tools to safeguard proprietary research datasets, maintaining their integrity and exclusivity in academic research context.

As previously discussed, there is an increasing sophistication of advanced crawlers that mimic human browsing patterns. These crawlers, often designed for real-time inference or data collection for generative AI, can bypass traditional detection mechanisms by appearing as legitimate users. Furthermore, the use of **shared IP ranges or adaptive behaviours** complicates their identification. On the other hand, there is concerns of triggering **false positives**, where legitimate users are inadvertently blocked or restricted due to misclassification by automated systems. This could negatively impact user experience and harm relationships with genuine website visitors.

([306]) Bot Management, Cloudflare Docs (accessed 14 March 20205).

To combat these concurrent challenges, Cloudflare advocates for a **collaborative approach** between technology providers, content creators, and regulators, emphasising the importance of transparency, fairness, and proportionality in implementing crawler blocking measures. Cloudflare notably aims at providing accessible user interfaces and resources to smaller organisations with limited technical expertise, as well as enhancing false-positive management systems, such as incorporating feedback loops, to further optimise detection algorithms ([307]).

3.8.2.2 Akamai's Bot Manager and Content Protector

Akamai Technologies, Inc., is a leading global provider of content delivery network (CDN) services, cybersecurity solutions, and cloud services. The company has developed two solutions to address its customers' need to protect against the massive bot requests, up to 70% of the overall traffic, on their sites ([308]).

Akamai states that it is leveraging its **widespread network** to continuously gather up-to-date intelligence on bot trends and technologies. This enables near real-time updates to its detection system, allowing it to **deploy mitigations as soon as new bot activity is identified**, through its **Bot Manager** and **Content Protector** solutions[309]).

- **Bot Manager** provides website owners with a bot traffic management solution. It leverages deep learning models trained on the basis of the 37 billion bot requests it processes on a daily basis through its network, including data from bot attacks targeting large enterprises across multiple industries. The system employs **AI-driven analysis** to assess incoming traffic and make a distinction between human and bot traffic, assigning a **bot-likelihood score** based on detected patterns and anomalies. To assign this score, the tool notably provides the client some **invisible-to-humans**

([307]) Cloudflare Helps Content Creators Regain Control of their Content from AI Bots, Cloudflare, 23 September 2024 (accessed 14 March 2025).

([308]) Bot Manager, Akamai (accessed 13 January 2025).

([309]) Ibid.

challenges, such as storing cookies ([310]) and executing JavaScript ([311]), and performs **user behaviour analysis, browser fingerprinting, HTTP anomaly detection**, and **high request rate detection**. To minimise false negatives caused by detection evasion tactics, such as bots mimicking browsers, the tool includes a module specifically designed to **detect browser impersonation** ([312]). It then supports the setting of automatic actions when the computed bot-likelihood score overcomes a defined threshold, such as **blocking**, **serving alternate content, serving challenges, slowing**, as well as real-time and historical **reporting** and the possibility to **compare bot traffic statistics** across Akamai customers ([313]).

- **Content Protector** is a solution that is specifically designed to target scrapers rather than bots in general. It uses detection techniques specifically tailored to the **methods and techniques used in scraper attacks**. The system evaluates traffic risk by identifying anomalies across multiple assessment levels ([314]):
    - **Protocol-level assessment**: Analyses how the client establishes a connection with the server, ensuring that the communication patterns align with those expected from common web browsers and mobile applications.
    - **Application-level assessment**: Determines whether the client can execute JavaScript-based business logic. When JavaScript is executed, Content Protector collects device and browser characteristics (fingerprint) and cross-checks them against protocol-level data.
    - **Behavioural analysis**: Monitors behavioural patterns in user interactions to identify anomalous activity indicative of scrapers.

([310]) Cookies are files that websites save on the client's device (such as phone or computer) when visited. They help websites remember things about the client itself, such as login information, or items in a shopping cart. This makes the experience smoother when returning to the site.

([311]) JavaScript is a programming language widely used to build websites and applications. When visiting a site, the web server can send to the client browser some JavaScript code to be executed to produce some results which, eventually, are visualised in the browser itself.

([312]) Bot Manager, Akamai (accessed 13 January 2025).

([313]) Ibid.

([314]) Content Protector, Akamai (accessed 13 January 2025).

- **Headless browser detection**: Detects headless browsers by analysing JavaScript execution patterns and identifying known indicators.

Based on the assessed risk levels, Content Protector enables different actions, such as **blocking, throttling, or issuing CAPTCHA challenges** to mitigate false positives.

3.8.3 Digital Fingerprinting

Fingerprinting is a computing concept that refers to mapping a large quantity of data into a unique identifier using various algorithmic processes. It can be applied at the level of a specific digital asset (file) to identify different copies of a digital file, such as a digital copy of a copyright-protected work. This file-level differentiation can be used in rights management to determine the potential source of a leaked or pirated file.

In the context of rights management and GenAI, fingerprinting can be utilised for both input and output identification. Unique identifiers allow for looking up a digital file and mapping it to some external information about the work, such as rights management data. Additionally, opt-out reservation notifications can, in principle, be embedded directly in a fingerprinting system as a form of metadata (see *Section 3.4.2.6*). Such applications would necessitate fingerprinting analysis on a file-by-file basis during the GenAI training process.

Section 4.3.3.2.1 provides the example of Google's Content-ID system, which is an application of fingerprinting measures for the purpose of rights management (though not opt-out specifically). For more details on using fingerprinting to identify GenAI output, please refer to *Section 4.3.3*. A comparison of the differences between watermarking and fingerprinting is discussed in *Section 4.3.3.3.*

3.9 Comparison between Non-Reservation Solutions

Table 3.9-1 presents the evaluation of the solutions listed in *Section 3.8* based on the criteria defined in *Section 3.7*. Each column represents a distinct solution, while each row is dedicated to a specific criterion. By providing a brief description in each cell, this table is designed to encapsulate the key details elaborated upon in *Section 3.8*.

As for the technical reservation mechanisms, the **location-based or asset-based** approaches (reported under the field 'Typology') have their respective advantages and limitations, already discussed in *Section 3.5.1*.

As described, these solutions provide limited differentiation based on potential uses (see the '**Use-differentiation**' field), primarily due to their non-reservation-based nature and their indirect relation with the regulatory provisions on TDM.

Furthermore, the **data poisoning tools** are limited with regard to 'TDM User-specificity', 'Versatility' and 'Granularity', as they have been specifically designed to be **effective only under certain conditions**. For instance, they rely on algorithms that exploit the unique characteristics of visual content.

Conversely, the technologies presented below demonstrate **greater robustness** compared to Reservation Solutions (assessed in *Section 3.4*), as their technical implementation is primarily focused on **enforcing** protection. Even in the case of data poisoning techniques, where media are not explicitly blocked from being ingested into the training process, the model is effectively prevented from learning and subsequently reproducing the distinctive features of the images.

While the technologies for data poisoning are **open-sourced**, the tools for blocking AI crawlers are developed and made available as services by major companies that play a central role in the internet infrastructure and provide a wide range of services beyond crawler management. Their market reach and service offering contributes to the popularity of the tools under evaluation, which is considered when assessing their level of '**Market Maturity**'.

	Glaze	Nightshade	Cloudflare's Bot Management Suite	Akamai's Bot Manager & Content Protector
Typology	Asset-based.	Asset-based.	Location-based.	Location-based.
TDM User Specificity	Works only against Stable **Diffusion-based AI systems**.	No.	Yes.	Yes.
Use-differentiation	Designed to prevent AI models from replicating an artist's distinctive style in generated outputs.	No.	No. Cloudflare's tools differentiate between web crawlers based on their behaviour and identity but do not allow content-specific differentiation based on the intended purpose of the scraped data.	Some support is provided in distinguishing scrapers from crawlers with other purposes (e.g., indexing). However, there is no option to customise bot management rules based on the purpose of scraping.

	Glaze	Nightshade	Cloudflare's Bot Management Suite	Akamai's Bot Manager & Content Protector
Granularity	Asset-level. Can theoretically be applied even to a single portion of an image.	Asset-level.	Different rules can be applied to **each HTTP request** that meets specific criteria. These criteria may include website URLs or more granular aspects of data exchange during client-server interactions.	Offers customisable rules for managing each data exchange during client-server interaction.
Versatility	Can protect only visual content.	Can protect only visual content.	As explained in the previous point on 'Granularity', the protection **can be applied to each piece of data exchanged during client-server interaction**, thus including all data formats.	Solution for managing general bot requests, featuring customisable rules. Additionally, it provides a specialised tool designed to protect content from scrapers.

	Glaze	Nightshade	Cloudflare's Bot Management Suite	Akamai's Bot Manager & Content Protector
Robustness	Researchers achieved a **93-96% success rate**, as evaluated through both artists' feedback and a CLIP-based metric. However, if **unprotected copies of the same copyrighted content exist**, they may weaken Glaze's effectiveness by allowing AI models to be trained on unaltered versions.	Tests using a CLIP classifier and human inspection from 185 participants showed an **80% success rate when accounting for the poisoned model's architecture and nearly 70% when treating it as a black box**. This highlights high transferability across models. Poisoning typically becomes effective after training on at least 50 samples of the same concept.	Cloudflare uses heuristics and machine learning to identify bots, even when they do not self-identify. While effective, the system still encounters **false positives** (legitimate users blocked) and **false negatives** (bots bypassing detection). To improve accuracy, Cloudflare provides a **'Bot Feedback Loop'**, allowing customers to report misclassifications and enhance its detection models.	Akamai continuously updates its machine learning models using data from its global network to detect bot attacks. The system is designed to stay **up to date with emerging threats**. Additionally, **CAPTCHA challenges** can be configured to reduce **false positive bot detection rate** when suspicious activity is detected.
Openness	From March 2023, Glaze has been distributed **for free** for both Windows and MacOS.	The tool is freely available.	Among its subscription options, Cloudflare also provides **free plans**.	Akamai offers **customised pricing plans**.

	Glaze	Nightshade	Cloudflare's Bot Management Suite	Akamai's Bot Manager & Content Protector
Ease of implementation	Artists have to learn how to use the Glaze software tool and need to **adjust the level of perturbation** they want to add to their images. The time required for protecting a single piece of art **should not exceed ten minutes**.	Installing Nightshade may require some technical expertise. However, it provides a graphical user interface with a form for customising poisoning parameters. A step-by-step guide is available on the official site.	Using the tools requires a base technical expertise.	Using the tools requires a base technical expertise.
Flexibility	Once applied, protection is difficult to modify retroactively, as digital content diffusion is hard to control. This applies equally to protected and unprotected images.	Once applied, poisoning effects are difficult to reverse, as the diffusion of digital content is hard to control. This applies equally to poisoned and unpoisoned images.	Reconfiguration is always possible.	Reconfiguration is always possible.

	Glaze	Nightshade	Cloudflare's Bot Management Suite	Akamai's Bot Manager & Content Protector
External Effects	Images may exhibit slight visual alterations, but these are typically minimal and customisable to balance protection with perceptual similarity.	Poisoned images may exhibit slight alterations, but these are typically minor and customisable, ensuring effectiveness while maintaining visual consistency.	Given the challenges in bot identification, **functionalities beyond data gathering for AI, such as indexing, could also be affected**. Additionally, **bandwidth consumption** may increase due to the tool implementing JavaScript challenges to verify the identity of the interacting client.	Given the challenges in bot identification, **functionalities beyond data gathering for AI, such as indexing, could also be affected**. Additionally, **bandwidth consumption** may increase due to the tool implementing JavaScript challenges to verify the identity of the interacting client.

	Glaze	Nightshade	Cloudflare's Bot Management Suite	Akamai's Bot Manager & Content Protector
Market Maturity	From its release to June 2023, Glaze had surpassed **740,000 downloads**, attracting the attention of both rights holders and government organisations (Shan et al., 2023b).	Nightshade was released in 2024 and has attracted the attention of **dozens of press publishers** ([315]).	Cloudflare has a **wide variety of customers worldwide**. They range from small businesses to large enterprises. There are more than 500,000 users of the AI audit tools.	Akamai declares to serve **more than 50% of the Global 500 companies**. Across its worldwide network, its tools allegedly intercept an average of 37 billion bot requests daily.

Table 3.9-1: Comparison between Non-Reservation based Solutions.

([315]) Publications and Media Coverage, Nightshade (accessed 15 February 2025).

3.10 Evolving Input Management practices and solutions

To ensure legal access to training data, TDM users and GenAI model developers use a combination of technical solutions that align with "X1B" of the compliance taxonomy. The measures are particularly critical in navigating the complex landscape of digital rights, and they help prevent the unlawful use of copyright-protected materials in training datasets under EU copyright law. In that respect, a number of TDM users have developed new practices or are reported to consider new technical tools.

Interviewees belonging to the stakeholder category of *Solution Providers* acknowledged that the growing use of AI applications will present **increasingly complex challenges in managing opt-out protocols**. These include **aligning copyright policies with diverse local laws across the world** and to exploring additional machine-readable opt-out solutions by consultations of AI developers, academics, rights holders and policymakers. It was further stated by others *Solution Providers* that multiple solutions will need to be developed as one-size-fits-all solutions are not feasible at this stage, and standardised solutions may be not appropriate yet.

From the point of view of developers, it is much easier to deal with location-based opt-out solutions, since AI developers do not always have access to detailed information about the copyright status, licensing, or ownership of the content available on the web. Moreover, there is often no single source of truth for copyright information. In conclusion, some *Solution Providers* emphasised the need to balance rights holders' needs with the practical limitations faced by AI developers.

3.10.1 Google Extended

In September 2023 Google introduced **Google-Extended**, a control that can be used in the context of the **Robots Exclusion Protocol (REP)** and the associated **robots.txt** file which allows website owners to block its AI chatbot Gemini and its AI development platform Vertex from scraping their content.

Google-Extended implements an **opt-out** mechanism nested within **Robots.txt**, which has been chosen because it's a standard well-known by rights holders. The service is available to both rights holders' organisations and individuals.

This platform allows rights holders to opt-out from AI training while **continuing the indexing of the content in Google's search engine**. Moreover, Google-Extended does not stop sites from being accessed and used in **Google's AI Overviews summaries.** To avoid this, rights holders would have to opt-out of being scraped also for search indexing purposes ([316]). During interviews it emerged that the AI Overview summaries are only a search functionality, unrelated to Generative AI. However, some rights holders indicated those as **particularly harmful,** as they often replace direct user visits to their websites, potentially impacting traffic, engagement, and revenue streams.

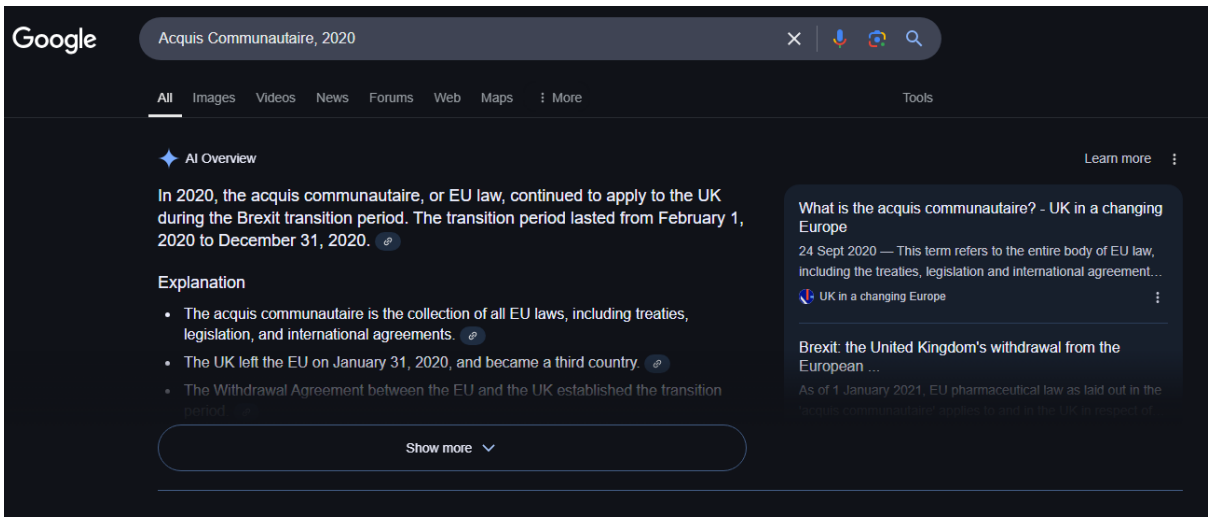


Figure 3.10.1-1: An example of how Google's AI Overviews appear in the internet browser.

([316]) News organisations are forced to accept Google AI crawlers, says FT policy chief, Press Gazette (blog), 6 November 2024 (accessed 14 March 2025).

3.10.2 Fairly Trained Certification

Fairly Trained ([317]) is a non-profit organisation offering a certification service to verify that Generative AI training has been conducted exclusively using copyright-safe approaches, augmenting the prestige of the companies owning the certification.

In particular, the training data in use must fall into one of the following categories ([318]):

- Be explicitly provided to the model developer for the purposes of being used as training data, according to a **contractual agreement** with a party that has the rights required to enter such an agreement;
- Be available under an **open license appropriate to the use-case**;
- Be in the **public domain globally**;
- Be **fully owned** by the model developer;
- Any **third-party models, open models, or synthetic data utilised in the product**, service, or models undergoing certification must **meet the same standards**. Specifically, any model used to build the certified model must also hold certification, and any synthetic data used for training must be generated by a certified GenAI system.

The certification is **reevaluated annually** for a feeand the applicant must prove to have **robust processes** for:

- Conducting **due diligence into the data** considered for being used for training purposes;
- **Keeping records of the training data** that was used for each model training.

The processes outlined above become increasingly complex when handling large volumes of data, even if the data is copyright-safe, underlining the importance of **scalability** as a key challenge for AI companies to manage.

([317]) Fairly Trained Launches Certification for Generative AI Models That Respect Creators' Rights, Fairly Trained, 17 January 2025 (accessed 21 November 2024).

([318]) Licensed Model Certification, Fairly Trained (accessed 17 January 2025).

Fairly Trained hosts a website page ([319]) which lists all the GenAI products and companies that have obtained its certification. As of January 2025, the list includes 19 entries, primarily featuring companies in the music sector, along with a few examples of certified LLMs and text-to-image generators. The prevalence of music companies reflects the involvement of key music industry players in endorsing this project, as well as the existence of two layers of rights (publishers' copyright and record labels' phonogram rights) in this sector.

3.10.3 OpenAI Media Manager

To consider rights holders reservations OpenAI employs a dual approach: a robots.txt directive serves as the primary opt-out mechanism for its GPTBot (addressing text-based content), while a separate, dedicated process governs DALL-E outputs.

OpenAI is also reported to be working on Media Manager – a tool designed to let creators and content owners declare their **ownership** of works and decide **how their content should be included or excluded in machine learning research and training** ([320]). This initiative should reportedly involve pioneering machine learning research to create a unique tool capable of identifying copyrighted text, images, audio, and video from various sources and aligning usage with creator reservations. OpenAI is also reportedly partnering with creators, content owners, and regulators to develop Media Manager, aiming to launch it by 2025 ([321]). , However specific details on whether Media Manager will allow opt-out based on content location or other criteria have not been disclosed yet.

([319]) Certified Models, Fairly Trained (accessed 17 January 2025).

([320]) Our approach to data and AI, Open AI (accessed 7 November 2024).

([321]) Ibid.

3.11 Institutional Support by IP Offices

The analysis in this Chapter of the issues surrounding GenAI inputs, and the various rights reservation and content access measures highlights potential opportunities for support by public institutions. These institutions might include national IP offices, Community-level institutions such as the EUIPO, and other national or supranational competent authorities that could contribute to the development of a robust and fair AI ecosystem supporting both the development to GenAI and the valuation of copyright and related rights. This section briefly discusses some of these opportunities.

3.11.1 Technical Support

Some interviewed stakeholders highlighted that given the complexity of EU Copyright Law and the operation of TDM exceptions, smaller AI companies may require support through **technical facilitations and ready-made solutions**.

Moreover, there is a broad consensus among interviewed **stakeholders** in favour of standardised opt-out solutions, rather than company-specific protocols and tags. For this purpose, an 'impartial authoritative organisation' could list appropriate standards to increase clarity and consistency in the administration of opt-out mechanisms. This could contribute to overcome the current situation where rights holders are hesitant to invest in the implementation of technical measures that may not be acknowledged by TDM users, while TDM users encounter new technical measures without clarity on which ones to implement.

Many interviewed stakeholders (belonging to the category of *Solutions Providers* in particular) highlighted the crucial role that IP offices could play in managing **identity and ownership verification functions for opt-out declarations**. Given their access to official records and established legal regulatory/policy jurisdiction, such institutions may be well-positioned to oversee this process effectively and provide backend support for verification of rights reservations.

The issue of technical support from public institutions was also raised by several stakeholders, suggesting that one option may be for institutions at both the national and Community level to

provide a framework for supporting a federated database systems that would facilitate collaboration among diverse stakeholders.

Federated registries or databases function by allowing multiple trusted entities to contribute to and maintain portions of a broader, distributed system (see *Annex XI.6*). A **federated search API** has been proposed by some *Solution Providers* to enable rights holders to verify whether their content has been used in AI training datasets.

Some stakeholders suggested that federated databases could be used to **aggregate information from a multitude of other primary databases**. As federated databases do not need to supplant individual databases, they may represent a possible solution which can balance centralised oversight with the decentralised principles supporting a range of innovative solutions driven by rights holders and AI developer market needs. In addition federated systems can enable **shared governance and participation**, whereby each entity retains control over its own database (of opt-out information) while adhering to common standards for synchronisation and data integrity. In this context, federated databases may have different advantages for copyright management as they reduce central control, maintain consistency, and provide a flexible mechanism for stakeholders across different jurisdictions to access and update data in real-time.

This is based on the idea that a federated approach may support **scalability** and address the specific needs of the different content sectors. A public institutions oversight over such federated databases could offer rights holders both autonomy and security in managing their data. Furthermore, **the involvement of public institutions may bring trust and certainty** to the ecosystem which would benefit smaller players that tend to be more risk averse.

In the context of GenAI and copyright management, federated registries would allow publishers, artists, and other rights holders to register their works, express opt-out reservations, and verify usage through a network of interconnected databases managed by various stakeholders, such as publishers, copyright offices, and other authorities. Each participating node within the federated system maintains its own data, but the system as a whole is synchronised to ensure consistency and prevent discrepancies in the management of rights and permissions. This means that if rights holders update the opt-out status of their works, it would be reflected across the entire federated system, enabling AI developers to verify compliance without relying on a single centralised database.

The potential for federated registries may extend beyond simply recording opt-out reservations. They offer dynamic, up-to-date rights information that is crucial for AI developers, who need to verify permissions before incorporating content into their training datasets. The federated model addresses one of the primary limitations of blockchain-based copyright management systems, namely the difficulty of making updates or corrections once data is recorded. In federated registries, updates can be performed more efficiently, ensuring that the most current rights information is always accessible.

It is however noted that federated databases may also be managed by **private entities**. The **Open Rights Data Exchange** by **Valunode** is one such example (see *Section 3.4.2.7*). Private solution provider **Liccium** (*Section 3.4.2.6*) also uses a federated approach, which provides a reference point for how a decentralised yet coordinated registry system might operate.

The role of the public institution in managing a federated database may include **establishing protocols**, **ensuring that registries are properly interconnected**, and overseeing that the data maintained by each node is **accurate and up to date**.

However, a significant technical challenge may be in ensuring **real-time synchronisation** across different nodes while **preserving data integrity**. This necessitates the development of secure APIs that allow for continuous updates without compromising data quality or exposing it to **unauthorised access**.

Federated registries must also accommodate the **specific needs of different content sub-sectors**, each of which may have unique requirements for data storage, access, and usage. Collaboration between private and public sector actors may thus play a vital role in standardising these processes and ensuring inter-sectoral harmonisation.

This distributed yet coordinated approach, supported by federated APIs and registries, could be one way to address some of the complexities of rights management in the rapidly evolving GenAI landscape. It may also facilitate cross-sector integration, ensuring that AI developers have the necessary tools to verify compliance while preserving the rights of content creators across multiple jurisdictions.

There are different approaches possible regarding the management of such databases. These will be fully analysed in the European Commission's Study to assess the feasibility of a central registry of Text and Data Mining opt-out expressed by rights holders ([322]).

3.11.2 Non-Technical Support

Aside from technical support, there are a number of potential roles for public sector institutions, including IP Offices. The assumption behind the potential roles listed below is that **rights holders groups are more likely to benefit from institutional support**, relative to AI commercial providers who likely have greater financial and organisational capacity to navigate copyright and AI issues that are at the core of their activities.

- *Repository of Measures* - Institutions can provide **public information on various solutions for managing rights-reservations**, including their key features and mechanisms. This could help rights holders to understand the possible varieties of solutions available, as well as their respective advantages and limitations. While the institution should not endorse any particular one (especially proprietary measures provided by private undertakings on a commercial basis), it may provide information on various solutions, point to reliable sources of information or host trainings (webinars) with major rights holders organisations (like CMOs or civil society organisations) which may give recommendations for specific content sectors.

- *REP Crawler agent identifier lists* - While many public resources exist to aggregate the user-agent identifiers for web crawlers, the public institution can serve as a platform to consolidate **lists of bot names provided directly by AI providers**, as well as **statistics about use of these bots**, and the proportion of internet domains that block them using protocols such as REP. While this information is available through other technical sources, the public institution can bring further trust and confidence in this information by contextualisation **specifically in relation to TDM rights reservations**.

([322]) The main purpose of this study is to assess both the opportunity and feasibility of developing a work-based registry of content identifiers and associated metadata that would support – whether centrally or within a federated network– the effective expression of Text and Data Mining (TDM) opt-outs for copyright-protected works and other subject matter and facilitate their identification by Artificial Intelligence (AI) developers. See Study to assess the feasibility of a central registry of Text and Data Mining opt-out expressed by rightsholders, European Commission, 22 January 2025.

- *Unilateral Declarations Repository* – A public institution may serve as a **repository of unilateral declarations** made voluntarily by large rights holders entities such as CMOs and publisher organisations. This may bring increased public visibility to such unilateral declarations.

- *Model Contractual Terms* - The public institution may serve as a **forum** for rights holders interest groups (such as CMOs and publisher organisations) **to share model contractual terms** suited for specific rights holders groups and sub-sectors, not only for TDM rights reservations, but also for ensuring that licensing, hosting, distribution, assignment, and representation agreements sufficiently address the issue of the capacity to make rights reservations (i.e., that a chosen opt-out mechanism can meet the *'by the right holder'* requirement when implemented by a party that is not the original author themselves). Such model terms can also include suggestions for terms and conditions of websites that include TDM opt-out language.

- *Licencing Reports* - The public institution may **track trends in direct-licencing** across different content sub-markets, specifically the **emerging norms and standard contractual practices,** as they inevitably emerge. As noted in the analysis of pricing dynamics in training data markets, norms and standards are still evolving regarding issues like one-time payments (as compared to ongoing royalties), and remuneration calculations based on per-token rates (as compared to a per-work basis). Making such trend reports publicly available also facilitates open participation in the development of these norms, and access to information for smaller rights holders groups which may not have access to expensive proprietary industry reports.

- *Public Education* - Aside from activities specifically aimed at assisting actual rights holders, the institution can serve an important public messaging function. This is important as end-users are also a key stakeholder in building trust in the overall AI ecosystem. Such educational outreach may focus on helping the public understand the complex interface between copyright law and AI services. Several interviewed stakeholders emphasised the need for **greater awareness among rights holders** regarding the mechanics of GenAI training and the opt-out mechanisms. A shared 'vocabulary/ontology' is needed so that the **difference between AI training and data**

mining is universally understood. On the side of AI companies, **adherence to voluntary codes, such as the GPAI Code of Practice** can help to mitigate risk-averse attitudes and encourage developers to remain open rather than shifting toward closed or proprietary approaches.

4 Generative AI Output

This chapter provides an overview of the final stages in the GenAI life cycle, including on the technical processes involved in output generation. Regarding copyright compliance and transparency issues, it investigates solutions aiming to meet legal requirements for AI-generated content, as well as strategies to prevent such content from infringing on copyright.

The following aspects are considered:

- The model's integration with **Retrieval-Augmented Generation (RAG)** technologies, which enables it to incorporate **external data beyond its training dataset**. This includes information retrieved from additional databases as well as publicly available content from websites and other resources across the internet.
- The indication of information, for example through metadata allowing for effective **provenance tracking of content**.
- The model's ability to generate content that falsely appears to the user as authentic, truthful or not generated by AI, and thus needs to be properly **marked** (via visible labelling, provenance tracking solutions or watermarking) or **detected** to be AI-generated.
- The **training data memorisation** phenomenon (discussed in *Section 3.2*), which could lead AI developers to implement **output filters** as a mitigation strategy.
- The probability that a model may **generate a copyright-protected work without having been exposed to it during the training process**, which may again lead AI developers to consider implementing output guardrails.
- The possibility of **filtering input prompts** when malicious requests are detected.
- The potential of using **machine unlearning** technologies to instruct the model to "forget" specific training data, thereby preventing it from producing related outputs in the future.

- The possibility of **editing** a model's knowledge with the purpose of correcting or updating the knowledge on which it has been trained.

4.1 Technical Analysis of Content Generation Methods and Phases

Once the training phase of a GenAI model has been completed, the content generation process goes through several stages, starting with the validation and deployment of the model, and ending with the output generation, which are described in this section.

Figure 4.1-1: Graphical overview of the main processes involved in the GenAI output development.

4.1.1 Model Validation and Deployment

Model validation and deployment are critical in the life cycle of GenAI systems, bridging the stages **between model training and real-world applications,** and testing technical reliability and compliance with ethical and legal frameworks.

4.1.1.1 Model Validation

During the validation phase, the model's developers **evaluate the model's performance** against predefined metrics. Techniques such as **cross-validation** ([323]) and **benchmarking** are employed to test the model's ability to 'generalise' across diverse datasets, often incorporating external benchmarks like GLUE for natural language models or ImageNet for computer vision tasks. **Adversarial testing** is also a key component, where the model is

([323]) Model cross-validation is a technique for evaluating a model's performance by splitting the data into multiple subsets, training the model on some subsets, and testing it on others. Training/validation/test is the subdivision commonly referred to by AI developers.

subjected to inputs designed to reveal vulnerabilities. These validation methods ensure the technical soundness of the model and play a role in mitigating risks such as extractable memorisation, which could inadvertently reproduce copyrighted material from training datasets (See *Section 3.2* for more details on memorisation).

4.1.1.2 Benchmark Datasets

Benchmark datasets are used to build a series of standardised tests that measure the **capabilities of AI models**, such as understanding and generating natural language, solving complex problems, and adapting to new tasks. AI researchers have created a plethora of benchmark tests, that provide a **framework for comparison** and which enable developers and users to quantitatively assess different AI models' performances.

The fast technological evolution causes benchmarks to quickly lose relevance. If benchmarks measure the wrong attributes or tasks, they can result in systems that perform exceptionally well in tests but falter in real-world applications.

To minimise this risk, benchmarks need to be continuously updated and reevaluated, which requires **additional data that is sufficiently diverse and does not overlap with training data** ([324]).

Some commonly used technical benchmarks include:

- **GLUE and SuperGLUE** – evaluating natural language understanding;
- **GPQA (Graduate-Level Google-Proof Q&A Benchmark)** – evaluation performances in solving complex science questions;
- **ImageNet** – object detection and image classification; and
- **SQuAD (Stanford Question Answering Dataset)** – featuring over 150,000 questions based on Wikipedia articles.

([324]) An Introduction to GenAI Benchmarks, Medium (blog), 10 April 2024 (accessed 14 March 2025).

Among these, the use of ImageNet may present potential **copyright issues**, for example, as it contains links to possibly copyright protected images. However, the Terms of Access ([325]) allow use of the database only for non-commercial research and educational purposes. Meanwhile, the licence for the distribution of Wikipedia articles (Creative Commons Attribution - ShareAlike) is compatible with the use made by SQuAD and the licence under which the dataset is distributed.

4.1.1.3 Deployment

Deploying a model means **setting it up in real-world systems** where it can be used, making sure it operates effectively and with a large number of users ([326]). This process involves deploying the GenAI model on a server infrastructure, ensuring its accessibility for user requests.

Once installed the system requires continuous maintenance and monitoring to evaluate the model's performance, while checking for unexpected changes, security breaches and system crashes.

4.1.2 Retrieval Augmented Generation (RAG)

The training of an AI model requires significant time, making it impracticable to repeat this process every time when updated training data becomes available. Considering this inherent limitation of training, coupled with an inability to provide primary sources within generated output, it is a common aspiration to base AI applications on dynamic information. An example of this includes the **use of up-to-date information at AI inference time,** i.e., during the actual generation process (see Arkko 2024).

([325]) See ImageNet website (accessed 14 March 2025).

([326]) The number of generative models' users is substantial. In February 2025, Salesforce conducted interviews to estimate the percentage of population using those technologies. The results showed that 73% of the Indian population surveyed uses GenAI, while this rate amounts to 49% for Australia, 45% for the USA and 29% for UK. See Top Generative AI Statistics for 2025, Salesforce, February 2025 (accessed 14 March 2025).

RAG techniques go beyond ingesting content in a model, they allow for indexing and later retrieval of relevant material. RAG is a GenAI technique that combines the power of LLMs with the accuracy of document retrieval mechanisms. By using external knowledge sources, RAG enables more up-to-date, factually grounded and contextually rich outputs, as opposed to relying purely on pre-trained data.

Applications of RAG include:

- 'Answer engines', providing users with concise and contextually relevant answers, as opposed to traditional search engines;
- Customer support, where chatbots provide up-to-date solutions;
- Healthcare, aiding clinical decisions and patient interactions;
- Legal research, ensuring compliance and effective legal arguments and;
- Education, offering real-time tutoring and research tools.

From a user's perspective, RAG enables **timely and personalised content generation by** allowing the inclusion of information that was not available at the time of the training. A company can also integrate its proprietary databases into the RAG system, enabling the GenAI model to generate responses informed by the organisation's specialised data.

From a technical view, RAG **as an approach is a compromise between involving the data in the training process and fetching** it from a database without applying further elaboration. Below is a detailed comparison between the three approaches – RAG, fetching and training.

4.1.2.1 RAG versus Fetching

While both RAG and fetching involve data retrieval, they differ significantly in application and complexity.

Fetching is a basic operation that retrieves raw data from a source, such as a database or an API, without applying any transformation or contextualisation. In contrast, RAG retrieves information but also processes and integrates it to generate coherent, context-aware, responses.

Furthermore, RAG relies on advanced machine learning techniques, such as indexing and language modelling, to dynamically contextualise information, whereas fetching operates through simple querying mechanisms. When data is fetched, it must be explicitly identified beforehand. In contrast, RAG systems autonomously determine which data to retrieve based on the GenAI system's input prompt.

Feature	RAG	Fetching
Processing	Retrieves & processes data before response	Only retrieves data
Use Case	Relevant examples are AI chatbots, Q&A models	API & database queries
Context Awareness	High (integrates retrieved data)	None (raw data only)
Complexity	High (needs ML models, indexing)	Low (basic querying)

Table 4.1.2-1: Differences between RAG and data fetching.

4.1.2.2 RAG versus Training

In contrast with an LLM model, which generates text on a probabilistic basis and patterns learned from training on large datasets (See *Section 3.1.6*), a **vector database as deployed in RAG** acts as a 'memory lookup' system of direct content retrieval. The following table highlights key differences between LLM model training and RAG systems:

Feature	RAG	GenAI Model Training
Purpose	Retrieving specific, contextual data to augment responses	Learning probabilistic patterns from large datasets
Data Handling	Accesses external databases for ad hoc context retrieval	Processes entire datasets to adjust model parameters
Output Basis	Augmented by directly retrieved data	Generated from learned probabilistic associations
Retention of Data	Retains external data sources for repeated use	Data is not stored post-training; only patterns remain coded into the model's weights
Legal Considerations	Extended storage of reproductions makes applications of copyright exceptions unlikely, unless TDM exceptions would apply	Use of works and databases can be covered if compliant with the conditions of the TDM exceptions

Table 4.1.2-2: Differences between data ingestion via RAG or GenAI model training.

4.1.2.3 Static RAG versus Dynamic RAG

There are several approaches and techniques for incorporating RAG into AI system deployment, which can be subdivided into **two main types of RAG systems**: **Static RAG and Dynamic RAG**.

Static RAG relies on predefined, **stable datasets stored locally** or in a fixed format, while Dynamic RAG incorporates **real-time data retrieval** from external sources, such as live links or APIs. The databases integrated into Static RAG solutions can still be modified during the system's functioning, but not in a systematic and automated way. The amount of data

incorporated into a Static RAG's database is **limited and defined** at each moment. While for Dynamic RAG, the possibility of using web crawlers and scrapers to expand data retrieval (see *Section 3.1.2.2*) make it limitless, with a **constantly varying** retrieval potential.

Dynamic RAG extends the functionality of traditional RAG systems by incorporating real-time external data sources, such as live URLs, dynamically updated databases, or APIs. **Unlike Static RAG**, which works with fixed datasets, **Dynamic RAG continuously fetches and processes external data during inference**, making it highly adaptable.

Current market trends suggest that dynamic RAG is a particularly important technique in the **evolution of 'online' search engine services**, and the emergence of '**answer engines**'. These refer to systems designed to directly provide users with concise and contextually relevant answers, as opposed to traditional search engines, which return a list of links to external content. Examples include AI-driven platforms like **Bing AI Chat** and **Google's Gemini**, which both use RAG-enabled processes to retrieve and integrate real-time data into conversational responses. Stakeholders, particularly in the press publishing sector, express concerns with RAG-enabled search engines which use and repurpose their content, amplifying the 'value gap' arguments that previously drove discussions on press publishers' rights during the legislative development of the CDSM Directive (see Recitals 54 and 55). There are some attempts from providers of 'answer engines' to address this issue ([327]).

4.1.2.4 Copyright Implications of RAG

Using external data may require licensing agreements to avoid infringing copyright, database rights, and other related rights, particularly in commercial applications. Furthermore, the cost structures associated with accessing dynamic databases or licensing external content can significantly influence the economic viability of RAG-based solutions. A key feature of the direct licensing landscape is the **growing number of licensing agreements specific to RAG applications**, as confirmed by publicly available licensing information (as seen in *Section 2.4.3.8*). Such agreements are particularly prevalent in the press publishing sector and are also observed in academic and scientific publishing.

([327]) For example, Bing AI Chat provides links and citations to its sources. Insights from stakeholder interviews indicate that, based on preliminary data, this integration enhances the value of traditional search functionalities.

There is **no clear reference to RAG as a form of TDM in the existing agreements between AI developers and rights holders.** The framing of these agreements specifically as "licenses," as opposed to "content access agreements" referred to training data, may reflect a distinction made by stakeholders between RAG and strict TDM applications (see *Section 2.4.3.9*).

RAG generally involves the representation of referenced information in the form of **vectorised embeddings stored in a database**, which are retrieved for inference. This is in contrast with standard AI training, which involves extracting 'patterns, trends and, correlations' from large datasets before encoding them into the model's parameters and weights. Thus, RAG differs from standard model training and content generation in both **how information is abstracted and represented**, and **how these vectorised representations influence the generative process**. The copyright implications of RAG might be understood in terms of the two components of RAG applications – information retrieval, and content generation.

In terms of whether reproductions of works during RAG's retrieval phase qualifies as TDM ([328]), it may depend on how the process of RAG generation is understood. Unlike AI model training, where works are reproduced to extract correlation and patterns then abstracted into model parameters and weights, RAG may involve a **more direct process of semantic information extraction** which is used to contextualise generative prompts. However, the CDSM definition does not further define the type of information that can be generated from a TDM process (*"is not limited to…"*), and a conceivable broad interpretation might include some RAG applications. This **issue** may eventually be settled through judicial interpretation, particularly as AI technologies evolve and RAG applications become more prevalent.

Counterintuitively, **static RAG application referencing locally hosted content *may* potentially trigger more copyright-restricted acts compared to dynamic RAG, which uses open internet scraping** due to a narrower space for possibly invoking copyright exceptions**.** This arises because locally hosted content often necessitates a longer **retention** of reproductions to enable ongoing reference, a requirement that may exceed the conditions of applicability of the CDSM Directive Article 4 TDM exception, as well as the (more strict) requirements for the applicability of the InfoSoc temporary reproduction exception (see

([328]) The CDSM Directive in Article 2 (2) defines TDM to mean *"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".*

Section 2.2.1.9). By contrast, scraping the open internet for context references in dynamic RAG typically retains content only temporarily, aligning more closely with potential for application of either TDM or temporary reproduction exceptions**.**

By **licensing content for AI-specific uses**, such as accessing content through dedicated APIs, rights holders can provide controlled access to their works for retrieval in a RAG context. These **APIs** could facilitate dynamic and secure access to licenced content while embedding usage restrictions and monitoring mechanisms to ensure compliance with copyright law and contractual terms.

4.1.3 Output Generation

'Output Generation' is the last stage in the GenAI process and described in this Chapter. The process of generating content differs by the specific Generative AI technology used, and may be categorised into four **main families of GenAI models** ([329]):

- **GANs (Generative Adversarial Networks)**
  In GANs, the **generator** starts with a random input, often a vector of noise, and gradually transforms it into a structured output. This transformation happens through layers that apply **learned mathematical operations to shape the noise** into something meaningful. The generator was instructed, during the training phase, by testing its generated samples against a module trained to discriminate between real and synthetic data (called a discriminator).

- **VAEs (Variational Autoencoders)**
  VAEs generate data by first **sampling a point from the latent space ([330])** coded into the model's parameters. During the machine learning phase, various features of the training data had been analysed with the aim of deriving the relative statistical distributions and coding them into the model's parameters. Thus, by randomly sampling from the latent space, the data obtained fits the statistical features of the

([329]) For a deeper insight into the resulting structure and components of the models after the training phase, see *Section 3.1.6*.

([330]) See the *Glossary* for a definition of "latent space".

training set. The sampled point is then **passed through a decoder**, which reconstructs it into a detailed, complete output. The decoder uses **learned data patterns** to transform the simple latent representation into something meaningful and realistic.

- **Diffusion Models**

  Diffusion models generate images by **starting with random noise** and refining them step by step to create a coherent output. The generation process is like **reversing a gradual corruption of data**: instead of adding noise to clean data (as during training), the model learns to remove noise from random inputs in stages. Each step slightly improves the clarity and structure of the data until the final output emerges. The number of denoising steps depends on the specific use case, but it typically ranges **from 50 to 1000 steps** to generate a single image.

- **LLMs**

  LLMs generate text **one token after another**. Given an input prompt, the model calculates the **most likely next token** based on **patterns it has learned** from large amounts of text. It adds this token to the sequence, then uses the updated sequence to predict the next token. This **iterative process** continues until the "**end token,**" a special token used internally by the model, is generated. *Figure 4.1.3-1* visualises the complete process, including the eventual additional retrieval of data from a RAG system (see *Section 4.1.2*).

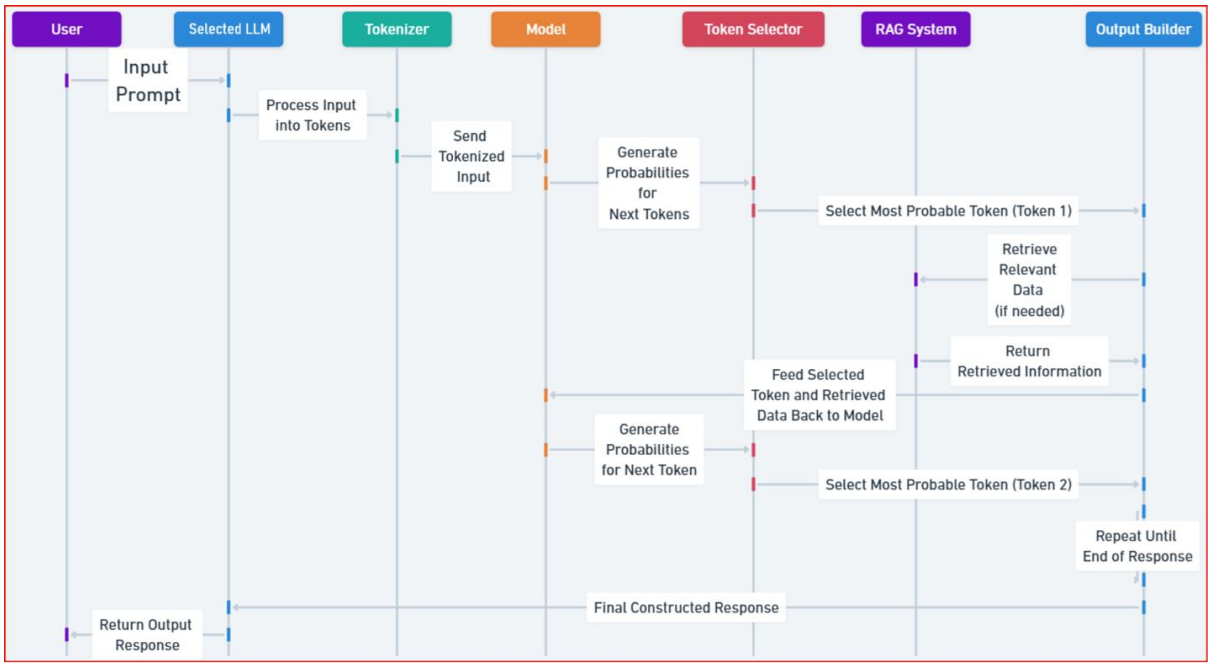


Figure 4.1.3-1: Diagram outlining the sequence of all possible interactions during GenAI model operation.

4.2 Criteria for Generative Transparency Solutions

This chapter aims to establish criteria for comparing the mapped generative transparency solutions, similar to the approach taken in *Section 3.3* for reservation measures. Based on the transparency obligations outlined in the AI Act, a set of criteria will be presented to evaluate and compare different solutions, including their advantages and limitations.

4.2.2 Legal Criteria for Transparency Measures

As discussed in *Section 2.2.3.4*, **Article 50 of the AI Act** sets out transparency obligations regarding **GenAI content**. These obligations are addressed to providers and deployers of AI systems that generate **synthetic content**, but not on the content's subsequent users ([331]).

The requirements for output transparency measures are set out in **AI Act Article 50(2),** which applies to synthetic content generally. This article states that:

> *"Providers shall ensure their technical solutions are **effective, interoperable, robust and reliable** as far as this is technically feasible, taking into account the specificities and limitations of various types of content, the costs of implementation and the generally acknowledged state of the art, as may be reflected in relevant technical standards."*

The mentioned requirements for output transparency measures – effectiveness, interoperability, robustness, and reliability – are not absolute and are to be met '*as far as this is **technically feasible**'*. Additionally, the AI Act recognises that these requirements may apply differently to different types of content, and that implementation costs need to be taken into account.

Recital 133 suggests that, for technical transparency solutions, *"Such techniques and methods can be implemented at the level of the **AI system** or at the level of the **AI model**,*

([331]) For example, once a GenAI system is used to create synthetic content, it may spread across all communication channels, both online and offline. Thus, the synthetic content may be viewed and further shared by a large number of people aside from the original end-user (i.e., the user of the deployed GenAI system who initiated the creation of the content).

including general-purpose AI models generating content, thereby facilitating fulfilment of this obligation by the downstream provider of the AI system." This is directed towards the different points on the AI value chain at which output transparency measures may be implemented.

Under Article 50(2), synthetic content is required to be *'marked in a **machine-readable** format and detectable as artificially generated or manipulated'*, while under **Article 50(4)** for deepfakes, the obligation is on deployers to *'disclose that the content has been artificially generated or manipulated'.* This points to a possible differentiation between machine-readable and human-readable transparency measures. Furthermore, the transparency obligation regarding deepfakes is at the level of the AI systems deployer, and not the AI systems provider, which places the obligation further downstream in the GenAI value chain.

Ultimately, Article 50(7) foresees the development of a Code of Practice to facilitate the effective implementation of the obligations regarding the detection and labelling of artificially generated or manipulated content. This Code is anticipated to set out best practices for output transparency measures. As this Code of Practice is relevant for providers and deployers of GenAI systems, it is distinct from the Code of Practice foreseen by Article 56 (which addresses copyright-compliance measures for the providers of general-purpose AI models).

4.2.3 Comparison Criteria for Transparency Measures

Measures for ensuring the transparency of GenAI output, following from the schema in *Section 2.5,* are measures under category "**X2**". Taking a holistic view, considering the criteria defined in *Section 3.3* for comparing training input opt-out measures and the legal analysis of **AI Act Article 50 and Recital 133**, the following criteria have been chosen for evaluating and comparing various output transparency measures:

- **Typology** – type of technical measure with reference to the examples in Recital 133;
- **Versatility** – ability to apply to different types of content (both in terms of content sub-sectors and file types);
- **Openness** – existence of any proprietary rights and licensing terms over the measure's enabling technologies;

- **Market maturity** – extent to which the measure has already been deployed in the market, or to which it has demonstrated proof of concept;
- **Human-readability** – ability of the measure to be easily understood by natural persons, to convey information about the content's nature;
- **Cost implication** – cost of deploying the measure at the level of AI systems (both financial cost and compute requirements);
- **Robustness** – ability of the measure to consistently apply across subsequent content distribution channels and the life cycle of a digital asset, including both intentional and unintentional manipulations or other unexpected situations;
- **Interoperability** – availability of public specifications or API for enabling the technology to be integrated with others;
- **Scalability** – the measure's capability of managing an increasing number of assets or users; and
- **Reliability** – the solution's capability to manage the transparency of GenAI output in a comprehensive and trustworthy manner over time.

To recall, AI Act Article 50(2) requires that *"Providers shall ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible, taking into account the specificities and limitations of various types of content, the costs of implementation and the generally acknowledged state of the art, as may be reflected in relevant technical standards."* However, neither Article 50(2) nor its supporting Recital 133 explicitly define the meaning of the terms 'effective', 'interoperable', 'robust', and 'reliable'.

Nevertheless, a possible understanding of these requirements can be derived from international standards, without prejudice to how these concepts may be ultimately interpreted and applied in the Code of Practice on Article 50. In particular, the International Organisation for Standardisation (ISO) has developed standardised definitions of AI concepts and terminology in **ISO/IEC 22989:2022** ([332]).

([332]) This ISO Standard defines **'robustness'** as *"ability of a system to maintain its level of performance under any circumstances"* (3.5.12), and **'reliability**' as the *"property of consistent intended behaviour and results"* (3.5.9; incorporated from standard ISO/IEC 27000:2018). These two definitions are consistent with the definitions in standard ISO/IEC TS 5723:2022 (on 'trustworthiness' vocabulary). ISO-IEC TR 24029-1:2021 further specifies that "*robustness properties demonstrate the degree to which the system performs with atypical data as opposed to the data expected in typical operations*". See ISO, ISO/IEC 22989:2022' (accessed 14 March 2025).

These AI-specific standards do not contain definitions for **'interoperable'**, though this term is referenced in various non-AI contents throughout the broader framework of EU digital law and data regulation ([333]). The Computer Programs Directive defines 'interoperability' as *"the ability to exchange information and mutually to use the information which has been exchanged"* ([334]).

Given the multi-stakeholder nature of the AI value chain (see *Section 2.5*), the obligations of AI Act Article 50 and the context of AI Act Recital 133, it is useful to expand on the concept of 'interoperability'. Measures may exhibit '**horizontal interoperability**' when they can be used by different stakeholders (such as different AI model providers or different systems deployers) at the same point of the value chain. Measures may also exhibit '**vertical interoperability**' where they can be applied by stakeholders at different points on the value chain. Vertical interoperability is important in the context of AI Act Recital 133, which suggests that measures implemented at the upstream GPAI model or system levels may be enough for a downstream system provider to fulfil its transparency obligations.

As for the requirement of **'effectiveness'**, understanding its meaning necessitates evaluating the outcomes produced by implementing a measure considering the objectives of Article 50. Applying this requirement to the measures analysed in this report is beyond the objectives of this study as this may involve an assessment of concrete solutions of the regulatory objectives.

The concepts of 'effectiveness', 'interoperability', 'robustness', and 'reliability' are broadly reflected in several of the above listed comparative criteria. In particular, these comparative criteria adopt a concept of 'interoperability' which focuses on 'horizontal interoperability', while 'vertical interoperability' conceptually overlaps with the criteria of 'robustness' and 'versatility'. Furthermore, aspects of the concept of 'effectiveness' are captured in various criteria including 'scalability' and 'reliability'. These criteria are to be understood in broad terms and should not be interpreted as evaluations of specific measures in terms of the legal requirements of Article 50.

([333]) For example, in the *'Interoperable Europe Act'* (Regulation (EU) 2024/903), the *'Digital Markets Act'* (Regulation (EU) 2022/1925.

([334]) See Recital 10, DIRECTIVE 2009/24/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 23 April 2009 on the legal protection of computer programs.

4.3 Generative Transparency Solutions

Solutions offering copyright information and related use conditions for content like music, books, or scientific works have been an important tool for facilitating the legal use and distribution of creative works. The rise of GenAI and the AI Act's obligations highlight the need to assess and refine these solutions to meet new requirements for transparency and compliance. Different approaches, which can be combined, have been developed to enhance the transparency of GenAI systems. Those belong to the class **X2** described in *Section 2.5* and can be divided in:

- **Provenance Tracking**: This approach seeks to certify the entire lifecycle of a digital asset, encompassing its creation and subsequent modifications. By clearly delineating the steps that may involve copyright protection and licensing, provenance tracking ensures a reliable record of the asset's history. This history is often encoded in a machine-readable format into the **content's metadata**. Some examples of solutions following this approach are C2PA and JPEG Trust.

- **Generated Content Detection:** This technique plays a critical role in promoting transparency towards consumers and safeguarding them against deception by identifying fraudulent or manipulated content, including so-called 'deepfakes' ([335]). This problem partially overlaps with copyright and related rights when artists who create performing art are cloned and their performances are recreated by using GenAI systems. The study provides a brief overview of the Generated Content Detection landscape, citing NVIDIA StyleGAN3 detector as one of the possible solutions.

- **Content-Processing Solutions**: These methods, including **watermarking and fingerprinting**, directly include information in or analyse the digital content itself. Fingerprinting is used to detect (unauthorised) copies by identifying unique patterns within the content, while watermarking embeds provenance information into the content itself to help prevent or detect unauthorised use. Both methods can be used to **detect copyrighted data to be filtered out when performing the input data**

([335]) A "deepfake" is defined in Article 3(60) of the AI Act as an "*AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful*".

collection for the GenAI system training. They may also be used on the output side for marking GenAI output (watermarking) or detecting if data contains copyrighted works (fingerprinting). Watermarking can be subject to a series of attacks aiming at removing the embedded information (see *Section 4.3.3.1.2*).

- **Membership inference attacks:** can be employed to determine whether a model has been trained on a specific data point, even if the model's training set has not been openly disclosed and thus, it is not possible to directly search for the presence of a specific work inside it.

4.3.2 Provenance Tracking

One of the central issues of GenAI is provenance tracking. Due to the rise of AI generated **deepfakes** provenance tracking and deepfake detection have become **closely related issues** (although deepfakes can and are still generated without using GenAI techniques).

Here we provide a detailed analysis of C2PA, a well-established protocol for content provenance, as well as of the pilot projects of JPEG Trust and TRACE4EU.

4.3.2.1 C2PA

C2PA (Coalition for Content Provenance and Authenticity) is a Joint Development Foundation project ([336]) to address the prevalence of misleading information online through the development of media standards for **certifying** the **provenance** of media content.

The specifications aim to support the global, voluntary adoption of digital provenance methods by fostering the development of a robust ecosystem of provenance-enabled applications tailored to diverse individuals and organisations. These specifications are designed to uphold security, privacy, and human rights standards ([337]).

([336]) It was formed through an alliance between Adobe, Arm, Intel, Microsoft and Truepic.

([337]) See C2PA Specifications, C2PA (accessed 28 November 2024).

It is crucial to note that C2PA specifications do not assess the truthfulness of provenance data. Instead, they focus on verifying whether the provenance information is properly linked to the associated asset, accurately structured, and untampered ([338]).

The protocol is compatible with nearly twenty file formats across **all media types**, such as jpeg, mp3, pdf and mp4. Some use-case examples of this technology are: (i) helping consumers check the provenance of the media; (ii) enhancing clarity around journalistic work; (iii) assisting intelligence; (iv) enhancing the evidentiary value of critical footage; and (v) enforcing disclaimer laws on edited images.

By embedding machine-readable assertions into media files, C2PA also enables the inclusion of Training and Data Mining Assertions (see *Section 3.4.2.3*).

4.3.2.1.1 C2PA: Technical functionality

C2PA provides **unique credentials to each author of provenance data** to bind statements of provenance data to instances of content. With the same credentials, the author can perform **late edits**, and the protocol ensures that the current versions of the asset and the provenance data are **up-to-date** and **cryptographically bounded**.

Trust decisions are made by the consumer of the asset based on the identity of the actor(s) who signed the provenance data, and the information contained in the data itself.

To enable customers to make informed decisions, it is crucial to prevent **impersonation attacks** targeting the authors of provenance data. This requires ensuring that the user digitally signing the C2PA manifest is genuinely the individual or entity they claim to be. For this reason, C2PA leverages **Certification Authorities´ (CAs)** real-word due diligence to ensure **digital credentials are only issued to identified actors**. Once the identity has been certified and associated with credentials, it can no longer be altered.

([338]) Ibid.

Certification Authorities and Their Role in Digital Provenance

CAs are fundamental to the establishment of trust within digital ecosystems, acting as **trusted third parties** that issue **digital certificates** to validate the identity of entities involved in digital interactions.

From a technical perspective, CA-issued digital certificates consist of an entity's public key, identification data, and the CAs' digital signature. This signature ensures that any attempt to alter the certificate is detectable.

The authentication process involves the use of CAs' public keys, embedded into applications such as browsers, to validate certificates in real-time. This allows software to automatically recognise whether a certificate is genuine, thereby facilitating seamless trust without direct user intervention.

In practice, however, complete verification of a CAs' authenticity is often not implemented to its full extent. Many software programmes depend on a **pre-determined list of "trusted" CAs**, which is embedded by the **software vendor**. As a result, end-users are ultimately **placing their trust in the software vendor** as much as in the CAs, which presents a potential vulnerability if the vendor's list is compromised or outdated.

Those lists of "trusted" CAs often include the most well-known CAs worldwide: there are few of such **root CAs** and they emit certificates for the other CAs, building an **hierarchy** of trustiness.

To mitigate these risks, **regular audits of CAs**, coupled with enhanced verification protocols, are imperative. Additionally, **decentralised approaches**, such as blockchain-based identity verification, offer promising alternatives to address the vulnerabilities inherent in centralised traditional CAs models.

The protocol also manages **nested assets**, i.e., content created using other works: those sources are referred as the "ingredients" and are signalled in the derived work's provenance data, which will also include the ingredients' provenance data.

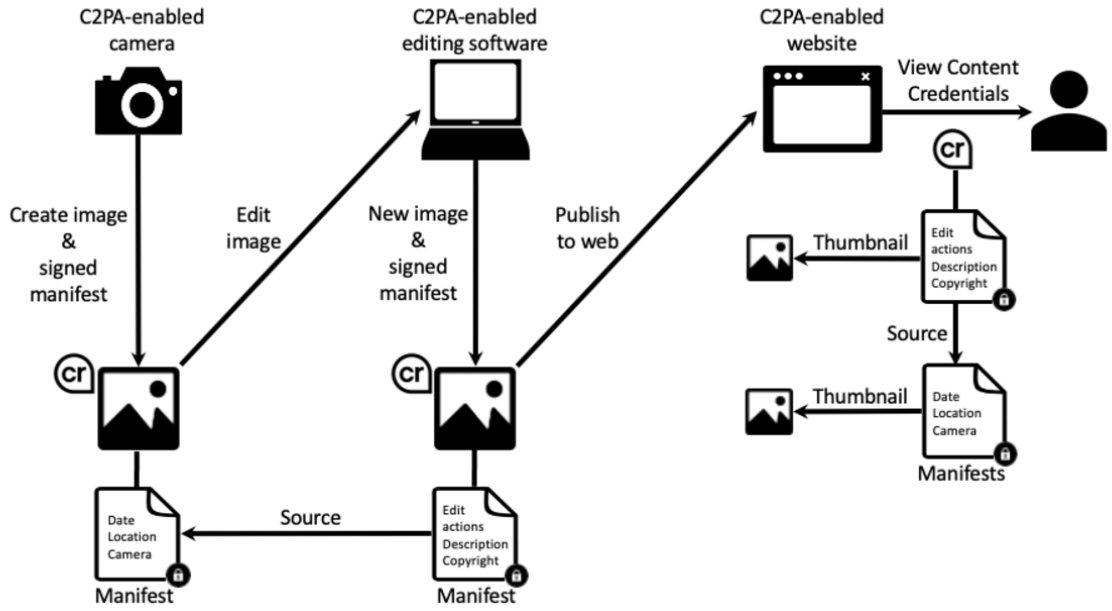


Figure 4.3.3:-1: Diagram representing a possible use case for C2PA, which includes certifying the content's whole history([339]).

For instance, a **C2PA-enabled camera or smartphone** can **automatically embed a C2PA manifest into captured photos**. This manifest is customisable and may include details such as the creator's name and the photo's copyright information. If the original image is edited using **C2PA-enabled software**, the application generates a new **manifest documenting** the date of the modifications and a summary of the changes made. When the final image is published online, a **C2PA-compatible website** or display device will recognise the signed manifest and display a **Content Credentials icon on the image**. Viewers can click the icon to access a complete history of the asset. For display devices that do not support C2PA, users can upload the image to a Content Credentials website to view the associated C2PA metadata.

If a malicious actor tampers with the asset, the altered version will no longer align with the data recorded in the manifest, signalling a red flag. Similarly, any unauthorised changes to the metadata will be clearly detectable ([340]).

([339]) Fighting Deepfakes With Content Credentials and C2PA, CMSWire.com, 13 March 2024 (accessed 1 December 2024).

([340]) Ibid.

More technical information on the C2PA protocol is provided in *Annex XIII*.

Given the increasing consumer sensitivity to AI-generated content, the official guide of the standard ([341]) suggests including detailed information in the C2PA manifest beyond just the basic **claim and data hash ([342]).** In particular, the following attributes are suggested:

- Links to the AI-ML model's Content Credential.
- Provenance details of the model's inputs.
- Information on the components and security of the environment where the model ran.
- Timestamps of the process.
- Explainability metadata to clarify why the model generated its results.

The official guide adds that, for generative AI outputs, the prompt used for input should also be included, as well as training and data mining assertion to clarify rights associated with the output. This level of detail builds greater trust in AI-ML outputs.

The standard also defines the syntax of **Training and Data Mining assertions**. By including such an assertion in the C2PA manifest, it can be used to *'indicate that the asset should not be used for either training or data mining purposes. The assertion is flexible and allows the author of the asset to specify whether each type of process – data mining, general AI training, or training specific to generative AI – is permitted, or not.'* ([343])

4.3.2.1.2 C2PA: Advantages and limitations

The main advantage of C2PA is its high level of interoperability, which is mainly based on the respect of the data formats it defines. All the specifications are openly available, enabling compliance, and the Content Authenticity Initiative open-sourced C2PA's APIs ([344]) allows for integration with other applications. C2PA is also already integrated with other protocols, such

([341]) See C2PA Specifications, C2PA (accessed 28 November 2024).

([342]) For the definition of 'data hash', see the *Glossary*

([343]) See C2PA Specifications, C2PA (accessed 28 November 2024).

([344]) Content Authenticity Initiative, GitHub (accessed 18 February 2025).

as **ISCC** (see Section 3.4.2.6), **TDMRep** (*Section 3.4.2.2*), **JPEG Trust** (*Section 4.3.1.3*) and is supported by many devices.

As all asset-based solutions that are leveraging hard-binding (see *Section 3.5.1*), C2PA may be vulnerable to **content metadata tampering**, including both modification and removal. C2PA manifests are **protected against modifications** due to the cryptographic techniques described above. A **removal** of the content metadata, including the C2PA manifest embedded within it ([345]), would lead to a complete loss of the provenance data associated with the content. This removal can be executed through various editing tools that allow the modification or stripping of embedded metadata. One potential solution is for platforms to **flag assets that lack manifests**. Developers are also collecting, through the contributions to the project's GitHub repository, a list of soft binding algorithms that may be used to retrieve a stripped C2PA manifest.

Another important aspect to consider is that C2PA certification **verifies the author of the provenance data but does not guarantee the authenticity of the content** itself. Even if the author is certified, they could still produce and sign false or manipulated content.

The dependence on **centralised CAs**, while a necessary part of the trust framework, introduces a point of vulnerability. Any compromise at the CAs level can result in the injection of falsified provenance data, thereby undermining the system's overall integrity. Malicious content producers or distributors can obtain manipulated certificates in various ways, such as by creating their own CAs for this purpose. Therefore, content consumers must verify the **authoritativeness of the CAs referenced in the C2PA manifest**, rather than relying solely on the presence of a CAs' signature.

C2PA also faces challenges in managing nested assets. Derived works often incorporate multiple source assets, termed as "ingredients," each of which must be tracked and verified through its entire lifecycle. **Maintaining the integrity of these nested components is complex** and requires the "hard-binding" of the components through cryptographic methods.

4.3.2.1.3 C2PA: Market Maturity (2024)

([345]) Ibid.

The adoption of **C2PA in cameras is expanding**, driven by the need for content authenticity in journalism and the fight against misinformation. Leading camera manufacturers like Sony, Canon, Nikon, Fujifilm, and Leica have introduced or announced plans to support C2PA ([346]).

The diffusion of the C2PA standard within **smartphones** remains limited compared to cameras, as no major smartphone manufacturer has announced full integration of C2PA technology yet.

The adoption of the C2PA standard in **image editing software** is gradually progressing, with several major programs already supporting the specification. Notably, **Adobe Photoshop** and **Lightroom** have integrated C2PA-compatible tools that allow for embedding digital signatures and metadata ([347]).

Some GenAI tools, such as OpenAI's DALL-E (text-to-image generation), are now **automatically** adding C2PA manifests to their output to provide the context of the generation ([348]). Major technology companies, publishers and manufacturers are also supporting C2PA, including **Google** ([349]), **Microsoft** ([350]), **Sony** ([351]), **Adobe** ([352]), **The New York Times** ([353]) and the **BBC** ([354]). For example, the Google Search integration with C2PA includes the functionality 'about this image', which presents to the user provenance information in a human-readable format.

([346]) C2PA Camera Support, C2PA (accessed 14 March 2025); Sony Completes Field Test for In-Camera Image Authentication Tech, New Atlas, 22 November 2023 (accessed 14 March 2025); Nikon Will Add C2PA Content Credentials to the Z6 III by Next Year, PetaPixel, 14 October 2024 (accessed 14 March 2025); Fujifilm to Bring C2PA Content Authenticity to X and GFX Cameras, PetaPixel, 16 May 2024 (accessed 14 March 2025).

([347]) Where does the photo come from? C2PA metadata as a key to content provenance, Digital Asset Management & Bildverwaltung (blog), 14 November 2024 (accessed 14 March 2025).

([348]) C2PA in DALL·E 3, OpenAI Help Center (accessed 1 December 2024).

([349]) How We're Increasing Transparency for Gen AI Content with the C2PA, Google, 17 September 2024 (accessed 1 December 2024).

([350]) Project Origin, Microsoft Research (blog) (accessed 1 December 2024).

([351]) Sony Delivers Highly Anticipated Firmware Updates Including C2PA Compliancy and Ensuring Authenticity of Images, Sony Europe, 28 March 2024 (accessed 1 December 2024).

([352]) C2PA Achieves Major Milestone with Google to Increase Trust and Transparency Online, Adobe Blog (accessed 1 December 2024).

([353]) Using Secure Sourcing to Combat Misinformation, New York Times, 5 May 2021 (accessed 1 December 2024).

([354]) Mark the good stuff: Content provenance and the fight against disinformation, 5 March 2025 (accessed 14 January 2025).

There exists the possibility to **manually** create a C2PA manifest using open-source tools. For example, 'c2patool' ([355]) is a **command-line tool**, which requires a gradual learning curve. Conversely, 'c2pa-rs SDK' ([356]) is a **software library** suitable for developers. Both these open-source packages **can be integrated into graphical software programs** or online platforms to provide a more user-friendly access to their functionalities.

4.3.2.2 IPTC Photo Metadata Standard

The IPTC Photo Metadata Standard is a standard for describing photos and enjoys widespread adoption across various sectors. Such as photo agencies. [...]The IPTC Core and IPTC Extension reportedly provides the possibility to describe the content of an image and supports the inclusion of information such as creation dates, creator names, and identifiers, as well as a flexible system for expressing rights information ([357]).

IPTC Photo Metadata Standard and C2PA are highly interoperable and share numerous similarities, including their dual functionality as mechanisms for binding reservation expressions to digital files and as provenance tracking solutions that enhance transparency in media. However, while the first was designed for providing ([358]), the second emphasises assertions and provenance (Mo et al., 2023). Moreover, IPTC Photo Metadata Standard's applicability is limited **to visual content only**.

4.3.2.3 JPEG Trust

JPEG is a joint working group between ISO and the International Electrotechnical Commission (IEC). It creates and maintains several standards for digital images. One of them is JPEG Trust, which defines a framework for establishing trust in media. This framework addresses aspects of **authenticity, provenance and integrity** through secure and reliable **annotation**

([355]) Contentauth/C2patool, Github website (accessed 1 December 2024).

([356]) mikecvet/C2PA Rust SDK Simple Walkthrough, GitHub blog (accessed 5 March 2025)

([357]) IPTC Photo Metadata Standard, IPTC (accessed 17 March 2025).

([358]) C2PA now supports both IPTC photo and video metadata, IPTC, 6 November 2022 (accessed 17 March 2025).

of media assets **along their life cycle**. The two key pillars guiding JPEG Trust's development are interoperability and trustworthiness. It is expected that the standard will evolve over time and be extended with additional specifications ([359]).

Currently JPEG Trust consists of three parts:

- The **Core Foundation** (ISO/IEC 21617-1), which officially became an **international ISO standard** in January 2025 ([360]) and supports **data provenance**.
- A second version of this standard, which is scheduled to be released in 2026, also includes **rights declaration.**
- JPEG Trust Profile Catalogue and JPEG Trust Media Asset Watermarking, which are currently under development and not covered by this report.

The standard is **media-agnostic**: although originated in the 'JPEG' ecosystem, implementers can embed trust manifests in audio, text, or other file types. This standard is aligned with **C2PA v1.4**. As a result, existing media assets that have C2PA-compliant provenance information are fully compatible with the JPEG Trust framework.

4.3.2.3.1 JPEG Trust Core Foundation: Data Provenance

This foundation handles three main areas: annotating **provenance information**, extracting and evaluating trust indicators, and handling privacy and security concerns.

Figure 4.3.1-2 illustrates the steps proposed to **assess the originality and trustworthiness** of a media asset.

([359]) JPEG Trust website (accessed 4 December 2024).

([360]) ISO/IEC 21617-1:2025, ISO (accessed 6 February 2025).

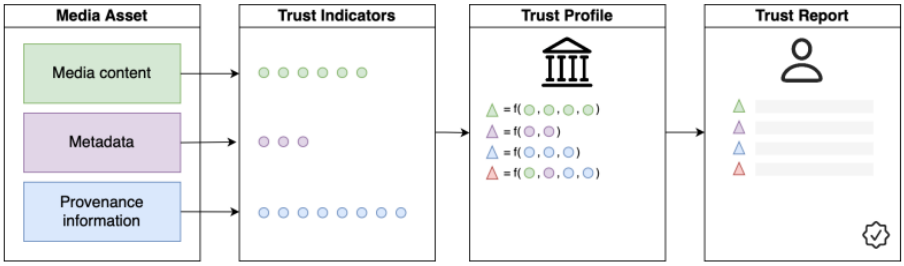


Figure 4.3.1-2: High-level schema of the procedure proposed by JPEG Trust for assessing the trustworthiness of a media asset (Caldwell et al., 2024).

To enable traceability, JPEG Trust includes in the media metadata a dedicated **Trust Manifest** containing a **list of actions** performed on the media, each reporting the **timestamp** along with information about what took place on the asset and what **software or hardware component** performed the action. Each entry of the Trust Manifest is called an assertion (or Trust Records) and reflects the syntax of C2PA. The number of Trust Records within a Trust Manifest is theoretically infinite. However, in practice, its scalability is constrained by the capacity to manage these records robustly at the JPEG Trust's implementation level, particularly given the fragmented nature of rights declarations across different sectors.

When a content is consumed, its trustworthiness is evaluated by analysing and cross-referencing the information embedded within the media to extract **Trust Indicators**. These indicators are computed taking into account the metadata, the media content itself and the provenance information embedded into the Trust Manifest. The reliability of this procedure is enhanced through the implementation of cryptographic techniques such as secure hashing **([361])** functions and digital signatures **([362])**.

Examples of Trust Indicators include the following: (Caldwell et al., 2024)

- From the media content: **Results from an external AI-Generated Content (AIGC) detector**, such as the probability that the asset was generated by AI based on a specific algorithm.

([361]) For the definition of 'data hash' see the *Glossary*.

([362]) For the definition of 'digital signatures' see the *Glossary*.

- From the Trust Manifest: **An assertion regarding the device used to capture the image**, including details about the camera that recorded and digitally signed the asset.

The resulting evaluation can be expressed in a **Trust Report** to make the information easily accessible and understood by the end user. An example of the report's bare content can be found in *Figure 4.3.1-3*. The standard does not specify the details of how the Report is displayed to the user, as these are implementation aspects determined by the software developed by companies adhering to JPEG Trust.

The JPEG Trust standard provides **reference guidelines to implementers**, but the standardisation body itself does not deliver an official **software service**. These guidelines allow software implementations based both on graphical user interfaces, which would favour the human-readability of the information, and APIs, which would allow managing a large quantity of content in a scalable manner. JPEG Trust itself does not mandate a single central repository.

```
metadata:
  name: Experimental Generative AI Profile
  issuer: JPEG Trust Committee
  date: 2023-10-31T20:23:30.668Z
  version: 1.1.0
  language: en  # this report has been generated in English

---
# Section 1

# section info, English strings extracted, except for description
-
  title: Example section title
  report_text: This is an example of a section description.

-
  # 'jpt.profile_compliance' is a reserved ID that can (optionally) be
used to signal the result of an overall binary profile compliance test
  id: jpt.profile_compliance
  report_text: This media asset is compliant with this profile.
  value: True

-
  # Need for multilingual reports probably exists, maybe we should
optionally allow the following (in addition/alternative to the references
as suggested above):
  id: content
  report_text:
    en: This content has not been modified
    es: Translation in Spanish
  value: True

-
  # Is Generative AI?
  id: aigc
  report_text:
      en: This media asset was produced by generative AI
      de: Translation in German
      zh: Translation in Simplified Chinese
  value: True

-
  # Any modifications?
  id: declaration_only
  report_text:
      en: No modifications took place after it was created
  value: True
```

Figure 4.3.1-3: An example of Trust Report's possible content. In this case, it identifies an AI-generated content ([363]).

4.3.2.3.2 JPEG Trust Core Foundation v2: Rights Declaration

The JPEG Trust committee is developing a second version of the standard, scheduled for release in 2026. This new version will expand the support for media tokenisation ([364]) such as declaration of **authorship, ownership and terms of use**. These include the **terms and**

([363]) ISO/IEC 21617-1:2025, ISO (accessed 6 February 2025).

([364]) For the definition of 'Media Tokenisation', see the *Glossary*.

conditions related to text and data mining, as this version is intended to provide an explicit means to embed opt-out declarations into digital assets. To achieve this purpose, the company built its work on the Dublin Core™ (ISO Standard for formatting to describe digital and physical resources' metadata), the Open Digital Rights Language (See *Annex XI.4*), and on **C2PA v2.1** (*Section 4.3.1.1*).

In the syntax designed to embed **TDM reservations** in the content metadata, JPEG Trust refers to some of the **categories for media usage** already defined in C2PA Training and Data Mining Assertion structure presented in the *Section 3.4.2.3*.

In particular, when the '**constrained**' value is selected for a TDM permission, the protocol foresees the use of a field named '**constraint_info**,' which can be used to freely write explanation texts or URLs. For example, this field could reference an ODRL object that encodes, in a machine-readable format, the conditions for legally accessing the associated asset. These conditions can also be specified for identified actors (see *Annex XI.4* for information on how ODRL can be used to code this information).

4.3.2.4 TRACE4EU

The Trace4EU ([365]) project addresses the vital need for the **traceability** of data, documents, and physical goods, which is essential across numerous sectors. This project, rooted in the **European Blockchain Services Infrastructure (EBSI)**, aims to develop an "**umbrella architecture**" leveraging existing EBSI services. The architecture will serve as the foundation for creating and implementing traceability application scenarios. By engaging pan-European stakeholders, the project also seeks to promote recommendations for further advancing the EBSI ecosystem. One of the project goals is to identify existing EBSI services and develop additional transparency services upon them ([366]). This initiative promises to transform traditional industries, making them more efficient, competitive, and resilient, while enhancing transparency for European citizens by enabling better tracking of commodities and data flows.

([365]) Trace4EU website (accessed 5 December 2024).

([366]) Project, Trace4EU blog (accessed 11 February 2025).

The **TRACE4EU consortium** comprises of **30 partners from 14 countries**, including research institutions, small and medium-sized enterprises, and government bodies. Through continuous collaboration with the European Blockchain Partnership (EBP), the EBSI core team, and various pilot projects (including the Open Rights Data Exchange described in *Section 3.4.2.7*), the consortium facilitates large-scale implementation throughout the European ecosystem.

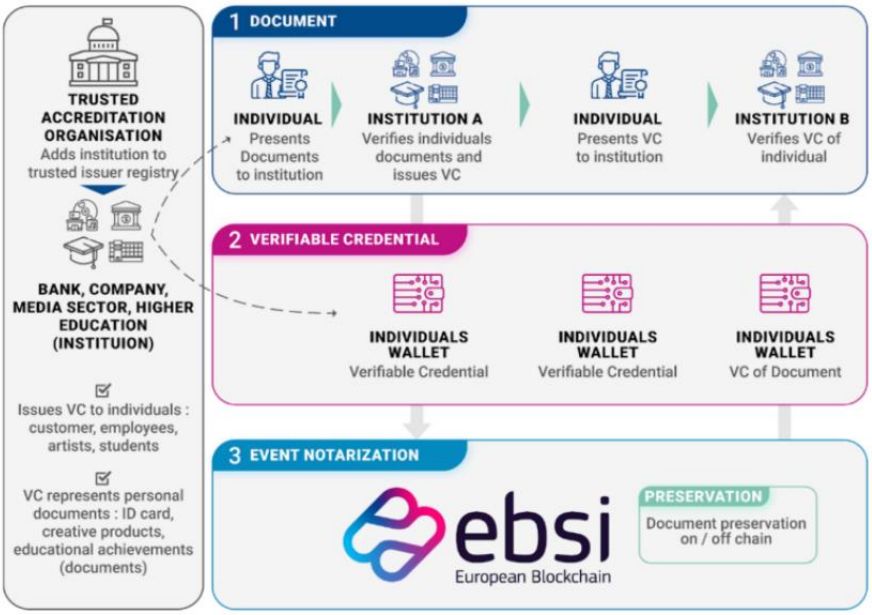


Figure 4.3.1-4: Framework of the core concept underlying the TRACE4EU Project.

4.3.2 AI-Generated Content Detection

Different techniques exist or are under development to support the detection of AI-Generated content.

4.3.2.1 AI-Generated Cloning Deepfake Detection

One of the known problems that harms artists who **create art through performances** (such as singers) is their **deepfake cloning** using GenAI.

In the field of cloning deepfake detection, some relevant techniques are based on the idea of developing person-of-interest (POI) or soft-biometric models. These models learn person-specific facial motion patterns based on head pose and facial action units for expressions. Trained on known authentic data (approximately an hour of video) for an individual, these models have demonstrated the ability to discriminate the real individual from deepfake impersonations (Agarwal et al., 2019, 2020; Christodorescu et al., 2024).

4.3.2.2 NVIDIA StyleGAN3-detector

One of the main pioneering projects in the field of AI-generated content detection is the experiment conducted in 2021 by **DARPA** ([367]) **SemaFor** ([368]) in collaboration with **NVIDIA** ([369])..

The experiment tested the ability to **detect** images from Generative Adversarial Networks (GANs) **without training data from the architecture**, mimicking the threat posed by adversaries that could develop novel GAN architectures. SemaFor performers demonstrated the ability to detect images from **StyleGAN3 ([370])** with high accuracy and no knowledge of the architecture (Christodorescu et al., 2024). They tested their product against a benchmark of both synthetic and authentic images, counting the true positives against the false positives.

Crucially, this experiment was conducted prior to NVIDIA's release of StyleGAN3, so training information could not have leaked. NVIDIA held the release of StyleGAN3 until the detectors were available and then both StyleGAN3 and the detectors were released publicly on the same day (Christodorescu et al., 2024), giving also an example of good practices from the ethical point of view.

([367]) DARPA (Defense Advanced Research Projects Agency) is an agency of the U.S. Department of Defense.

([368]) DARPA's SemaFor (Semantic Forensics) program focuses on developing technologies to detect, attribute, and understand manipulated media. It aims to combat misinformation and ensure the integrity of digital content.

([369]) NVlabs / StyleGAN3 Synthetic Image Detection, Github blog, 23 October 2024 (accessed 14 March 2025).

([370]) StyleGAN3 is a generative adversarial network (GAN) developed by NVIDIA, designed for high-quality image synthesis with improved geometric consistency and control. It is particularly notable for its ability to generate realistic, highly detailed images with fewer distortions.

The results of the experiment not only show that the existing detection algorithms are effective at identifying images from StyleGAN3, but also suggest the forensic research field is advancing on the more difficult problem of detecting images from previously unseen generators (Christodorescu et al., 2024).

4.3.2.3 Deezer: AI-Generated Music Detection

In January 2025 Deezer, a global music platform available in more than 180 countries, deployed a cutting-edge AI music detection tool. The new technology can detect music created from several generative models, such as Suno ([371]) and Udio ([372]), and is designed to generalise across similar AI music generators, provided relevant training data is available ([373]).

The software has been released **open-source** on GitHub ([374]). The training dataset is publicly available as well (Defferrard et al., 2017). This initiative aligns with the researchers' objective of promoting transparency in AI detection systems, which are often proprietary, thereby complicating independent verification and limiting the feasibility of appeals.

The tool is designed to detect AI-generated audio across both vocal and instrumental components, spanning multiple music genres (Afchar et al., 2024). It achieved **99.8% lab accuracy** in distinguishing synthetic music but faces **limits in robustness (noise/re-encoding vulnerabilities), generalisability (across autoencoders families), and mixed-content analysis**, prompting future improvements in adversarial defence, interpretability, and model adaptation.

A technical description and evaluation are available in *Annex XIV*.

([371]) Suno AI website (accessed 22 February 2025).

([372]) Udio website (accessed 22 February 2025).

([373]) Deezer Deploys Cutting-Edge AI Detection Tool for Music Streaming, Deezer Newsroom, 24 January 2025 (accessed 22 February 2025).

([374]) Deezer – Deepfake Detector, Github (accessed 23 February 2025).

4.3.3 Content-Processing Solutions (Watermarking and Fingerprinting)

4.3.3.1 Watermarking

Watermarking is the technique of modifying the digital asset to **embed information about the content's provenance**. *Annex XV* provides a technical description of the procedures used to apply and verify a watermark ([375]).

Some methods **embed the watermark and encoder into the parameters of a GenAI model** such that the watermark is intrinsically present in generated content (Fernandez et al., 2023).

There are a variety of techniques applicable to all media types (text, audio and visual content). As discussed later in this section, it is apparent that existing watermarking methods are vulnerable to various attacks that limit their effectiveness. This is particularly true within strict regimes such as TPR@1%FPR ([376]). Nonetheless, the literature does not entirely rule out the possibility of developing reliable watermarking techniques in the future (Christodorescu et al., 2024).

The discussion below contains, as an example, a detailed description of Google's watermarking technology, **SynthID**. Ongoing efforts are being made to continuously enhance the robustness of watermarking algorithms.

For instance, the system developed by **Imatag ([377])** employs state-of-the-art techniques to offer a watermarking solution that remains effective regardless of the different generative technologies to which it is applied. This system can be used to detect AI generated content ([378]). It can also integrate with C2PA (*Section 4.3.1.1*) by enabling the embedding of a

([375]) For a more extensive analysis on watermarking technologies and use cases, see Automated Content Recognition: Discussion Paper – Phase 1, EUIPO, November 2020, and Automated Content Recognition: Discussion Paper – Phase 2, EUIPO, September 2022.

([376]) TPR@1%FPR (True Positive Rate at 1% False Positive Rate) is a metric often used to evaluate models that perform binary classification (i.e., distinguishing between positive and negative). TPR and FPR are measured experimentally and are respectively the frequency with the models correctly and wrongly recognises the positive cases (i.e., assigns correctly the label "positive" and assigns wrongly the label "negative"). Then, TPR@1%FPR measures the TPR fixing the FPR at 1% and is used in critical scenarios where false positives have to be minimized in order to avoid rising false alarms.

([377]) See IMATAG website (accessed 18 March 2025)

([378]) See Label4.ai website (accessed 31 March 2025).

URL linking to a C2PA manifest. This facilitates the association of provenance-certifying information with a digital asset through a **soft-binding** approach.

Referring to the classification of possible approaches for binding digital assets to their relevant information, as outlined in *Section 3.5.1*, watermarking generally constitutes a **hard-binding** mechanism. However, as exemplified by Imatag's solution, this technique can be seamlessly integrated with other provenance-tracking solutions, enhancing its applicability and robustness, as there is theoretically no limit to the variety of information which can be embedded into a media content through watermarking.

4.3.3.1.1 Learning-Based Watermarking vs. Non-Learning-Based

Watermarking methods can be categorised into non-learning-based and learning-based. The former manually design **encoder** and **decoder**; while the latter uses neural networks as encoder/decoder and trains them using deep-learning techniques. In the **image and audio domains**, non-learning-based watermarking methods have been studied for several decades, while learning-based watermarking methods were proposed in the last several years. In the **text** domain, both non-learning-based and learning-based methods were proposed more recently (Christodorescu et al., 2024). For this reason, according to experts' interviews, ensuring the effectiveness of text watermarking is still very difficult. In particular, it has emerged that text watermarking techniques can be easily bypassed using AI-powered text rewriting.

The unique advantage of **learning-based** watermarking is that they are more **robust against post-processing that aims to remove the watermark** in the content. Such enhancement is obtained through training at the same time both the encoder and the decoder in an adversarial setup, putting between them a **post-processing layer**. The post-processing layer modifies the watermarked content produced by the encoder, and the decoder is trained to detect the watermark (Christodorescu et al., 2024). *Figure 4.3.3-1* shows an example of **concurrent adversarial training** of both the AI-based watermarking encoder and decoder for image watermarking systems.

Further information about the existing machine learning watermarking methods can be found in *Annex XVI*.

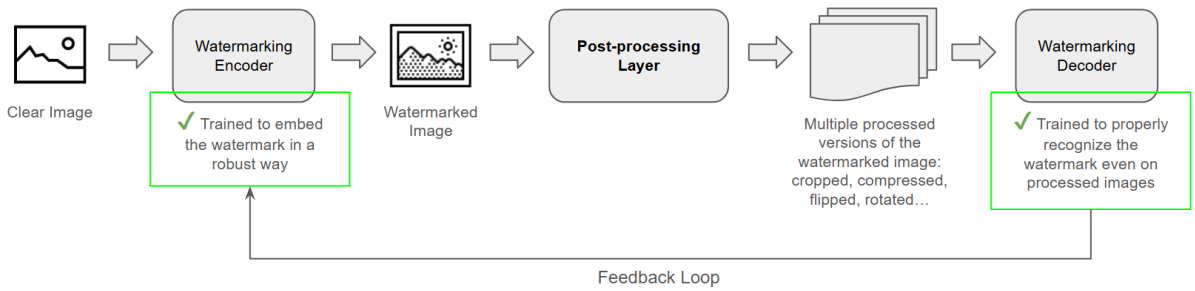


Figure 4.3.3-1: Schema outlining the adversarial training at the base of machine learning-based watermarking systems.

4.3.3.1.2 Attacks to Watermarks

There is a range of attacks that can make this technology less reliable (Christodorescu et al., 2024):

- **Common content post-processing**: these operations, while possibly **not malicious** in intent, can inadvertently remove the watermark. Common **image** post-processing operations include compression, resizing, cropping, and colour adjustments; typical **text** post-processing involves paraphrasing, word insertion, word deletion, and structural modifications; and popular **audio** post-processing includes compression, filtering, and re-recording. However, prior studies showed that **learning-based watermarking methods can be robust** against common post-processing as they can leverage adversarial training, so this issue mainly regards non-learning-based watermarks;

- **Diffusion purification attack**: This method adds random noise to the watermarked content and then uses a special AI model to clean it up, making it look like the original. In particular, it iteratively introduces **Gaussian noise** to the content and then utilises **denoising diffusion models** to undo the Gaussian noise in order to get an output that is similar to the input (the more the iterations, the more the similarity). This technique has been largely studied for removing an images' watermark; and

- **Adversarial post-processing**: an attacker can remove a watermark by making **subtle, imperceptible changes to an image**. They do this by **training an AI model**

to mimic watermark detection and then use it to alter the image in a way that removes the watermark. This can work even without knowing how the watermark system functions.

In conclusion, interviews with stakeholders reveal that copyright holders often account for the diminished economic returns associated with watermark removal. In fact, **the resulting financial losses are typically small**.

4.3.3.1.3 Google SynthID

SynthID is a tool developed by Google DeepMind which **embeds digital watermarks** into AI-generated images, audio, text or video. It started as a standalone tool and now is being integrated across Google products.

SynthID was initially introduced for AI-generated images created with **Imagen**, a generative AI tool for producing high-quality images; now it is also available as part of **Google Cloud's Vertex AI platform**, specifically for customers using Imagen through Vertex AI. This integration allows businesses and developers to generate and manage watermarked AI images securely.

Google has stated intentions to expand SynthID's reach to other tools and platforms as part of its broader AI responsibility strategy. In particular, SynthID Text (Dathathri et al., 2024), the module dedicated to text watermarking, has been open-sourced to foster research into techniques for effectively handling textual content. This decision reflects the recognition that advancements in text watermarking lag behind those in other content formats.

Google has different policies for products like **YouTube** and **Google Ads**, where the creators have to **explicitly disclose** when their content includes altered or synthetic media that depicts real people, places, or events. Labels appear within the content description and sometimes on the media itself, especially for sensitive content.

4.3.3.2 Fingerprinting

In the context of transparency of GenAI, fingerprinting involves **generating and storing in an external database a unique identifier,** commonly referred to as a fingerprint or hash ([379]). ([380])This string is computed from the content itself, taking into account its peculiar characteristics and patterns. For example, it is possible to condense a brief segment of video or audio into a string that encapsulates its key characteristics ([381]).

Fingerprinting **does not alter the content**, and it does not require inserting any marker into the content in advance. The process of **identification** consists of calculating the fingerprint of the content to be identified and comparing it with a list of known fingerprints. According to stakeholders interviewed, fingerprinting could **facilitate copyright holders' remuneration** by linking specific pieces of GenAI output to the training content from which they were derived, based on matching fingerprints.

Along with the solutions presented below, also the **ISCC proposed by Liccium** leverages a fingerprinting approach (see *Section 3.4.2.6*).

4.3.3.2.1 Google's Content-ID

Content-ID is a **key component of YouTube's business model**, addressing the issue faced by **large rights holders** owning '*exclusive rights to a substantial body of original material that is frequently uploaded to YouTube.*' ([382])

By scanning original content, extracting key features, and storing them as compressed "fingerprints," Content builds a **reference database of copyrighted works**. Newly uploaded YouTube videos are then compared against this database. If a match is detected, the right

([379]) Science policy brief – Generative AI Transparency Identification Machine Generated Content, European Commission, 21 May 2024 (accessed 21 January 2025).

([380]) For a more extensive analysis on fingerprinting technologies and use cases, see Automated Content Recognition: Discussion Paper – Phase 1, EUIPO, November 2020, and Automated Content Recognition: Discussion Paper – Phase 2, EUIPO, September 2022.

([381]) Watermarking vs. Fingerprinting, Actus Digital (accessed 21 January 2025).

([382]) How Content ID Works, YouTube Help (accessed 12 February 2025).

holder is notified and given three options: (1) remove the video, (2) claim all future ad revenue generated by the video, or (3) allow the video to remain while tracking its viewership statistics.

The technology behind Content-ID can identify not only exact copies but also **modified and distorted versions** of the content (Eriksson, 2023). Meanwhile, the system deals with false positives by producing a list of potential claims to be manually reviewed each time some uncertainty about a match is detected.

In 2018, Google claimed that rights holders selected to claim all future ad revenue generated by the video (Option 2) in 90% of the cases. In that year, YouTube claimed to have facilitated the payment of over $3 billion to rights holders who chose this option ([383]).

An **API** ([384]) is available to rights holders, enabling them to automate the procedure of uploading content while interacting with YouTube's rights management system. YouTube also provides in-person support Content-ID users, especially when the system is uncertain on a match. In those cases, Content-ID generates a **list of potential claims to be reviewed manually**. This manual intervention helps to enhance the robustness of the solution by reducing false claims ([385]).

The possibility to automate this process allows **scalability** on the rights holders side. However, the solution is based on a comprehensive database of fingerprints, which could hinder the overall technology's scalability as the number of stored fingerprints increases.

4.3.3.2.2 Audible Magic's Automated Content Recognition

Audible Magic developed an **Automated Content Recognition (ACR)** tool which enables recognising and preventing unauthorised use of copyrighted media content. According to Audible Magic website ([386]), the system claims a **match rate exceeding 99%**, with virtually zero false positives and a **service uptime surpassing 99.7%**. It supports various social media platforms where users can upload content, including Twitch, SoundCloud, ShareChat,

([383]) How Google Fights Piracy - Report, Google blog, 8 November 2018 (accessed 29 January 2025).

([384]) YouTube Content ID API, Google for Developers (accessed 17 March 2025).

([385]) YouTube Copyright Transparency Report, Google Transparency Report (accessed 12 February 2025).

([386]) Identification, Audible Magic (accessed 29 January 2025).

Dailymotion, Bolo Indya, and Suno. The technology can also be integrated with **Amazon Interactive Video Service (IVS)** ([387]), allowing for streamed content recognition.

Audible Magic specialises in fingerprinting and identification for **audio** files, but its solutions are also adaptable to **video** content (with or without audio). The company provides support for both media types through dedicated and distinct services**. In the context of GenAI, the solution is also able to identify AI-generated works if they incorporate segments from copyright protected works.** Audible Magic also actively supports compliance with Article 17 of the CDSM Directive, through the integration of **royalty reporting and payment administration** through their Administration Service.

The core infrastructure, illustrated in the *Figure 4.3.3-2* below, computes fingerprints of media content and stores them as compressed representations. These **stored fingerprints** are then referenced during the identification process of an unknown file.

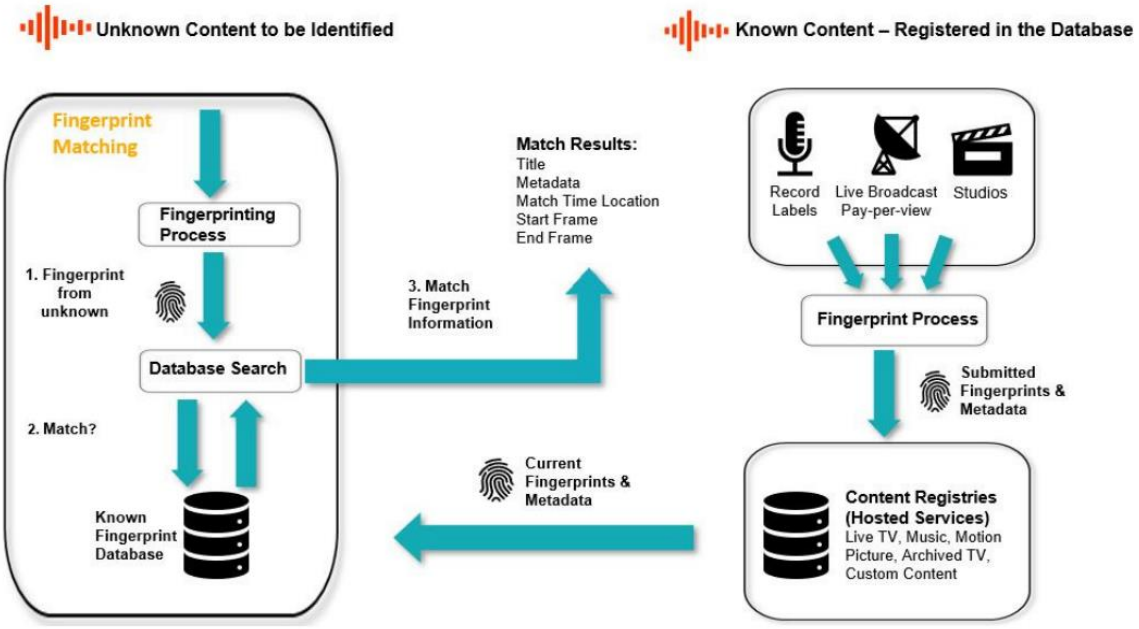


Figure 4.3.3-2: Schema illustrating the base principle behind Audible Magic's technology: when unknown content has to be identified, its fingerprint is computed and compared against the large database of fingerprints of known contents ([388]).

([387]) Identifying Copyrighted Content in Live Streams with Audible Magic | S3 E04 | Streaming on Streaming, Community AWS (accessed 17 March 2025).

([388]) Audible Magic's Content Identification, Audible Magic (accessed 29 January 2025).

The system can **identify content despite manipulations** in rate, pitch and tempo. It also handles ambient noise and **clips as short as 5 seconds** ([389]).

In 2020, Audible Magic claimed that it had the capacity to identify over 25 million media assets stemming from 1000 video suppliers and 140,000 record labels worldwide, with its registry growing by approximately 250,000 new registrations per month ([390]).

4.3.3.3 Differences Between Watermarking and Fingerprinting

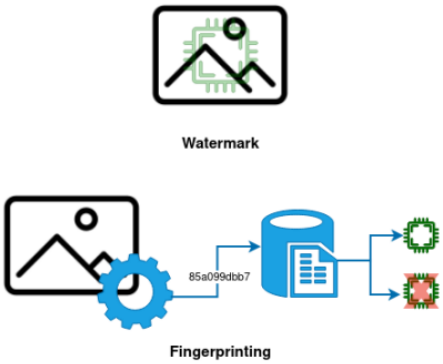


Figure 4.3.3-3: Representation of the base concepts behind watermarking and fingerprinting **([391]).**

Watermarking is more effective at **tracing and identifying** a specific content file or stream that has been previously marked, whereas fingerprinting is more suitable for **recognising a specific piece of content** and can be used in identifying copyright protected material, either in the training data or in the generated output. Watermarking can embed a unique mark into every copy of a piece of content, allowing illegal copies to be traced back to their original

([389]) Ibid.

([390]) Ibid.

([391]) Science policy brief – Generative AI Transparency Identification Machine Generated Content, European Commission, 21 May 2024 (accessed 21 January 2025).

source. In contrast, fingerprinting can only determine that a given piece of media is "identical or very similar" to the original content ([392]) ([393]).

Moreover, fingerprinting offers a distinct advantage for **forensic analysis**. At any time, a fingerprint can be computed for a given piece of content and compared against another. This flexibility is absent in watermarking, as **content that has not been pre-marked cannot be detected retroactively** ([394]).

Watermarking allows for content recognition with more certainty than fingerprinting, although **both methods are vulnerable to content modification** ([395])**.**

In terms of cost, **watermarking typically requires greater human intervention during the marking process** and involves significant logistical effort, which may include obtaining an original copy, creating the watermark, and managing its distribution. In contrast, **fingerprinting involves a far simpler marking process**. However, **fingerprinting may require more human effort during the detection phase**, as the results often require manual inspection for validation ([396]).

4.3.4 Membership Inference Attacks

Membership inference attacks can be used to determine whether a specific data sample has been used to train a model. However, it is important to note that current attack techniques exploit models' vulnerabilities to evaluate their impact. These methods do not produce deterministic results but yield only probabilistic outcomes.

Some stakeholders suggested that **retaining and authenticating user prompts** could significantly enhance the copyright enforcement of GenAI. By securely preserving the **exact**

([392]) Watermarking vs. Fingerprinting, Actus Digital (accessed 21 January 2025).

([393]) For a more extensive analysis on watermarking and fingerprinting technologies, and use cases, see Automated Content Recognition: Discussion Paper – Phase 1, EUIPO, November 2020, and Automated Content Recognition: Discussion Paper – Phase 2, EUIPO, September 2022.

([394]) Watermarking vs. Fingerprinting, Actus Digital (accessed 21 January 2025).

([395]) Ibid.

([396]) Ibid.

prompts used to generate outputs in a **certified manner**, courts could more effectively assess whether an allegedly infringing output was specifically elicited, **for instance, by naming a protected work or style within the prompt**. This approach would provide **crucial evidentiary support** in disputes concerning GenAI, offering a more direct means of establishing intent and liability in cases of copyright infringement.

The **Loss Threshold Attack** (Yeom et al., 2018) is the simplest membership inference attack. It assumes that models are trained to optimise the value of an objective function on their training set and therefore training examples produce the best results when this function is computed on them (Carlini et al., 2023). This approach thus requires knowledge of the objective function used during training, which may not always be available.

4.4 Comparison between Generative Transparency Solutions

Table 4.4-1 presents a structured comparison between **selected solutions** described in *Section 4.3*, evaluated based on the criteria outlined in *Section 4.2.2*. Each column is assigned a distinct technology, which is evaluated, on a row-by-row basis, against the predefined criteria. The cells contain brief discussions, while more details can be found in *Section 4.3*.

Watermarking is a technology that can be integrated in several solutions. For the sake of comparison, **Google's Synth-ID** is reported as an example. The **IPTC Photo Metadata Standard** (*Section 4.3.1.2*) has not been reported due to its similarity with C2PA.

Additionally, the following table includes the evaluation of **Liccium's TDM.ai protocol**, previously classified in this report as a reservation mechanism addressing the input phase of the GenAI development process (see *Section 3.4.2.6*). The reason for its inclusion is that, theoretically as still under development, it could be used for binding provenance data to digital assets, as well as for tagging generative output.

	C2PA	JPEG Trust	Watermarking (e.g., Google's Synth-ID)	Liccium's TDM.ai protocol
Typology	Enables **provenance tracking** through **metadata** binding.	Enables **provenance tracking** through **metadata** binding.	Embeds digital **watermarks** into AI-generated content.	Enables storing of **provenance information** inside federated registries. This **metadata** is (soft-) bound to digital assets through the ISCC. Theoretically, those unique identifiers can be used for emitting **detailed summaries of the data used for training**.

	C2PA	JPEG Trust	Watermarking (e.g., Google's Synth-ID)	Liccium's TDM.ai protocol
Versatility	It supports **image**, **video**, and **audio** files, while also providing **partial coverage for text** file formats such as PDF and HTML.	Primarily designed for **JPEG files**, but its high level of interoperability with other standards allows it to support **all file formats**.	Designed to embed watermarks into **images, videos** and **audio**.	The ISCC can be computed for a wide range of file formats, including **common text, image, audio, and video extensions**. The full list of supported formats is available on Liccium's website ([397]).
Openness	It is an **open standard**. Some **open-source tools** have been developed to enable C2PA manifests management.	It is an **ISO standard** ([398])**.**	Synth-ID Text has been **open-sourced** ([399])**,** whereas the modules working with other file formats are **proprietary**.	The tool suite for the protocol implementation is **open-source** and available on GitHub.

([397]) Generating ISCC Codes, Liccium (accessed 15 November 2024).

([398]) ISO/IEC 21617-1:2025, ISO (accessed 6 February 2025).

([399]) SynthID: Tools for watermarking and detecting LLM-generated Text, Google AI for Developers (accessed 12 February 2025).

	C2PA	JPEG Trust	Watermarking (e.g., Google's Synth-ID)	Liccium's TDM.ai protocol
Market Maturity	Some **major camera manufacturers** have already integrated C2PA into their products. **Widely used image editing software** and **AI models for image generation** (e.g., Dall-E) automatically embed C2PA manifests into processed images.	As of January 2025, JPEG Trust has been formalised as an ISO standard. However, its adoption and integration into real-world applications remain in early stages, making market maturity difficult to evaluate.	As of February 2025, it is integrated with **several of Google's products**, such as Gemini AI and Google Photos.	As of February 2025, the protocol is still **under development**.

	C2PA	JPEG Trust	Watermarking (e.g., Google's Synth-ID)	Liccium's TDM.ai protocol
Human-Readability	C2PA manifests require **compatible verification tools** for human readability, as they are primarily designed for machine interpretation.	Metadata is primarily machine-readable but can be interpreted by humans. Once trust indicators are extracted, typically via **dedicated verification tools**, the resulting **trust report** is designed to be highly human-readable.	Watermarks embedded via Synth-ID can only be detected using **specific tools for watermark detection**.	Metadata associated with the digital would be accessible in a **human-readable way via Liccium's platform**.

	C2PA	JPEG Trust	Watermarking (e.g., Google's Synth-ID)	Liccium's TDM.ai protocol
Cost implications	Managing C2PA manifests incurs computational costs due to the cryptographic operations required for their generation, verification, and maintenance. Those costs are borne by the entity directly managing C2PA manifests within the digital assets, meaning that both AI developers and rights holders are affected. At scale, storage and bandwidth costs may also be significant, particularly for large datasets with frequent updates.	The computational costs associated with **cryptographic operations** (e.g., cryptographic signature to ensure metadata integrity) are non-negligible.	**Optimised** for minimal computational overhead.	Even if the necessary tools are open-source, **generating the ISCC** for an asset can incur costs due to the cryptographic nature of the operations. Furthermore, **rights holders** must establish appropriate **storage for their verifiable credentials**, while **AI developers** must conduct potentially costly **searches into federated registries** for each asset when retrieving an ISCC.

Robustness	The **specifications were designed** with consideration for potential threat scenarios. Security against metadata tampering is reinforced through the use of appropriate **cryptographic algorithms**. However, **metadata removal remains unaddressed**. Given the **continuous evolution** of advanced attacks, robustness will also depend on improvements to the cryptographic mitigation measures.	Metadata integrity is preserved using cryptographic techniques, ensuring that **tampering is detected and reflected in a lower trust score**. Metadata removal is not covered by the standard. The overall implementation robustness will depend on the software design choices made by the adhering companies.	Overall, the robustness depends on the continuous development of the technology and on how the company will "feed" and develop its mechanisms. Text content: the technology works with at least three sentences and its robustness increases with the text length. Some significant weaknesses exist due to the ease with which text can be manipulated. Audio content: the technology is robust against many modifications and automatically checks different segments across the same audio.	Overall, the robustness depends on the continuous development of the technology. Unlike hard-binding, soft-binding mechanisms mitigate metadata tampering risks. **Images** remain identifiable through ISCC despite **slight modifications**, while **text** content is estimated to tolerate **up to 20% alterations**.

	C2PA	JPEG Trust	Watermarking (e.g., Google's Synth-ID)	Liccium's TDM.ai protocol
			Visual content: the technology is designed to remain detectable even after common image editing operations ([400]).	
Interoperability	It is an open standard, providing **public specifications** to enable compliance. For example, some compatible protocols are TDMRep, Liccium's TDM.ai and JPEG Trust.	Interoperability is one of the main pillars guiding JPEG Trust's development. It is an open standard, providing **public specifications** to enable compliance. It has been built upon other existing standards: ODRL, C2PA and Dublin Core.	**Synth-ID Text** has been **open-sourced**, allowing broad interoperability with other applications. The modules dedicated to **other content types** have no associated public APIs.	The **open availability** of the suite for implementing TDM.ai makes it possible to integrate this solution across applications. Moreover, the protocol already supports **C2PA** and **W3C recommendations for verifiable credentials**.

([400]) SynthID, Google DeepMind (accessed 5 February 2025).

	C2PA	JPEG Trust	Watermarking (e.g., Google's Synth-ID)	Liccium's TDM.ai protocol
Scalability	The Content Authenticity Initiative offers some **APIs for creating, signing and parsing C2PA manifests**. Thus, manual operations can be automated, enhancing scalability. Due to the presence of cryptographic operations, processing a large quantity of digital content could result in high computational time.	The standard does not specify the practical implementation details of the software used to assess the Trust Indicators and generate the Trust Report. Consequently, the **design choices made by companies implementing the standard** will impact the overall scalability of the solution.	**Synth-ID Text** has been tested on around 20 million of the models' generative outputs – with and without watermark. The users did not perceive any relevant difference between the two, demonstrating the solution's capability at a high scale (Dathathri et al., 2024). The modules dedicated to **other content types are scalable within Google's environment** and integrated in an increasing number of Google's products.	AI developers' need to consult Liccium's federated registries and perform similarity checks when **searching for a given ISCC.** This may be a limitation to the scalability of the solution as the number of managed digital assets increases.

	C2PA	JPEG Trust	Watermarking (e.g., Google's Synth-ID)	Liccium's TDM.ai protocol
Reliability	If metadata are completely removed from the content, so would the C2PA manifest and the related protection. The protocol only guarantees that the relative C2PA manifests contain the original information from their original authors, but does not guarantee that those authors are reliable themselves.	The standard is designed to take into account all the aspects of the digital asset (content, metadata, history) with the aim of computing **reliable trust indicators**.	Given the stated limitations (e.g., heavily modified images or too short pieces of text), Google Synth-ID is designed for offering a reliable solution, with the possibility to be tested at large scale.	TDM.ai provides a reliable solution due to its ability to **identify assets independently of their location** and **securely store associated metadata**, ensuring consistency and traceability.

Table 4.4-1: Comparison between Transparency measures.

4.5 Content Regurgitation

The issue of '**models as infringement**' and **infringing output** are often linked to technical discussions on '**memorisation**' (see *Section 3.2*) and '**overfitting**' (see the *Glossary* for the definition of 'overfitting'). These two issues are technically distinct but often occur together *"due to the fact that while overfitting is a consequence of learning a feature which appears numerous times in the training set, memorisation can happen even with a single example contained only once in the training data set"*([401]).

A specific type of infringement, informally referred to as **'The Snoopy Problem',** is also gaining widespread attention. This issue arises when a generative model inadvertently reproduces **copyrighted characters (or other copyright protected content)** despite not having been explicitly trained to do so([402]). This occurs because models are trained on various representations of a character or may retrieve related material during inference, allowing them to generate new depictions that do not directly replicate any single existing representation. However, despite not copying a specific image verbatim, these outputs can still constitute copyright infringement if the character itself is protected by intellectual property rights. In the next chapters some studies on this exceptional issue are presented.

4.5.1 Models Creating Infringing Reproductions

There are two distinct issues with AI systems that may be producing content that infringes on copyright. The first issue, occasionally raised, is that once AI models/systems are trained, they **potentially constitute of infringing reproductions of works**, irrespective of whether the model is actually used to produce generative output. The second issue is that, regardless of whether an AI model reproduces a work, when it generates content, some of this **output may infringe upon copyright** on a case-by-case basis. This can happen both **when the copyrighted work related to the infringement is present or not in the original training dataset**. The first case is referred to as

([401]) Hense and Mustać (2024), p. 287.

([402]) Generative AI's "Snoopy Problem" Makes Avoiding Copyright Infringement a Challenge, Fast Company, 25 March 2024 (accessed 14 March 2025).

memorisation (see *Section 3.2*), while the second refers to the possibility of a model generating an already existing work.

Copyright compliance regarding training data inputs is directly referenced in legal provisions, namely in **AI Act Article 53**. Additionally, output transparency measures are directly connected to the obligations of **AI Act Article 50**. The issue of infringing output, however, may be generally inferred from the general principles of **copyright law**. Furthermore, the distinction between the 'AI models as infringing reproduction' issue and the issue of infringing generative content is important because they invoke different copyright relevant acts.

As mentioned above, there is a view held by some stakeholders, rights holders in particular, that AI models themselves are infringing reproductions ([403]). If that legal theory is correct then it has major implications for the AI ecosystem. Not only would the creation of AI models represent unauthorised reproductions of an incredibly large number of works, but the distribution, deployment, and use of these models may amount to unauthorised distribution (and possible communication to the public of) protected works.

The starting point of this view is that both LLM models and diffusion models analyse content through **tokenisation of content** to extract probabilities, correlations, and patterns with subsequent tokens.

This debate is connected to technical discussions on the phenomenon of '**memorisation**' of content within models which can lead to 'regurgitation', i.e., production of output which explicitly reproduces some training input. For more details about copyright implications of memorisation, see *Section 3.2.4.*

An often-repeated adage stated by AI developers is that phenomena like memorisation and regurgitation *'...are **not features but bugs ([404])'***, alluding to the fact that these outcomes are not intentional by design, and developers are actively investing in research and development to minimise them. The extent to which these 'bugs' manifest in AI systems and the question of whether AI models are reproductions also impacts the debate on if AI training constitutes TDM.

([403]) In particular, in an August 2024 report commissioned by the rightsholder organisation 'The Copyright Initiative' Dornis and Storber argue that protected works are stored 'inside' a model's parameters, constituting unauthorised reproductions, even if these reproductions are not immediately perceptible. See: Dornis, Tim W. and Stober, Sebastian, Copyright Law and Generative AI Training - Technological and Legal Foundations (Urheberrecht und Training generativer KI-Modelle - Technologische und juristische Grundlagen); 2024.

([404]) In computer science, a bug is an error, flaw, or fault in software or hardware that causes it to behave unexpectedly or incorrectly.

Finally, some interviewed stakeholders signalled the need to bind input prompts to models' outputs, so that they could be used as additional proof in court cases.

4.5.2 Plagiaristic Output Safeguards

The mitigations to address plagiaristic outputs, belonging to the category **X3** defined in *Section 2.5*, are described below. These safeguards apply to both the **input and output** of GenAI systems. Indeed, filters can be designed to either block prompts that contain plagiaristic requests or to prevent plagiaristic outputs at a post-generation stage.

It is important to recognise that when a model generates infringing content, the user may not always be aware, leading potentially to **unintentional infringements**. This can be mitigated by implementing alert systems notifying users when a generated output has a likelihood of being plagiaristic. Such detection tools, while not necessarily state-of-the-art, can serve as cost-effective measures to reduce the risk of unintentional copyright violations. The sections below summarise various technical measures in this regard.

4.5.2.1 Examples of Available Solutions

Text-Matching and Similarity Algorithms: Integrated plagiarism checkers using **similarity detection algorithms** aim to check if generated output violates IP rights. By employing mathematical and statistical techniques ([405]), these tools compare generated content against a vast corpora of known materials. Plagiarism checker platforms, such as Copyscape and Turnitin, are being adapted to work directly with generative models, flagging high-risk content that may require human review.

In particular, **Copyscape** is a tool to check if new content, whether written by humans or generated by AI, includes text from elsewhere on the Internet. It has been widely adopted already by many digital publishers ([406]).

([405]) Some examples are: n-gram analysis, cosine similarity checks and sentence-level semantic matching.

([406]) Testimonials, Copyscape (accessed 6 January 2025).

Turnitin is a plagiarism detection software that can scan student's works against a large database of academic works, publications, and other materials on the internet. It has been integrated into platforms such as Moodle and Canvas.

Some AI service providers have developed proprietary content-checking modules. These operate by assessing the output's similarity to known copyrighted works and filtering out content that exceeds a defined similarity threshold.

Microsoft's multi-layered safety approach includes **several layers of mitigations,** including for copyright-related issues. The technology is based on a secondary and **independent AI system** analysing the protected model's input and output, as schematised in *Figure 4.5.4-1.*

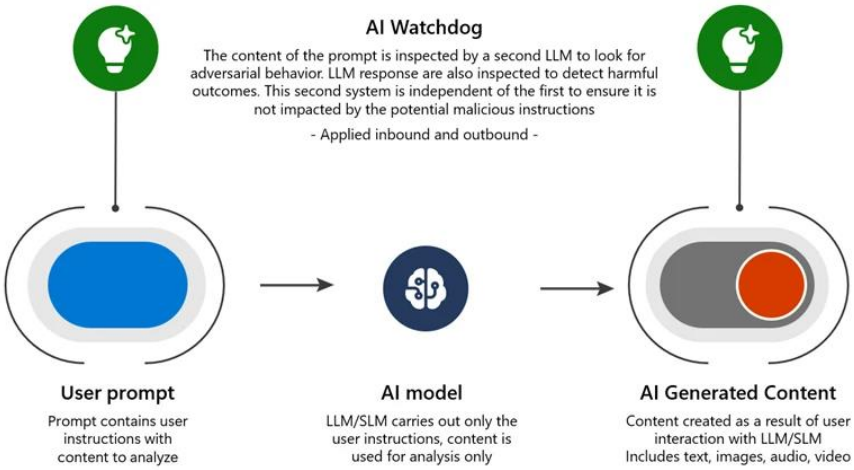


Figure 4.5.4-1: Diagram outlining the architecture of the multi-layered safety approach implemented into Microsoft's GenAI systems ([407]).

On the input side, Microsoft leverages **prompt filtering and rewriting**, while the **output is checked against an AI-powered block list**.

([407]) How Microsoft Discovers and Mitigates Evolving Attacks against AI Guardrails, Microsoft Security Blog, 11 April 2024 (accessed 6 January 2025).

Copyright Delta ([408]) is a company pioneering in song detection and protection software. In anticipation of upcoming regulations requiring transparency in AI training datasets, the company's platform is designed to manage derivative information. This includes generating comprehensive summaries of all copyrighted material used in training. Additionally, Copyright Delta provides a trusted timestamping service, enabling creators to document and verify ownership at any stage of the creation process. This service helps establish a secure and transparent record of intellectual property rights.

Software plagiarism detection requires a more nuanced approach, as traditional text-matching techniques alone are insufficient to identify copied code. Plagiarism can be concealed through techniques such as renaming variables and functions, reordering instructions, or inserting redundant lines of code. As a result, the source code may look different while still producing the same final software. Automatic software plagiarism detection tools are widely used, and advancements in GenAI have further improved their ability to identify obfuscated or transformed code fragments. Modern plagiarism detection tools incorporate advanced code analysis techniques, such as abstract syntax tree (AST) analysis and control flow analysis, which are also fundamental in compiler design ([409]).

Stakeholders' interviews revealed that, in practice, an external audit mechanism is often added to the GenAI system to ensure the effectiveness of the output guardrails.

4.5.2.1.1 Copilot's Duplication Detection Filter

Copilot, a well-established GenAI product between code developers, allows users to enable a duplication detection filter. Both Microsoft and GitHub provide indemnity for their Copilot products: if any suggestion made by Copilot to a user is challenged as infringing on third-party intellectual property rights and the same user has the filter enabled, then the contractual terms will indemnify the user.

([408]) Copyright Delta website (accessed 6 February 2025).

([409]) A code compiler is a software program designed to parse source code and convert it into a binary representation required for execution. During this translation process, compilers employ sophisticated techniques to analyse the code and identify potential errors.

With the filter set to “Block”, the generated lexemes are compared against the ones indexed in the public code repositories on GitHub. A generated text containing more than 65 lexemes of matching content in GitHub’s public repositories (about 150 characters, without considering whitespaces), will not be suggested to the user ([410]). Below is a schematic of the data flow during Copilot’s functioning, in which a proxy external to the LLM acts as a protection between the LLM and the user.

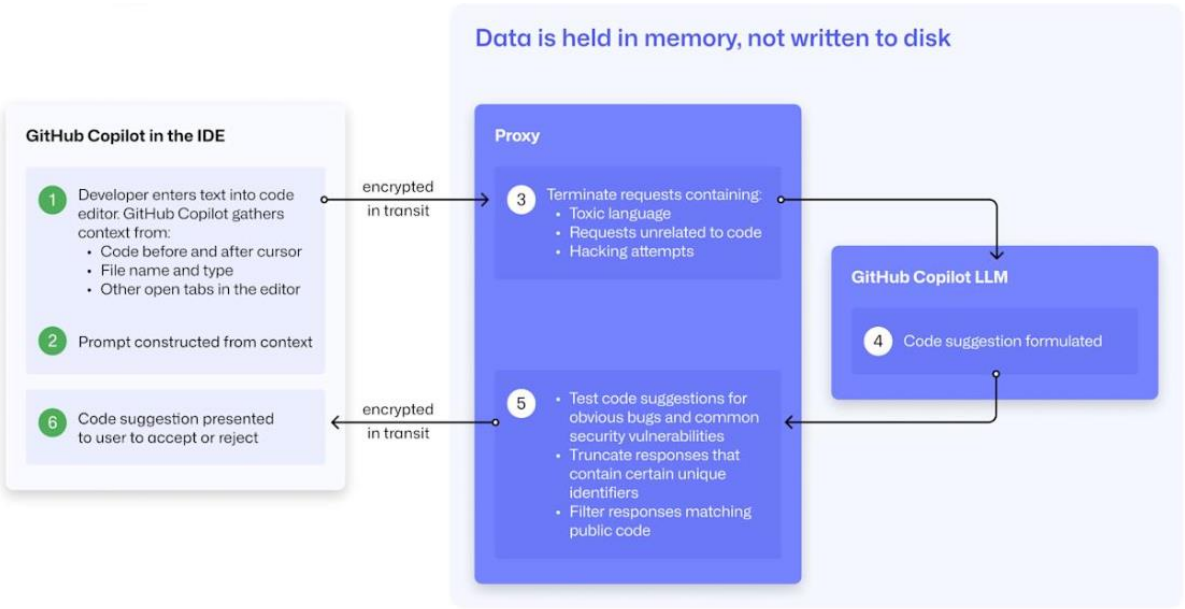


Figure 4.5.4-2: Diagram outlining the functioning principle of GitHub Copilot’s input and output safety guardrail ([411]).

4.5.2.1.2 Originality.ai’s Plagiarism Checker

Originality.ai ([412]) offers an online tool for checking if **text content** is similar to already-existing works, by comparing it with extensive content databases.

([410]) ‘Responsible AI Adoption’.

([411]) How GitHub Copilot handles data, Github resources (accessed 14 January 2025).

([412]) Originality AI website (accessed 29 January 2025).

The company reports achieving 90% accuracy for global plagiarism detection and 74% for patchwork plagiarism, plagiarism from several sources combined. These results are achieved through advanced machine learning and plagiarism checking algorithms, which take into account the **different forms of plagiarism** (e.g., global, paraphrase, unintentional) and the techniques used to disguise it. Moreover, they leverage **multilanguage** capabilities by checking originality across languages (this functionality builds upon the Google search engine) and by updating their algorithms to match the evolving plagiarism technologies.

The tool also enables users to generate and share plagiarism reports (e.g., to verify that a piece of writing is authentic).

The company claims that it has been endorsed by some of the well-known publishing companies, such as The New York Times and The Guardian ([413]).

4.5.2.2 Solutions Under Development

Contextual Rewriting AI: To avoid near-exact replication of input data, contextual rewriting modules are being developed to operate in tandem with GenAI models. These modules ensure that generated content is sufficiently transformed from the original source. Instead of simply replicating, the AI learns to abstract ideas and generate content with novel wording and structure.

Advanced Semantic Distillation: Another evolving technology involves semantic distillation, which ensures that outputs retain the thematic or informational value of the source material without replicating specific phrasing. This technology is intended to work at the generation stage, continuously assessing outputs to ensure compliance with originality requirements.

4.5.2.2.1 COPYCAT: Evaluating Text-to-Image Models' Tendencies to Generate Copyrighted Characters

Copyrighted characters pose a difficult challenge for image generation services: in 2024 a study (He et al., 2024) demonstrated that visual models can generate figures closely resembling famous

([413]) Ibid.

characters even if their names are not explicitly mentioned in the input prompt. As seen before in this report, one lawsuit in China has resulted in liability for a GenAI system that generated the copyrighted character Ultraman (see *Section 2.3.3.1*).

He et al. (2024) introduced a benchmarking suite called **COPYCAT to assess runtime mitigation strategies** implemented by some of the leading GenAI models. They applied it to the following models: **Playground v2.5** (Playground AI), **Stable Diffusion XL** (Stability AI), **PixArt-α** (PixArt AI), **DeepFloyd IF** (DeepFloyd), **DALL·E 3** (OpenAI), and **VideoFusion** (Runway).

To perform the evaluation of the Model Under Test (MUT), COPYCAT defines two evaluations to be computed, with the aid of the model GPT-4V:

- **DETECT**: This evaluation measures how frequently GPT-4V correctly identifies copyrighted characters from a predefined list of 50 characters based on images generated by the MUT in response to corresponding textual descriptions; and
- **CONS**: it evaluates how well the generated image **aligns with the key characteristics of the copyrighted character using the VQAScore ([414])**.

COPYCAT prescribes to iterate over the list of 50 characters, prompting the MUT with descriptions, check the generated image with the aid of GPT-4V, and compute DETECT and CONS, which represent the performance of the MUT in avoiding generating copyrighted characters. An optimal mitigation strategy should aim to **reduce DETECT** scores—indicating fewer instances of unauthorised character replication—while **maximising CONS** scores to ensure the generated output retains artistic coherence and usability.

Their findings revealed that strategies like "**prompt rewriting**" are insufficient when used as standalone guardrails. Although DALL·E rejects user requests explicitly mentioning copyrighted characters and rewrites prompts into more generic descriptions, researchers were still able to generate visual representations closely resembling copyrighted characters (He et al. 2024). In *Figure 4.5.4-5* some examples of successful extractions are reported:

([414]) VQAScore (Visual Question Answering) is a way to measure how well an AI model performs at answering questions about images. It compares the model's answers to the correct ones using a scoring system reflecting the fragmented nature of human expectations.

Figure 4.5.4-5: Examples of generated images representing copyrighted characters. They demonstrate that existing mitigations may be ineffective against extraction attacks (He et al., 2024).

The paper proposes to couple "prompt rewriting" with "**negative prompting**". This approach involves not only specifying what the model should include in its generated output but also explicitly defining **what elements should be excluded**—such as key features associated with copyright-protected characters. (Further analysis of this method is available in *Annex XVII*).

4.5.2.2.2 Prompt Rewriting-Enhanced Genericization (PREGen)

Building upon the work presented in the previous chapter, another study (Chiba-Okabe & Su, 2024) proposes **quantifying the level of originality** in data with the aim of avoiding copyrighted generations. The underlying principle is that, as a work becomes widely disseminated and frequently utilised, its perceived originality diminishes in legal considerations of copyright.

In particular, PREGen is designed to enhance the effectiveness of the mitigation strategy that involves both **prompt rewriting and negative prompting**.

In addition to these mitigation techniques, their method adds **'genericization'** to a model's output by **internally producing more samples**, estimating the originality of each sample, and selecting **the one with the lowest estimated originality for the final output**. For computational efficiency, the

originality estimate of a sample can be **cross computed** by measuring the **distance** (CLIP ([415]) was selected as the distance metric) to the other internally produced samples.

The algorithm first modifies the input prompt to a **clean prompt**, removing references to copyrighted elements, using a LLM. It further generates **multiple variations of the clean prompt**. The effectiveness of prompt rewriting is enhanced by further incorporating a **negative prompt**. Subsequently, each generated input prompt is fed into the generative model. Finally, the algorithm outputs the generation that has the **lowest originality**.

The researchers utilised **COPYCAT**, as detailed in *Annex XVII*, to evaluate their proposed technique.

4.6 Unlearning

Once data is embedded into a **model's weights** through training, its removal can be complex because there is no way to precisely isolate specific information from the parameters. Data erasure is only feasible before training, affecting individual dataset entries (Cooper et al., 2024). Sometimes, there is no other solution than retraining the model without the data to be erased. Current techniques, such as **machine unlearning, have not been proven effective** for large-scale foundation models (that are often the base of contemporary GenAI models). Furthermore, the **techniques employed in model training** may influence the feasibility and the complexity of unlearning (Zhang, Xia, et al., 2024).

An effective method would remove unwanted knowledge while maintaining **locality**, i.e., preserving non-targeted knowledge and the model's reasoning ability. With limited research in that field, it is unclear if existing methods are suitable (Avoiding Copyright Infringement via Machine Unlearning, 2024).

Zhang et al. (2024) proposes a categorisation of the unlearning methods currently available:

- **Exact Machine Unlearning**: involves the targeted removal of specific data points from a model through an **accelerated retraining process**. Methods such as **SISA** provide exact unlearning utilising dataset partitioning and re-training only the affected segments, thus

([415]) Contrastive Language–Image Pretraining (CLIP); see the *Glossary* for more details.

minimising computational overhead. Despite its efficiency, this approach is **resource-intensive** and may inadvertently introduce fairness issues by disproportionately impacting certain subsets of data.

- **Approximate Machine Unlearning**: Approximate methods **adjust model weights** to simulate the effect of removing specific data without full retraining. Techniques include the calculation of data influence on model parameters or storing training updates for selective rollback. In particular, the first approach is followed by some techniques emerged from recent studies: **Stable Sequential Unlearning (SSU)** and **Approximate Unlearning with Idiosyncratic Expressions Replacement.** However, these methods are prone to **over-unlearning**, which can significantly impair the model's overall efficacy.

Annex XVIII provides a general introduction to the technical aspects of unlearning, including the summarised descriptions of several unlearning techniques, along with machine learning concepts which are common between machine unlearning and model editing approaches. For further details about the mentioned unlearning approaches, see also *Annex XIX*, *Annex XX* and *Annex XXI*.

4.7 Model Editing

"*In order to respond to changes in the world (e.g., new heads of state or evolving public sentiment on a particular topic) or correcting for instances of underfitting or overfitting the original training data, the ability to quickly make targeted updates to model behaviour after deployment is desirable.*" (Mitchell, Lin, Bosselut, Manning, et al., 2022)

Model editing seeks to modify a model's parameters and adjust the learned information. Unlike unlearning, this approach does not erase knowledge from the model but rather **updates or corrects specific learned information** while preserving the overall structure and functionality of the model.

Model editing could, theoretically, be achieved through **fine-tuning** techniques (see *Section 3.1.4*), using **pairs of inputs and their corresponding updated desired outputs** as training data. However, as Mitchell et al. (2022) notes, this approach risks causing the model to **overfit** to the fine-tuning data (see the Glossary for the definition of 'overfitting').

In response, alternative approaches have been explored. Since these methods do not modify the original model directly but instead introduce side paths to alter its behaviour, they can be referred to as **"band-aid" solutions** (Zhang, Finckenberg-Broman, et al., 2024), a sort of interim corrective measure. Model editing methods, such as those proposed by Mitchell et al. (2022), preserve the original model intact while storing modifications separately. The model's output is then generated by combining the original model's predictions with the stored modifications.

Moreover, when implementing unlearning strategies in models, it is crucial to ensure that the edits effectively modify the target data and that these modifications do not lead to a **decline in the model's overall performance**. This consideration is essential and should not be underestimated when evaluating the effectiveness of an editing strategy.

Yao et al. (2023) conducted a survey on existing model editing methods (see *Figure 4.7-1*). Among the possibilities, Semi-Parametric Editing with a Retrieval-Augmented Counterfactual (SERAC) and Model Editor Networks using Gradient Decomposition (MEND) emerge for their high scores. MEND belongs to the first category of editing methods, i.e., the ones aiming at fine-tuning model's parameters, while SERAC belongs to the "band-aid" solutions since it leverages an external memory. The two approaches are further detailed below.

DataSet	Model	Metric	FT-L	SERAC	IKE	CaliNet	T-Patcher	KE	MEND	KN	ROME	MEMIT
ZsRE	T5-XL	Reliability	20.71	**99.80**	67.00	5.17	30.52	3.00	78.80	22.51	-	-
		Generalization	19.68	**99.66**	67.11	4.81	30.53	5.40	89.80	22.70	-	-
		Locality	89.01	98.13	63.60	72.47	77.10	96.43	**98.45**	16.43	-	-
	GPT-J	Reliability	54.70	90.16	**99.96**	22.72	97.12	6.60	98.15	11.34	99.18	99.23
		Generalization	49.20	89.96	**99.87**	0.12	94.95	7.80	97.66	9.40	94.90	87.16
		Locality	37.24	**99.90**	59.21	12.03	96.24	94.18	97.39	90.03	99.19	99.62
COUNTERFACT	T5-XL	Reliability	33.57	**99.89**	97.77	7.76	80.26	1.00	81.40	47.86	-	-
		Generalization	23.54	**98.71**	82.99	7.57	21.73	1.40	93.40	46.78	-	-
		Locality	72.72	**99.93**	37.76	27.75	85.09	96.28	91.58	57.10	-	-
	GPT-J	Reliability	99.90	99.78	99.61	43.58	**100.00**	13.40	73.80	1.66	99.80	99.90
		Generalization	97.53	**99.41**	72.67	0.66	83.98	11.00	74.20	1.38	86.63	73.13
		Locality	1.02	**98.89**	35.57	2.69	8.37	94.38	93.75	58.28	93.61	97.17

Figure 4.7-1: Comparison between different model editing techniques, based on three metrics: (1) Reliability, which is the effectiveness of the edit, (2) Generalisation, which is the capability of the edit to correctly influence related model's generations, and (3) Locality, which is the method's capability to correctly avoid influencing model's generations which are unrelated to any edit record (Yao et al., 2023).

4.7.1 Model Editor Networks using Gradient Decomposition (MEND)

MEND (Mitchell, Lin, Bosselut, Finn, et al., 2022) is a technique designed to make quick and precise adjustments to the behaviour of a pre-trained AI model by using a single example of how the model should respond (i.e., the pair composed by the input and the new desired output).

The Gradient: A Key Mathematical Tool for Fine-Tuning

The gradient is a mathematical concept that measures how a **function's output changes in response to small variations in its input**. It is represented as a vector that points in the direction of the steepest increase of the function, with its magnitude indicating the rate of change.

In the context of machine learning, the gradient is crucial for optimising models training through a process known as **gradient descent**. This is an iterative method aiming at adjusting model parameters to minimise a **loss function**, which quantifies the model's performance.

Computing **the gradient of the loss function** involves calculating the partial derivatives of the loss function with respect to each parameter. These derivatives quantify the sensitivity of the loss function to small changes in each parameter. The computed gradient is represented as a matrix whose dimensions are proportional to the number of model parameters.

During training, the model updates its parameters in the direction that reduces the loss, moving opposite to the gradient, thereby improving accuracy over successive iterations.

MEND's approach includes to transform the **gradient** by applying a **low-rank decomposition,** a mathematical technique that simplifies complex matrices, enabling efficient computation. In AI models, it reduces gradient complexity, making weight adjustments computationally feasible.

Furthermore, the method is based on developing small, auxiliary **editing networks**. Those are **trained to transform the gradient in a way that captures the necessary weights' adjustments** while avoiding overfitting or disrupting the model's broader functionality. Once trained, the networks forming the MEND infrastructure enable rapid edits to the pre-trained model's behaviour without requiring additional extensive training. These networks work **alongside** the original model to implement targeted modifications without altering the model itself.

An assessment of this approach under the lens of the software qualities mentioned by the AI Act is provided in *Annex XXII.*

4.7.2 Semi-Parametric Editing with a Retrieval-Augmented Counterfactual Model (SERAC)

SERAC (Mitchell, Lin, Bosselut, Manning, et al., 2022) was designed as a solution to address the **scalability challenges** of model editing. In other words, it aims to prevent the degradation of the whole model's performance that can occur when many edits are applied to a single model.

This solution enables the storage of edits within an **explicit memory system ([416]),** allowing the model to reason over these edits and adjust its predictions accordingly. Additionally, SERAC incorporates an **AI-based classifier** trained to identify whether an incoming input corresponds to one or more edits stored within its explicit memory. If a match is detected a separate component, SERAC's **counterfactual model**, generates the output instead of the base model, integrating the relevant edit records into the response. The resulting infrastructure is summarised in *Figure 4.7.4-1.*

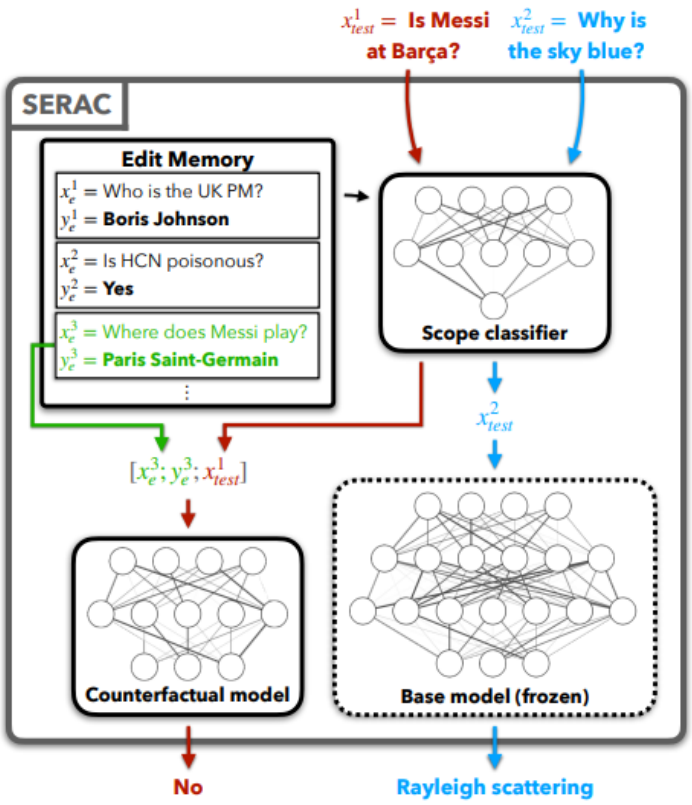


Figure 4.7.4-1: SERAC's high-level architecture ([417]).

([416]) Explicit memory refers to structured data storage mechanisms that operate independently of the model's parameters.

([417]) The blue annotations illustrate the data flow when the scope classifier determines that the input prompt is unrelated to any edit record. In this scenario, the final system output corresponds to the base model's generation. Conversely, the red annotations indicate the pathway followed when the input prompt is identified as related to an existing edit record. In

SERAC appears to perform well, possibly as it does not rely on the gradient, a complex mathematical operator that significantly increases the computational cost of the editing process. Moreover, it prescribes simultaneous adjustments to all model parameters within the gradient matrix: this introduces the risk of degrading the existing model's parameter delicate configurations, which are themselves the product of a sophisticated training process.

SERAC adopts a **'gradient-free'** memory-based approach. However, its ability to leave the base model unchanged could introduce potential **security vulnerabilities.** Since the original information remains encoded within the base model, albeit overridden by SERAC, this may render it susceptible to extraction attacks.

A more in-depth assessment of SERAC with respect to the software qualities highlighted by the AI Act is provided in *Annex XXIII*.

this case, the counterfactual model intervenes, generating an updated output that incorporates the relevant edit information (Mitchell, Lin, Bosselut, Manning, et al., 2022).

4.8 Contractual Indemnification

With the emergence of the AI technology boom, a number of GenAI service providers initially included clauses in their terms and conditions to protect themselves from liabilities arising from users' wrongdoing. For example, the terms and conditions for image generation service Midjourney, stating that *"You understand and agree that we will not be liable to you or any third party for any loss of profits, use, goodwill, or data, or for any incidental, indirect, special, consequential or exemplary damages, however they arise."* ([418])

More recently, major AI model developers have been introducing copyright indemnification clauses within their terms and conditions under which they accept liability under certain conditions. An example of this is the GitHub Customer Agreement, which also covers the AI-driven Github Copilot for code generation. In 2022, the GitHub Agreement stated that it would defend a customer against third-party intellectual property claims made against a paid product ([419]). Stock image provider Shutterstock was also an early adopter of user indemnification provisions for its GenAI services. Shutterstock's approach follows user generation of AI images with an internal experts review process, before images are cleared for commercial use and backed by indemnification protection ([420]). Adobe also introduced user indemnification for its Firefly service (text-to-image GenAI model), which states that it would cover claims that allege that Firefly output directly infringes a third party's intellectual property ([421]).

Another development was the September 2023 announcement by Microsoft regarding its 'Copilot Copyright Commitment' ([422]). While GenAI indemnification provisions were previously included by GitHub (a subsidiary of Microsoft), the new Commitment claimed that Microsoft would defend a commercial customer of its various Copilot AI and Bing Chat Enterprise services, in case of third-party claims of copyright infringement in generated output. In November 2023, this was expanded

([418]) Terms of Service, Midjourney (accessed 14 March 2025).

([419]) General terms, Github (accessed 14 March 2025).

([420]) Enjoy peace of mind with full legal protection on AI-generated images, Shutterstock (accessed 14 March 2025).

([421]) Firefly Legal FAQs – Enterprise Customers, Adobe, 13 September 2023 (accessed 14 March 2025).

([422]) Microsoft announces new Copilot Copyright Commitment for customers, Microsoft blogs 7 September 2023 (accessed 14 March 2025).

into Microsoft's 'Customer Copyright Commitment' which also covered its Azure OpenAI Service ([423]). This indemnification is subject to the customer complying with the guardrails and content filters built into the products and services. The specific guardrails that must be complied with depend on the specific use case, with specific mitigation practices set out for code generation (i.e., use of Microsoft's GitHub Copilot) ([424]). Additionally, Microsoft requires its commercial customers to adhere to a 'Generative AI Services Code of Conduct', which includes obligations to build any downstream applications (i.e., AI systems based on Microsoft models) with certain responsible AI mitigation requirements ([425]). The Microsoft Universal Licencing Terms for Online Services also sets out a specific conditions for the Customer Copyright Commitment to apply ([426]).

In October 2023, Google announced the introduction of its GenAI legal indemnification ([427]). Google summarises that *"if you (customer) are challenged on copyright grounds, we (Google) will assume responsibility for the potential legal risks involved"*. It describes its indemnification approach as being 'two-pronged', relating to both the use of training data, and generative output. The company suggests that indemnity for IP issues that may arise from its use of training data was always covered by its 'general services indemnity', but it has made this more explicit in response to demand from its customer base. Indemnity protection for generative output applies to the use of 'Duet AI' (later rebranded as 'Gemini') in Google Workspace and Google Cloud services, under certain conditions ([428]).

Many other major players in the AI development market have incorporated user indemnification clauses. IBM announced that it would offer intellectual property indemnity for its 'Granite' foundation

([423]) Microsoft Azure AI, data, and application innovations help turn your AI ambitions into reality, Microsoft blog, 15 November 2023 (accessed 14 March 2025).

([424]) Customer Copyright Commitment Required Mitigations, Microsoft learn, 21 May 2024 (accessed 14 March 2025).

([425]) Microsoft Enterprise AI Services Code of Conduct, Microsoft learn, 4 January 2025 (accessed 14 March 2025).

([426]) The customer must: (i) not circumvent the product's measures such as content filters and prompt restrictions, (ii) not have used the output with constructive knowledge that it is infringing, (iii) has rights to use any input used to customise the model, (iv) raise a claim specific for commercial trademark use, and (v) have implemented all required mitigation measures. See For Online Services', Microsoft Licensing – Product terms, (accessed 14 March 2025).

([427]) Shared fate: Protecting customers with generative AI indemnification, Google Cloud blog, 13 October 2023 (accessed 14 March 2025).

([428]) Google's indemnity is subject to customers following 'responsible AI practices', including not intentionally creating infringing generative output, not bypassing other output-control measures within the model, and ceasing use after third-party infringement claims See: Google Cloud Terms Directory, Service Specific Terms (Last modified February 6, 2025), Service Terms, Section 19(i) Generative AI Services: Additional Google Indemnification Obligations. See Service Specific Terms, Google Cloud (accessed 14 March 2025).

models ([429]) while Amazon Web Services (AWS) Customer Agreement also includes an intellectual property indemnification clause ([430]). Furthermore, OpenAI has implemented contractual terms (which it has branded as 'Copyright Shield') which provides for generative output indemnity for ChatGPT Enterprise customers ([431]).

A few observations can be made based on a review of the indemnification provisions of major AI. First, like general indemnification provisions in various fields such as product liability, users are required to follow specific usage guidelines to be covered. This includes complying with model guidelines and internal measures for mitigating infringing output, as well as using output in good faith (i.e., without constructive knowledge of infringement). Second, while many services provide for a general indemnification clause (which may or may not cover copyright infringement of generative output), several market-leader developers have explicitly drafted provisions which refer to generative output copyright liability. Third, there appears to be a trend where broad indemnification provisions are granted for commercial and enterprise users, but not necessarily users of free/non-subscription based services.

Based on the interviews conducted during this research and publicly available information, there are not yet any known instances of users relying on these indemnification clauses. This appears consistent with the trend in the litigation landscape, where rights holders initiate legal actions directly against AI companies, and not necessarily end-users.

([429]) IBM Announces Availability of watsonx Granite Model Series, Client Protections for IBM watsonx Models, IBM, 28 September 2023 (accessed 14 March 2025).

([430]) AWS Customer Agreement, AWS, 5 March 2025 (accessed 14 March 2025).

([431]) Similar to the Microsoft and Google indemnification conditions, OpenAI's indemnity does not apply in specific conditions, where (i) a customer has constructive knowledge of infringement, (ii) an end-user has bypassed restriction measures, (iii) the output was modified (or used in combination with third-party services), (iv) the customer did not have the rights for the inputs or fine-tuning files used to generate the output, (v) the infringement related to commercial trademark use, or (vi) the content is from a third party offering. See Service terms, OpenAI (accessed 14 March 2025).

4.9 Institutional Support by IP Offices

As with the issues related to GenAI input, public institutions such as national and supranational intellectual property offices can provide **valuable support** in relation to GenAI output issues. Possible opportunities include:

- ***Documenting measures related to output*** - Similar to listing and documenting various TDM opt-out measures, institutions can provide public information on measures used by GenAI providers to mitigate potential infringing output, as well as measures used to detect and identify synthetic output. This could provide information on developing technologies and good practices to address some of the copyright-related risks of GenAI development to lawmakers and regulatory bodies. They could also facilitate multi-stakeholder discussions on emerging copyright-related risks, in a way that would support solution-driven approaches to the benefit of all relevant stakeholders. Smaller AI developers, in particular could be made aware of the **state-of-play of key protocols and measures** at their disposal. End users may also be able to better understand the technical measures built into GenAI systems to **improve their own user experience** and ability to choose a GenAI deployer based on personal preferences. Furthermore, rights holders could be made aware of the safeguards built into different competing models, which may equip them with useful information to further **inform their strategic licensing decisions**. As with the case of opt-out solutions, the role of the public institution is to provide information in the promotion of transparency and awareness, and not to necessarily endorse any particular solution (especially proprietary measures provided by private undertakings on a commercial basis).

- ***Facilitating end user understanding of GenAI Terms and Conditions of Services*** - Public institutions may also serve as centralised points for end user information on the terms and conditions of major GenAI models. Specifically, end-users should be made aware of **'responsible usage guidelines'** and 'reasonable service use' provisions in GenAI systems' terms of usage, which may explicitly forbid a user from using the service to intentionally infringe on third party intellectual property rights. Such public information could also provide details and explain **indemnification clauses** that major GenAI models incorporate in their terms and conditions.

- ***Building public awareness*** - Similar to their potential role in addressing GenAI input issues, public institutions can serve help raise awareness on GenAI output issues. Public education should aim to not just build awareness of the copyright implications of GenAI output, but to also develop a sense amongst the end user base for the need to balance creative uses of GenAI systems with respect for intellectual property rights. Public institutions may wish to partner with other organisations involved in end user outreach to incorporate these copyright-specific concerns into **wider AI literacy efforts**. Beyond awareness of service terms and conditions, and potential legal consequences for infringement, efforts might be made to familiarise end users with the principles of responsible AI usage, including recommended practices for **ethical and responsible prompt engineering**. Furthermore, public institutions may have a critical role to play in public education efforts regarding the identification and interpretation of generative or manipulated output, particularly when it comes to deepfakes. This may involve educational efforts on the use of **assistive content detection tools**, and the more general principles of information literacy (and awareness of issues with deepfakes in particular) as they relate to interactions with generative content.

- ***Trend tracking*** - As with GenAI inputs, public institutions may play an important role by **reporting on market developments**. This may include reports on new models and architectures, and the unique challenges they create for addressing plagiaristic output, but also tracking new measures and protocols for addressing infringing output generally. The institutions may also track and report on trends of enforcement of intellectual property rights against plagiaristic output, to promote **increased visibility of the legal consequences** of intentionally misusing GenAI systems for infringing purposes.

- ***Technical forums for AI interoperability and content detection*** - The legal obligations under the AI Act to ensure detectability of generative output specifically reference the need for these technological solutions to be effective, interoperable, robust, and reliable. Public institutions may serve as forums to bring together AI developers and technical solution providers to promote information-sharing which is necessary to **ensure interoperability** of content detection and labelling measures. This may be an important service given the importance of an authoritative and trusted facilitator to ensure that proprietary interests are not compromised when competing AI systems share information with each other to promote interoperability of measures, and consistency in deployment across various AI systems. Public institutions could also play a role in pushing for standardised watermarking, content

attribution methods and guidelines on how GenAI output should be disclosed and labelled across jurisdictions. If requested by the industry, the public authority may also play a role as the custodian of standardised APIs used to ensure interoperability and cross-platform information sharing regarding the nature of synthetic content.

5 Conclusion

This study explores developments in GenAI from the perspective of EU copyright law. In particular, it aims to identify, explore and analyse key trends at the interface between copyright law and GenAI technologies, with specific focus on technical measures used within the AI ecosystem to address copyright management issues. The subject matter is also considered in the context of the EU legislation on AI, namely the copyright relevant obligations. This study is structured around three main components – Technical background, GenAI inputs, and GenAI outputs.

The Technical Background documents the evolution of the GenAI sector focusing on the development and deployment of key technologies and model architectures. These developments are taking place within the legal environment of EU copyright law, and in particular the provisions of the CDSM Directive *(Directive (EU) 2019/790)*, and the EU AI Act *(Regulation (EU) 2024/1689)*. The CDSM Directive provides for exceptions to the exclusive reproduction and extraction rights of copyright (and database) owners, which allows for TDM activities to take place without the authorisation of rights holders.

In the case of commercial (non-scientific research) TDM, rights holders can 'opt-out' their works from the scope of this exception, by expressing a reservation of rights that must meet specific legal criteria. The interpretation and application of these criteria may play a significant role in the strategic approach of rights holders, which in turn could influence the data acquisition processes undertaken by GenAI developers. The AI Act includes an obligation for providers of general–purpose AI (GPAI) models to put in place a policy to comply with EU copyright law, including to identify and respect the reservation of rights from the text and data mining exceptions. Additionally, providers of GenAI systems must ensure that generative output is marked in a machine-readable form and is detectable.

The development of GenAI from a copyright perspective is currently shaped by litigation between rights holders and GenAI providers in different jurisdictions. In the EU, four publicly known cases of litigation have been identified, three in Germany and one in France. The September 2024 judgement of the Hamburg Regional Court in the case *Kneschke vs. LAION* is the first legal ruling in the EU in a private dispute concerning copyright and AI training. In this case, the Court determined that LAION (a major provider of text-image datasets used for GenAI training) benefited from the TDM exception

for scientific research under Article 3 CDSM. At the same time the Court made several *obiter dicta* comments which provide potential insights into the ways the legal requirements of Article 4, TDM rights reservations, may be applied by courts in the future.

The LAION case also highlights concerns about potential 'data laundering', where the development of datasets is performed through TDM under the broad exception of scientific research. The TDM exception for scientific research is not affected by the reservations expressed by rights holders, even if such datasets are subsequently used for commercial purposes. Recent months have seen major developments in direct licensing markets where rights holders and GenAI developers entered into agreements for the use of copyright-protected works. Several direct licensing agreements have been announced publicly, though their exact contractual terms have not been disclosed. Nevertheless, analysis of the market dynamics suggests a number of potential drivers for direct licencing markets including expectations of future data drought, the added value of metadata and annotation associated to content that rights holders can provide, the relative negotiating power of contracting parties, and the emergence of content aggregation services which serve as commercial intermediaries for smaller rights holders who seek to access the AI training data market. As the market continues to develop, norms regarding pricing and contractual issues may emerge, including the framing of standard contractual terms, pricing benchmarks, and bases for remuneration.

A critical role is also played by data curators, dataset providers, and platforms supporting dataset distribution, which create a new intermediary ecosystem between rights holders and AI developers for the development and access to training datasets. However, a key challenge in this new ecosystem is the need for improved clarity and accuracy in dataset licensing terms.

Retrieval Augmented Generation (RAG) is technology gaining importance in the field of GenAI as it enhances Generative AI based services. By integrating real-time information retrieval, RAG helps contextualise users' prompts and improve both the performance and relevance of model outputs. While RAG has its own distinct copyright challenges, it represents a strategic licensing opportunity for rights holders in different sectors, starting with press, scientific, and academic publishing.

A key process for collecting training data is 'web scraping', where specific tools (called scrapers) are used to automate the mining of digital data and content from publicly available online sources. With most AI companies resorting to web scraping to gather training data for their models, many measures for managing access to copyright-protected works focus on addressing such activities. The Robots Exclusion Protocol (REP) is a *de facto* industry standard for managing web scraping, and it is widely

deployed by websites to manage access to web crawlers and scrapers, including those used by AI companies for TDM purposes. A widely acknowledged limitation of REP as a rights reservation mechanism is its inherent lack of granularity and specificity regarding permitted uses. It requires website managers to actively configure and maintain restrictions, making implementation inconsistent across different sites. Furthermore, REP is a non-binding protocol from a technical point of view, relying entirely on voluntary compliance by scrapers, which undermines its enforceability as a technical safeguard. Finally, it necessitates the public disclosure and identification of the scrapers used by different entities, as well as information on their specific purposes in case the same entities are using several crawlers.

No single opt-out mechanism has emerged as a standard. Instead, combinations of different legally-driven measures and technical measures are used by rights holders to express their TDM rights reservations. Legally-driven measures include unilateral declarations, licensing constraints, and website terms and conditions, while technical measures include various forms of metadata and content provenance protocols. These technical measures are generally characterised as either 'location-based' (applied to a specific copy of a digital asset as hosted in a particular location), or 'asset-based' (applied to the digital asset more broadly and replicated in every copy of that asset). This study compared the various reservation measures across a set of seventeen key criteria to highlight their respective advantages and limitations.

While the exact technical process of training and content generation varies depending on a model's architecture, there are significant concerns from copyright holders that some models can 'memorise' training data and subsequently generate outputs infringing their rights. In response to these concerns, as well as to the risk that end-users might intentionally use GenAI systems to infringe copyright, some model providers are implementing measures to mitigate the likelihood of generating infringing content. These measures include forms of automated input-output comparison, *ex ante* prompt filters, *ex post* output filters, as well as legal indemnification for users. Furthermore, emerging approaches such as model 'unlearning' and model editing are being tested in research and early-stage implementations. While some initial deployments exist, their scalability and effectiveness in large-scale commercial applications remain under evaluation.

To comply with the AI Act's obligations that generative output must be detectable as artificially generated or manipulated content, technical measures are used by model and system providers to support such transparency. These measures include different protocols for provenance tracking, detection of generative content, content tagging and identification solutions such as watermarking

and digital fingerprinting, and member inference attacks. This study compared various output transparency measures against ten key criteria and concluded that each measure is associated with its own respective advantages, but also limitations in the current context.

Given the complexity of the AI ecosystem, there is potential for public institutions such as intellectual property offices to provide technical and/or non-technical support. Non-technical support may take the form of public awareness initiatives, tracking key technical and commercial developments within GenAI markets, facilitating stakeholder dialogue and cooperation, and documenting the various legally-driven and technical measures used to address copyright issues related to both GenAI input and GenAI output. Technical support may take the form of solutions to address the shortcomings of the current technical market and technical developments.

6 References

Afchar, D., Meseguer-Brocal, G., & Hennequin, R. (2024). Detecting music deepfakes is easy but actually hard (No. arXiv:2405.04181). arXiv. https://doi.org/10.48550/arXiv.2405.04181

Ahmad, Z., Jaffri, Z. ul A., Chen, M., & Bao, S. (2024). Understanding GANs: Fundamentals, variants, training challenges, applications, and open problems. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-19361-y

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., & Baraniuk, R. G. (2023). Self-Consuming Generative Models Go MAD (No. arXiv:2307.01850). arXiv. https://doi.org/10.48550/arXiv.2307.01850

An, B., Ding, M., Rabbani, T., Agrawal, A., Xu, Y., Deng, C., Zhu, S., Mohamed, A., Wen, Y., Goldstein, T., & Huang, F. (2024). WAVES: Benchmarking the Robustness of Image Watermarks (No. arXiv:2401.08573). arXiv. https://doi.org/10.48550/arXiv.2401.08573

Baack, S. (2024). *A Critical Analysis of the Largest Source for Generative AI Training Data:* Common Crawl. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2199–2208. https://doi.org/10.1145/3630106.3659033

Bacon, J., Michels, J. D., Millard, C., & Singh, J. (2018). *Blockchain demystified: A technical and legal introduction to distributed and centralised ledgers*. *Richmond Journal of Law & Technology*, *15*(1).

Baio, A. (2022, August 30). *Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator.* Waxy.Org. https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/

Bertram, T., Bursztein, E., Caro, S., Chao, H., Chin Feman, R., Fleischer, P., Gustafsson, A., Hemerly, J., Hibbert, C., Invernizzi, L., Kammourieh Donnelly, L., Ketover, J., Laefer, J., Nicholas, P., Niu, Y., Obhi, H., Price, D., Strait, A., Thomas, K., & Verney, A. (2019). Five Years of the Right to be Forgotten. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security,* 959–972. https://doi.org/10.1145/3319535.3354208

Borghi, M. (2021). Exceptions as users' rights? in E. Rosati (ed.) *Routledge Handbook of EU Copyright Law.* Routledge. https://doi.org/10.4324/9781003156277

Borghi, M., & Karapapa, S. (2015). Contractual restrictions on lawful use of information: Sole-source databases protected by the back door? *European Intellectual Property Review*, 37(8).

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2020). *Machine Unlearning*. arXiv. https://doi.org/10.48550/arXiv.1912.03817

Bygrave, L. A., & Schmidt, R. (2024). Regulating Non-High-Risk AI Systems under the EU's Artificial Intelligence Act, with Special Focus on the Role of Soft Law (SSRN Scholarly Paper No. 4997886). Social Science Research Network. https://doi.org/10.2139/ssrn.4997886

Caldwell, S., Temmermans, F., & Rixhon, P. (2024). *JPEG Trust White Paper* (ISO/IEC Patent). https://ds.jpeg.org/whitepapers/jpeg-trust-whitepaper.pdf

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., & Wallace, E. (2023). Extracting Training Data from Diffusion Models (No. arXiv:2301.13188). arXiv. https://doi.org/10.48550/arXiv.2301.13188

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2023). Quantifying Memorization Across Neural Language Models (No. arXiv:2202.07646). arXiv. https://doi.org/10.48550/arXiv.2202.07646

Chiba-Okabe, H., & Su, W. J. (2024). Tackling GenAI Copyright Issues: Originality Estimation and Genericization (No. arXiv:2406.03341). arXiv. https://doi.org/10.48550/arXiv.2406.03341

Christodorescu, M., Craven, R., Feizi, S., Gong, N., Hoffmann, M., Jha, S., Jiang, Z., Kamarposhti, M. S., Mitchell, J., Newman, J., Probasco, E., Qi, Y., Shams, K., & Turek, M. (2024). *Securing the Future of GenAI: Policy and Technology* (No. 2024/855). Cryptology ePrint Archive. https://eprint.iacr.org/2024/855

Cooper, A. F., Choquette-Choo, C. A., Bogen, M., Jagielski, M., Filippova, K., Liu, K. Z., Chouldechova, A., Hayes, J., Huang, Y., Mireshghallah, N., Shumailov, I., Triantafillou, E., Kairouz, P., Mitchell, N., Liang, P., Ho, D. E., Choi, Y., Koyejo, S., Delgado, F., … Lee, K. (2024). Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice (No. arXiv:2412.06966). arXiv. https://doi.org/10.48550/arXiv.2412.06966

Crowell&Moring, Directorate-General for Communications Networks, Content and Technology (European Commission), IMC University of Applied Sciences Krems, Philippe Rixhon Associates, Technopolis Group, & UCLouvain. (2022). *Study on copyright and new technologies: Copyright data management and artificial intelligence*. Publications Office of the European Union. https://data.europa.eu/doi/10.2759/570559

Dathathri, S., See, A., Ghaisas, S., Huang, P.-S., McAdam, R., Welbl, J., Bachani, V., Kaskasoli, A., Stanforth, R., Matejovicova, T., Hayes, J., Vyas, N., Merey, M. A., Brown-Cohen, J., Bunel, R., Balle, B., Cemgil, T., Ahmed, Z., Stacpoole, K., … Kohli, P. (2024). *Scalable watermarking for identifying large language model outputs.* Nature, 634(8035), 818–823. https://doi.org/10.1038/s41586-024-08025-4

Defferrard, M., Benzi, K., Vandergheynst, P., & Bresson, X. (2017). FMA: A Dataset For Music Analysis (No. arXiv:1612.01840). arXiv. https://doi.org/10.48550/arXiv.1612.01840

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *IEEE Conference on Computer Vision and Pattern Recognition.* https://doi.org/10.1109/CVPR.2009.5206848

De Wilde, P., Arora, P., Buarque, F., Chin, Y. C., Thinyane, M., Stinckwich, S., Fournier Tombs, E., & Marwala, T. (2024). *Recommendations on the use of synthetic data to train AI models*.

Dornis, T. W., & Stober, S. (2024). Copyright Law and Generative AI Training—Technological and Legal Foundations (Urheberrecht und Training generativer KI-Modelle—Technologische und juristische Grundlagen) (SSRN Scholarly Paper No. 4946214). Social Science Research Network. https://papers.ssrn.com/abstract=4946214

Dou, G., Liu, Z., Lyu, Q., Ding, K., & Wong, E. (2025). Avoiding Copyright Infringement via Large Language Model Unlearning (No. arXiv:2406.10952). arXiv. https://doi.org/10.48550/arXiv.2406.10952

Dusollier, S. (2020). The 2019 Directive on Copyright in the Digital Single Market: Some progress, a few bad choices, and an overall failed ambition. *Common Market Law Review*, 57(4).

Eldan, R., & Russinovich, M. (2023). Who's Harry Potter? Approximate Unlearning in LLMs (No. arXiv:2310.02238). arXiv. https://doi.org/10.48550/arXiv.2310.02238

Eriksson, M. (2023). We Want Your Tools! Or Do We? On Digitized Cultural Heritage Archives And Commercial Content Identification Tools. https://viewjournal.eu/articles/10.18146/view.319

Fernandez, P., Couairon, G., Jégou, H., Douze, M., & Furon, T. (2023). The Stable Signature: Rooting Watermarks in Latent Diffusion Models (No. arXiv:2303.15435). arXiv. https://doi.org/10.48550/arXiv.2303.15435

Fritz, J. (2024). The notion of 'authorship' under EU law—who can be an author and what makes one an author? An analysis of the legislative framework and case law. *Journal of Intellectual Property Law & Practice*, *19*(7), 552–556. https://doi.org/10.1093/jiplp/jpae022

Gerken, M. (2022). Facilitating the implementation of the European Charter for Regional or Minority Languages through artificial intelligence. Council of Europe Publishing. https://edoc.coe.int/en/minority-languages/11416-facilitating-the-implementation-of-the-european-charter-for-regional-or-minority-languages-through-artificial-intelligence.html

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., Rosa, G. de, Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., & Li, Y. (2023). Textbooks Are All You Need (No. arXiv:2306.11644). arXiv. https://doi.org/10.48550/arXiv.2306.11644

Hamann, H. (2024). 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. *JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law*, *15*(2), Article 2. https://www.jipitec.eu/jipitec/article/view/407

He, L., Huang, Y., Shi, W., Xie, T., Liu, H., Wang, Y., Zettlemoyer, L., Zhang, C., Chen, D., & Henderson, P. (2024). Fantastic Copyrighted Beasts and How (Not) to Generate Them (No. arXiv:2406.14526). arXiv. https://doi.org/10.48550/arXiv.2406.14526

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. van den, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Sifre, L. (2022). Training Compute-Optimal Large Language Models (No. arXiv:2203.15556). arXiv. https://doi.org/10.48550/arXiv.2203.15556

Ioannidis, D., Kepner, J., Bowne, A., & Bryant, H. S. (2024). Are ChatGPT and Other Similar Systems the Modern Lernaean Hydras of AI? (No. arXiv:2306.09267). arXiv. https://doi.org/10.48550/arXiv.2306.09267

Ippolito, D., Tramèr, F., Nasr, M., Zhang, C., Jagielski, M., Lee, K., Choquette-Choo, C. A., & Carlini, N. (2023). Preventing Verbatim Memorization in Language Models Gives a False Sense of Privacy (No. arXiv:2210.17546). arXiv. https://doi.org/10.48550/arXiv.2210.17546

Jiang, H. H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J., & Gebru, T. (2023). AI Art and its Impact on Artists. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 363–374. https://doi.org/10.1145/3600211.3604681

Jiménez, J. & Arkko, J. (2024). AI, Robots.txt. *Internet Architecture Board (IAB) Workshop*, September 2024, Washington, DC, USA.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models (No. arXiv:2001.08361). arXiv. https://doi.org/10.48550/arXiv.2001.08361

Koster, M., Illyes, G., Zeller, H., & Sassman, L. (2022). *Robots Exclusion Protocol (Request for Comments No. RFC 9309)*. Internet Engineering Task Force. https://doi.org/10.17487/RFC9309

Kowalski, K., Volpin, C., & Zombori, Z. (2024). Competition in generative AI and virtual worlds. Publications Office. https://data.europa.eu/doi/10.2763/679899

Leistner, M., & Jussen, R. (2025). The Flattening of Creative Industries? A Closer Look at Copyright Protection of AI-Based Subject Matter (SSRN Scholarly Paper No. 5080250). Social Science Research Network. https://doi.org/10.2139/ssrn.5080250

Levendowski, A. (2018). How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem. *Washington Law Review*, 93(2), 579. https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2

Liesenfeld, A., & Dingemanse, M. (2024). Rethinking open source generative AI: Open-washing and the EU AI Act. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1774–1787. https://doi.org/10.1145/3630106.3659005

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., & Hooker, S. (2023). The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI (No. arXiv:2310.16787). arXiv. https://doi.org/10.48550/arXiv.2310.16787

Madiega, T. (2023). General-purpose artificial intelligence. EPRS | European Parliamentary Research Service.

Mehrjardi, F. Z., Latif, A. M., Zarchi, M. S., & Sheikhpour, R. (2023). A survey on deep learning-based image forgery detection. Pattern Recognition, 144, 109778. https://doi.org/10.1016/j.patcog.2023.109778

Meng, F., Yao, Z., & Zhang, M. (2025). TransMLA: Multi-Head Latent Attention Is All You Need (No. arXiv:2502.07864). arXiv. https://doi.org/10.48550/arXiv.2502.07864

Mezei, P. (2024). A Saviour or A Dead End? Reservation of Rights in The Age Of Generative AI. European *Intellectual Property Review*, 46(7).

Mitchell, E., Lin, C., Bosselut, A., Finn, C., & Manning, C. D. (2022). Fast Model Editing at Scale (No. arXiv:2110.11309). arXiv. https://doi.org/10.48550/arXiv.2110.11309

Mitchell, E., Lin, C., Bosselut, A., Manning, C. D., & Finn, C. (2022). Memory-Based Model Editing at Scale (No. arXiv:2206.06520). arXiv. https://doi.org/10.48550/arXiv.2206.06520

Mo, J., Kang, X., Hu, Z., Zhou, H., Li, T., & Gu, X. (2023). Towards Trustworthy Digital Media In The Aigc Era: An Introduction To The Upcoming IsoJpegTrust Standard. *IEEE Communications Standards Magazine*, *7*(4), 2–5. IEEE Communications Standards Magazine. https://doi.org/10.1109/MCOMSTD.2023.10353009

Naik, I., Naik, D., & Naik, N. (2023). Chat Generative Pre-Trained Transformer (ChatGPT): Comprehending its Operational Structure, AI Techniques, Working, Features and Limitations. *IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, 1–9. https://doi.org/10.1109/ICTBIG59752.2023.10456201

Novelli, C., Casolari, F., Hacker, P., Spedicato, G., & Floridi, L. (2024). Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. *Computer Law & Security Review*, *55*, 106066. https://doi.org/10.1016/j.clsr.2024.106066

OECD. (2025). *Intellectual property issues in artificial intelligence trained on scraped data* (33rd ed., OECD Artificial Intelligence Papers) [OECD Artificial Intelligence Papers]. https://doi.org/10.1787/d5241a23-en

Pagallo, U., & Ciani Sciolla Lagrange Pusterla, J. (2023). Anatomy of web data scraping: Ethics, standards, and the troubles of the law. European Journal of Privacy Law & Technologies, 2. https://doi.org/10.2139/ssrn.4707651

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., & Launay, J. (2023). The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only (No. arXiv:2306.01116). arXiv. https://doi.org/10.48550/arXiv.2306.01116

Péter, M. (2024). A Saviour or A Dead End? Reservation of Rights in The Age Of Generative AI. EUROPEAN INTELLECTUAL PROPERTY REVIEW, 46(7), Article 7. https://publicatio.bibl.u-szeged.hu/35184/

Peukert, A. (2024). *Regulating IP exclusion/inclusion on a global scale: The example of copyright vs. AI training* (SSRN Scholarly Paper No. 4905400). Social Science Research Network. https://papers.ssrn.com/abstract=4905400

Publio, G. C., Esteves, D., Ławrynowicz, A., Panov, P., Soldatova, L., Soru, T., Vanschoren, J., & Zafar, H. (2018). *ML-Schema: Exposing the Semantics of Machine Learning with Schemas and Ontologies* (No. arXiv:1807.05351). arXiv. https://doi.org/10.48550/arXiv.1807.05351

Rep. Beyer, D. S. (2023, December 22). H.R.6881 - 118th Congress (2023-2024): AI Foundation Model Transparency Act of 2023 (2023-12-22). https://www.congress.gov/bill/118th-congress/house-bill/6881

Rep. Schiff, A. B. [D-C.-28. (2024, September 4). H.R.7913 - 118th Congress (2023-2024): Generative AI Copyright Disclosure Act of 2024 (2024-04-09). https://www.congress.gov/bill/118th-congress/house-bill/7913

Rosati, E. (2021). *Copyright in the digital single market: Article-by-article commentary to the provisions of directive 2019/790*. Oxford University Press.

Rosati, E. (2024). Infringing AI: Liability for AI-Generated Outputs under International, EU, and UK Copyright Law. European Journal of Risk Regulation, 1–25. https://doi.org/10.1017/err.2024.72

Rosati, E. (2021)., Linking and Copyright in the Shade of VG Bild-Kunst. 58(6) *Common Market Law Review* 1875-1894.

Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial intelligence: A modern approach* (3rd ed). Prentice Hall.

Samuelson, P. (2021). Withholding Injunctions in Copyright Cases: The Impact of eBay (SSRN Scholarly Paper No. 3801254). Social Science Research Network. https://papers.ssrn.com/abstract=3801254

Sarkis, A. (2023). Training Data for Machine Learning. O'Reilly Media, Inc.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., & Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models (No. arXiv:2210.08402). arXiv. https://doi.org/10.48550/arXiv.2210.08402

Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., & Zhao, B. Y. (2023). Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models (No. arXiv:2302.04222). arXiv. https://doi.org/10.48550/arXiv.2302.04222

Shan, S., Ding, W., Passananti, J., Wu, S., Zheng, H., & Zhao, B. Y. (2024). Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. 807–825. https://doi.org/10.1109/SP54263.2024.00207

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2024). The Curse of Recursion: Training on Generated Data Makes Models Forget (No. arXiv:2305.17493). arXiv. https://doi.org/10.48550/arXiv.2305.17493

Sinitsin, A., Plokhotnyuk, V., Pyrkin, D., Popov, S., & Babenko, A. (2020). Editable Neural Networks (No. arXiv:2004.00345). arXiv. https://doi.org/10.48550/arXiv.2004.00345

Stieper, M., & Denga, M. (with Universitäts- Und Landesbibliothek Sachsen-Anhalt & Martin-Luther Universität). (2024). The international reach of EU copyright through the AI Act. Institut für Wirtschaftsrecht. https://doi.org/10.25673/116949

Strowel, A., & Ducato, R. (2021). Artificial intelligence and text and data mining: A copyright carol. In E. Rosati (ed.) *The Routledge Handbook of EU Copyright Law*. Routledge. https://doi.org/10.4324/9781003156277-19

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA:

Open and Efficient Foundation Language Models (No. arXiv:2302.13971). arXiv. https://doi.org/10.48550/arXiv.2302.13971

Turing A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data (No. arXiv:2211.04325). arXiv. https://doi.org/10.48550/arXiv.2211.04325

Wang, J. T., Deng, Z., Chiba-Okabe, H., Barak, B., & Su, W. J. (2024). An Economic Solution to Copyright Challenges of Generative AI (No. arXiv:2404.13964). arXiv. http://arxiv.org/abs/2404.13964

Wang, Y. C., Xue, J., Wei, C., & Kuo, C.-C. J. (2023). An Overview on Generative AI at Scale With Edge–Cloud Computing. *IEEE Open Journal of the Communications Society*, *4*, 2952–2971. https://doi.org/10.1109/OJCOMS.2023.3320646

Weizenbaum, J. (1976). *Computer Power and Human Reason. From Judgement to Calculation*. W.H. Freeman & Company.

World Intellectual Property Organization (WIPO) (2024). Generative Artificial Intellligence. Patent Landscape Report. Geneva: WIPO. https://doi.org/10.34667/tind.49740

Yao, Y., Wang, P., Tian, B., Cheng, S., Li, Z., Deng, S., Chen, H., & Zhang, N. (2023). Editing Large Language Models: Problems, Methods, and Opportunities (No. arXiv:2305.13172). arXiv. https://doi.org/10.48550/arXiv.2305.13172

Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting (No. arXiv:1709.01604). arXiv. https://doi.org/10.48550/arXiv.1709.01604

Zhang, D., Finckenberg-Broman, P., Hoang, T., Pan, S., Xing, Z., Staples, M., & Xu, X. (2024). Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions (No. arXiv:2307.03941). arXiv. https://doi.org/10.48550/arXiv.2307.03941

Zhang, D., Xia, B., Liu, Y., Xu, X., Hoang, T., Xing, Z., Staples, M., Lu, Q., & Zhu, L. (2024). Privacy and Copyright Protection in Generative AI: A Lifecycle Perspective. *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, 92–97. https://doi.org/10.1145/3644815.3644952

Zhao, Z., Duan, J., Xu, K., Wang, C., Zhang, R., Du, Z., Guo, Q., & Hu, X. (2024). Can Protective Perturbation Safeguard Personal Data from Being Exploited by Stable Diffusion? *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24398–24407. https://doi.org/10.1109/CVPR52733.2024.02303

7 Glossary

7.1 Abbreviations

AI – Artificial Intelligence

CA – Certification Authority

CDSM – Copyright in the Digital Single Market (Directive)

CJEU – Court of Justice of the European Union

CMO – Collective Management Organisation

GAN – Generative Adversarial Network

GPAI – General Purpose AI Model

GPT – Generative Pre-trained Transformer

GPU – Graphics Processing Unit

IPTC – International Press Telecommunications Council

ISCC – International Standard Content Code

LLM – Large Language Model

RAG – Retrieval Augmented Generation

REP – Robots Exclusion Protocol

TDM – Text and Data Mining

ToS – Terms of Service

TPM – Technological Protection Measures

UGC – User Generated Content

VAE – Variational Autoencoders

7.2 Concepts

Artificial Neural Network (ANN)	a network of interconnected algorithms designed to process information based on external inputs. The term originates from the design of the network, which draws inspiration from the structure of biological neural networks in the brain.
Attention Mechanism	in Generative AI, is a mechanism that enables models to focus on the most relevant parts of input data when making predictions. It works by comparing every token in a sequence with each other to calculate a score (attention weight) that indicates their relative importance, improving context understanding and coherence, especially in sequence-based tasks like natural language processing (e.g., Transformers and self-attention in GPT).
Contrastive Language-Image Pretraining (CLIP)	a machine learning model designed to establish relationships between images and text. It processes visual and textual data to associate images with corresponding descriptions and vice versa. The model can evaluate similarities within images by transforming them into text representations and computing cosine similarity between the resulting text embeddings. Additionally, it can assess the similarity between an image and a given text description.
Convolutional Neural Network (CNN)	a type of AI model architecture based on a network of software components (called “neurons”) organised into layers. Each neuron is like a mathematical function, taking more inputs and producing one output based on the value of the model’s parameters. Each layer receives as input the output of the neurons of the previous layers and combines these during its elaboration. CNNs use convolutional layers that apply "filters" to an input image, identifying patterns like edges, textures, or shapes, which are then combined across layers to extract hierarchical features of the image. CNNs are used extensively in the context of image generation.
Data Hash	strings generated using a cryptographic algorithm, which takes another string as input. The cryptographic link between the input and output strings is crucial for detecting tampering with the original data. For this reason, data hashes (i.e., the output of the cryptographic algorithm) are always present alongside the data they protect (i.e., the input string), ensuring integrity verification.
Diffusion model	generative model capable of producing realistic images. This is achieved by instructing Diffusion Models to start with random noise and gradually transforming it into a clear image.
Digital Signature	mathematical scheme used to verify the authenticity and integrity of a digital message or document. It is generated using a private key and can be verified by anyone with the corresponding public key. This ensures that the message was created by the legitimate sender (authentication) and has not been altered (integrity).

Fine-tuning	the process of taking a pre-trained general-purpose model and further training it on a specific, smaller dataset tailored to a particular task or domain. This approach allows the model to adapt its base knowledge to meet the needs of a specific application, improving performance while requiring less computational effort compared to training a model from scratch. In case of GenAI the starting model for fine-tuning is a foundation model, while in non-GenAI it can be, for example, a pre-trained Vision Transformer (ViT). The latter is pre-trained on massive image datasets and can be fine-tuned to specific image recognition tasks.
Fingerprinting	the process of associating unique identifiers with content to trace its origin, verify authenticity, or prevent misuse. These identifiers are derived from peculiar characteristics of the referred media and stored externally from it. This technique enhances accountability and helps detect unauthorised use of AI outputs. The key difference between fingerprinting and watermarking is that fingerprinting calculates a unique identifier for the original and, on demand, for copies of the asset, whereas watermarking typically involves embedding a (unique) identifier in the original which is then also present in all copies of the asset.
Foundation models	large-scale machine learning models trained on vast amounts of diverse data that can be adapted (fine-tuned) for a wide range of tasks. Due to their dimensions, training a foundation model is very expensive. For that reason, they often serve as a starting point for building more specialised generative systems, eliminating the need to retrain a new entire model from scratch each time. Indeed, the machine learning phase concerning the creation of a foundation model is referred to as 'pre-training'. Subsequent fine-tuning is necessary for making the outputs more coherent and aligned with human preferences. Further details can be found in *Section 2.1.2.2*.
Generative AI	the subset of the Artificial Intelligence technologies focusing on data generation. It leverages machine learning techniques to extract statistical features from a large amount of data. Those learned patterns are then combined to generate new content conditioned by the user's request, also with a certain degree of randomness.
Generative AI Input	the part of the generative AI development process which uses input data such as during training and fine-tuning activities. The details of those activities vary depending on the generative model's type.
Generative AI Model	mathematical framework designed to perform specific tasks, such as recognising patterns or generating text. It processes input data, applies learned parameters, and produces output with different formats (i.e., text, visual, audio or even multimodal content) depending on the model's architecture. It differentiates from other AI models in the capacity of generating new, unseen, output. A working generative model is often obtained by fine-tuning a large, generic, Foundation Model (FM). A non-exhaustive list of currently popular models is available in *Annex IV*.
Generative AI output	the part of the generative AI development process which produces output data. Data generation is conducted during the model's utilisation by the end user and includes elaboration of the input prompt as well as the generation process itself. The latter may vary depending on the generative

	model's type. External data retrieval at inference time (RAG technologies) is related to this part.
General Purpose AI (GPAI)	generative AI technologies capable of generalising among a vast number of tasks without having been trained specifically for these. At the core of their functioning, they have Foundation Models (FMs). An example is ChatGPT, a chatbot capable of making translations, summaries, answering user's questions and more.
Image Features	distinctive visual elements or details, such as shapes, colours, or textures which GenAI models extract from images during training. These features are then coded into the models' parameters, which store the information about the statistical distribution of the training data's features. This statistical description of the training images serves to generate new, unseen, visual content having features which fit into the learned distributions.
Labelled data	in machine learning, it is a pair composed by the content and the tags (or labels) used to categorise the content itself. Labelled data is used in supervised machine learning, where the model under training learns to assign, for each item of the predefined set of labels, the probability that a content can be flagged with it.
Latent space	in machine learning, it is a condensed representation of data that retains only the key features defining its underlying structure. Effectively modelling latent space is crucial for the development of Generative AI systems.
Machine learning	technique used to train models that enable systems to learn patterns and make predictions or decisions without being explicitly programmed. Machine learning is referred to as Deep learning when the training process utilises artificial neural networks (ANNs).
Media Tokenisation	the process of converting media rights into digital tokens on a blockchain, potentially enabling new transparent rights management systems. (Note: media tokenisation is a different concept than text tokenisation; the latter is described in *Section 3.1.3.1*).
Natural Language Processing (NLP)	branch of artificial intelligence focused on enabling computers to understand, interpret, generate, and respond to human language in a meaningful way.
Objective Function	mathematical function depending on many parameters which have to be tuned in order to maximise/minimise the function's value. The objective function's optimisation is at the base of the machine learning processes.
Overfitting	a model learns patterns specific to its training data too well, resulting in poor generalisation to unseen data. This means the model produces outputs that are overly tailored to the training set, limiting its ability to perform effectively in real-world scenarios.
Parameter	refers to a numerical value within a model that the system adjusts during training to learn patterns from data. Parameters, such as weights in a neural network, define how the input data is transformed into an output. The more parameters a model has, the greater its capacity to learn

	complex relationships, though it also requires more computational resources.
Reinforcement Learning (RL)	machine learning approach based on the goal of maximising an objective function through trial and error. In GenAI, RL, especially Reinforcement Learning from Human Feedback (RLHF), is used to refine models for better output quality.
Retrieval-Augmented Generation (RAG)	combines retrieving relevant information from external sources (like databases, documents, or APIs) with generating responses using a model, thereby improving accuracy and contextual relevance of the output. RAG can support various GenAI applications that generate real-time content, for example by connecting a Foundation model directly to social media feeds, news platforms, or other frequently updated information sources.
Reward Model	system used in reinforcement learning to quantify the desirability of an action (i.e., a single piece of the generative model's output) by assigning a numerical reward. It guides the trained model's learning process by indicating which behaviours lead to better outcomes, enabling it to optimise for maximum cumulative rewards.
Self-attention Mechanism	see Attention Mechanism.
Supervised Learning	Machine Learning (ML) technique where models learn from labelled data by minimising prediction errors.
Text Tokenisation	the process of breaking text into smaller units, like words, subwords, or characters, that the model can process. It is essential for processing and generating text.
Token	the smallest unit of data that a model processes. In texts, a token could be a single word, sub word, or character
Training	in AI, it is the process of performing Machine Learning to develop a model.
Transformer	neural network architecture designed for sequence data, such as text. It is not a general model itself, but it can constitute one of the principal components of the LLM's architecture. It uses self-attention mechanisms (see the definition above) to process all input tokens simultaneously, capturing relationships between them, regardless of their distance.
Watermarking	technique to embed information (often content metadata) in the content itself. It can be visible or invisible, with a different degree of content's modification. An evident example is overlying an image with the author's logo.
Web crawling	the process of systematically exploring the web by starting with a set of seed URLs and following hyperlinks to discover new pages, often with the goal of building an index of the web. Crawlers prioritise discovering and indexing as many pages as possible rather than extracting specific data.
Web scraping	the process of extracting specific data from websites by retrieving their content and parsing it to obtain the desired information, such as text, images, or structured data. It focuses on targeted data extraction from one

or more predefined pages and is often used to collect data for analysis or integration into other systems.

8 Annexes

Annex I: Research Team

University of Turin

Prof. Maurizio Borghi	*Project Lead*
Dr Bryan Khan	*Senior Researcher*
Ms Anna Arnaudo	*Researcher*
Mr Riccardo Raso	*Researcher*
Prof. Marco Ricolfi	*Scientific Advisor*
Prof. Antonio Vetrò	*Scientific Advisor*
Prof. Riccardo Coppola	*Scientific Advisor*

European Union Intellectual Property Office Observatory

Mr Antoine Aubert	*EUIPO Co-Project Lead*
Mr Ziga Drobnic	*EUIPO Co-Project Lead*
Mr Stephan Edelbroich	*EUIPO Technical Expert*
Ms Chikemka Abuchi-Ogbonda	*EUIPO Trainee*
Mr Raffaele Darroch	*EUIPO Trainee*

Annex II: Stakeholder Interviewees

This annex outlines the stakeholder classification framework defined prior to conducting the semi-structured interviews within this research project. The categorisation aims to systematically capture the diverse roles stakeholders hold within the AI and copyright ecosystems, facilitating structured analysis and interpretation.

During initial scoping, it became clear that standard classifications would not fully encapsulate institutions involved in cultural heritage digitisation. This prompted the creation of an additional category (**C3 – Cultural Heritage Institutions**), distinct from traditional content providers (**B0**). Institutions in **C3** manage extensive digital collections predominantly for preservation, research, and public access rather than commercial exploitation. Unlike typical content providers (e.g., news publishers, music rights organisations), these entities neither monetise nor directly commercialise their collections. However, they remain significantly exposed to AI-related practices (such as large-scale data scraping), engage actively in copyright and AI policy discourse, and implement AI-driven digitisation techniques, including OCR and metadata extraction.

The table below summarises these pre-defined categories and indicates the number of stakeholder interviews subsequently conducted for each group.

Category Code	Description	Counter (No. of Interviews)
A1: AI Tools	Organisations providing software solutions, tools, and frameworks supporting AI applications without directly developing foundational models.	7

A2: AI Developers/Makers	Entities actively designing, training, and deploying AI models, including large language and multimodal systems.	6
B0: Content Providers	Organisations primarily involved in creating, managing, or commercially distributing creative content (e.g., publishers, media companies, authors' societies).	12
C1: Civil Society	Non-governmental organisations, advocacy groups, and think tanks addressing ethical, social, and governance challenges related to AI technologies.	2
C2: Government & Policy Institutions	Public bodies, regulators, and governmental institutions responsible for AI-related policy, copyright enforcement, and intellectual property regulation.	2
C3: Cultural Heritage Institutions	Institutions managing digitised cultural heritage and public-domain archives, focusing primarily on preservation, research, and public access rather than commercial purposes.	1

Annex III: Stakeholder Interview Templates

This annex presents the structured templates employed as flexible frameworks guiding the semi-structured interviews conducted within this research. The templates were designed to ensure comprehensive coverage of relevant themes, while allowing respondents the freedom to elaborate on topics most pertinent to their role and experience.

III.1 TEMPLATE A: CONTENT INDUSTRIES

This template applies to the copyright sector and rights holders' organisations.

1. Interviewee and Organisation Background

1.1 Interviewee Background

- Please state the name of the organisation and briefly describe its core business.
- What is your role, and how are you involved in the work on copyright and GenAI?

1.2 Organisation Background

- How would you describe your organisation's position within the GenAI Value/Data Chain?
- How is your content being used/licensed as part of GenAI training or output?
- Are you using/developing GenAI tools as part of your activities?
- What potential services along the GenAI Value/Data Chain (either upstream or downstream) would increase the economic viability of your organisation's business strategy?
- Is your organisation currently engaged in (or pursuing) any contracts or data-sharing practices with other stakeholders in the GenAI Value/Data Chain?

2. Input-Management Measures [X1]

- Does your organisation have any specific (public) position regarding copyright management and the Text and Data Mining (TDM) exception covered by Art. 4 CDSM?

- Is your organisation currently involved in internal discussions on developing a position on copyright management and TDM?
- Has your organisation developed, introduced, or experimented with mechanisms or technical solutions to express the opt-out for TDM?
- *(For associations/CMOs only):* Has your organisation delivered instructions/recommendations to its membership regarding copyright management and TDM?
- What was the decision-making process for adopting or dismissing particular measures?
- Does your organisation have monitoring processes to ensure implemented measures or solutions are effective for GenAI uses?
- Does your organisation see these measures/solutions potentially becoming a standard across your or other content sectors?
- Is your organisation involved in any discussions or standardisation forums related to developing TDM opt-out mechanisms or technical standards?

3. Infringing Output Measures [X2]

- Has your organisation experienced GenAI output clearly infringing its copyright or related rights?
- Does your organisation have specific positions or guidelines on mitigating potential copyright infringement risks from GenAI output?
- Does your organisation support or use specific technical solutions/good practices to reduce or eliminate GenAI-generated infringing content?

4. Output Labelling Measures [X3]

- Does your organisation have specific positions regarding technical solutions (e.g., watermarking) or standards for labelling GenAI output?
- Does your organisation use any such standards or solutions?

III.2 TEMPLATE B: SOLUTION PROVIDERS (NON-DEVELOPERS)

This template applies to organisations developing TDM opt-out or GenAI output labelling solutions.

1. Interviewee and Organisation Background

1.1 Interviewee Background

- Please state the name of your organisation and briefly describe its core business.
- What is your role, and how are you involved in copyright and GenAI?

1.2 Organisation Background

- How would you describe your organisation's position within the GenAI Value/Data Chain?
- What services/inputs does it use from other stakeholder groups, and who are its clients/users?
- What services or standards along the GenAI Value/Data Chain is your organisation developing?
- Is your organisation currently supported by/cooperating with other stakeholders in the GenAI Value/Data Chain?

2. Input-Management Measures [X1]

- Briefly describe the TDM opt-out solution(s) your organisation is developing.
- What measures has your organisation taken to adapt its TDM opt-out solutions to copyright holders or AI developers?
- *(For each solution):*
    - Describe the decision-making process for development and implementation.
    - At what stage of content creation or distribution is the TDM opt-out solution implemented?
    - Do you foresee current or future adoption across content sectors or AI developers?
    - What preliminary results exist regarding its effectiveness?
    - Describe advantages and limitations of your solution (feasibility, maturity, implementation costs, accuracy, upgradability).

3. Infringing Output Measures [X2]

- Does your organisation have specific positions regarding the potential infringement of rights by GenAI output?

- Does your organisation develop technical solutions/good practices to reduce the risk of infringing GenAI output?

4. Output Labelling Measures [X3]

- Does your organisation develop technical solutions (e.g., watermarking) or standards for labelling GenAI output?
- Do you foresee current or future adoption of these solutions by GenAI developers?
- Describe advantages and limitations of your solution for identifying GenAI output (feasibility, maturity, implementation costs, accuracy, upgradability).

5. EUIPO & Authority Promotion [Closing-Set]

- Do you see a role for EUIPO or other relevant authorities in supporting/promoting your system?
    - If yes, what specific contributions could they make?

III.3 TEMPLATE C: AI DEVELOPERS

This template applies to AI model/system developers and dataset providers.

1. Interviewee and Organisation Background

1.1 Interviewee Background

- Please state the name of your organisation and briefly describe its core business.
- What is your role, and how are you involved in copyright and GenAI?

1.2 Organisation Background

- Describe your organisation's position within the GenAI Value/Data Chain.
- What services/inputs are you using from other stakeholders, and who are your clients/users?

- What potential services along the GenAI Value/Data Chain would enhance your business strategy?
- Is your organisation engaged in data-sharing contracts/practices with other GenAI stakeholders?

2. Input-Management Measures [X1]

- Does your organisation undertake in-house TDM activities? If yes, state the TDM process type used.

2.1 In-House TDM (if applicable)

- What measures have you implemented for TDM opt-out compliance?
- *(For each measure):*
  - Decision-making process
  - Stage of TDM implementation
  - Potential adoption as industry standard
  - Preliminary effectiveness and over-filtering results
  - Advantages and limitations (feasibility, maturity, costs, accuracy, upgradability)

2.2 Third-Party AI Training Data (if applicable)

- Mechanism for receiving/sharing third-party data
- Licensing terms, including liability waivers
- Policies/checks for copyright compliance

2.3 Providing AI Training Datasets (if applicable)

- Mechanism for dataset availability
- Licensing terms, including liability waivers
- Policies/checks ensuring copyright compliance

2.4 AI Training Process

- Briefly describe primary AI training methodologies used.
- Do you incorporate unlearning techniques?

3. Infringing Output Measures [X2]

- Specific position regarding potential rights infringement by GenAI output
- Mechanisms considered/used for mitigation
- Liability provisions in your service terms and conditions

4. Output Labelling Measures [X3]

- Mechanisms used/considered for identifying/marking GenAI output
- Potential for these mechanisms as industry standards
- Advantages and limitations (feasibility, maturity, costs, accuracy, upgradability)

III.4 TEMPLATE D: INSTITUTIONAL & CIVIL SOCIETY ORGANISATIONS

This template applies to civil society groups, advocacy organisations, and public authorities.

1. Interviewee and Organisation Background

1.1 Interviewee Background

- Briefly describe your organisation's mission.
- Your role and involvement in copyright and GenAI

1.2 Organisation Background

- Describe your organisation's position within the GenAI Value/Data Chain.

2. Input-Management Measures [X1]

- Specific positions regarding copyright management/TDM
- Specific positions regarding public licences (e.g., open source)
- Promotion/development of TDM opt-out solutions
- Analysis of existing opt-out mechanisms (advantages, limitations, feasibility, costs, accuracy)

3. Infringing Output Measures [X2]

- Specific positions regarding potential rights infringement by GenAI output

4. Output Labelling Measures [X3]

- Specific positions regarding GenAI output labelling standards/watermarking

5. Role of the Organisation [X4]

- Existing or potential role in raising awareness or informing about opt-out mechanisms or GenAI output identification
- Role supporting compliance of GenAI solutions with EU copyright rules

Annex IV: Non-exhaustive list of Generative models and their employments

Model	Company	Type	Uses (examples)
GPT	OpenAI	Large Language Model	Products for the public: • ChatGPT: text generation Products for enterprises: • Zoom AI Companion: workplace efficiency
Gemini	Google DeepMind	Large Language Model	Products for the public: • Gmail AI: writing assistance Products for enterprises: • My Volkswagen App: virtual assistant ([432])
DALL-E2	OpenAI	Diffusion model	Products for the public: • ChatGPT: text-to-image generation Products for enterprises: • Microsoft Designer: streamline content creation
IMAGEN	Google	Diffusion model	Products for the public: • Google Gemini: text-to-image generation
Gen-2	RunwayML	Diffusion model	Text-to-video generation
Codex	OpenAI	Large Language Model	Products for the public: • GitHub Copilot: AI coding assistant Products for enterprises:

([432]) Volkswagen Integrates AI into the myVW Mobile App with Google Cloud, VW US Media Site, 24 September 2024 (accessed 6 November 2024).

			• Microsoft Copilot for Developers: enterprise software development support
Stable Diffusion	Stability AI	Diffusion model	Text-to-image generation
Claude	Anthropic	Large Language Model	Products for public: • Anthropic's Claude AI: general purpose conversations; Products for enterprises: • Claude Assist: customer service;
DeepSeek-R1	DeepSeek	Large Language Model	Products for the public: • AI-powered text generation and assistance.
Sora	OpenAI	Diffusion model	Products for the public: • Video generation from text. Products for enterprises: • AI-powered advertising and cinematic content creation.
Midjourney	Midjourney	Diffusion model	Products for the public: • AI-generated art and photography Products for enterprises: • Used in game development, advertising, and design studios for concept art
LLaMA	Meta	Large Language Model	Products for the public: • Available via open-source implementations like LLaMA.cpp, Ollama, and third-party chatbots Products for enterprises: • Used in Meta AI's internal projects

Mistral	Mistral AI	Large Language Model	Products for the public: ● LeChat: French AI chatbot Products for enterprises: ● AI-powered customer service platforms and business intelligence tools
VALL-E	Microsoft	Text-to-Speech model	Products for the public: ● Microsoft Edge Read Aloud: realistic voice reading Products for enterprises: ● Azure AI Speech: enterprise voice synthesis, automated dubbing, virtual assistants

Table IV-1: Non-exhaustive list of Generative models and their current applications.

Annex V: OSI Open-Source Definition

Training Data Disclosure and Definition of Open-Source AI

The Open Source Initiative (OSI) is an international non-profit standards organisation who serves as the 'steward' for the definition and standards of 'open source software'. After substantial community engagement, OSI published the 'Open Source AI Definition V1.0', which is based in the long established principles of 'Free Software' (developed by the Free Software Foundation), including the freedom to 'Study how the system works and inspect its components' and 'Modify the system for any purpose, including to change its output' ([433]).

The Definition sets out that the 'preferred form of making modifications to a machine-learning system must include' disclosure of data information, code, and system parameters. With respect to 'data information', this includes:

"Sufficiently detailed information about the data used to train the system so that a skilled person can build a substantially equivalent system. Data Information shall be made available under OSI-approved terms. In particular, this must include: (1) the complete description of all data used for training, including (if used) of unshareable data, disclosing the provenance of the data, its scope and characteristics, how the data was obtained and selected, the labeling procedures, and data processing and filtering methodologies; (2) a listing of all publicly available training data and where to obtain it; and (3) a listing of all training data obtainable from third parties and where to obtain it, including for fee."

The OSI notes that the decision to exclude certain data from the definition of Open Source AI is necessary for a variety of reasons, including differences in laws between jurisdictions, sector-specific concerns (e.g., medical data), privacy, protection of indigenous knowledge, and definitions of public domain ([434]). For the purpose of this definition, the OSI suggests that training data can be categorised into four classes of

([433]) The Open Source AI Definition – 1.0, Open Source Initiative (accessed 14 March 2025).

([434]) FAQ, Smafulli, 29 October 2024 (accessed 14 March 2025).

data, based on their legal constraints, all of which can be used to train Open-Source AI systems:

- Open training data: data that can be copied, preserved, modified and reshared. It provides the best way to enable users to study the system. This must be shared.

- Public training data: data that others can inspect as long as it remains available. This also enables users to study the work. However, this data can degrade as links or references are lost or removed from network availability. To obviate this, different communities will have to work together to define standards, procedures, tools and governance models to overcome this risk, and Data Information is required in case the data becomes later unavailable. This must be disclosed with full details on where to obtain it.

- Obtainable training data: data that can be obtained, including for a fee. This information provides transparency and is similar to a purchasable component in an open hardware system. The Data Information provides a means of understanding this data other than obtaining or purchasing it. This is an area that is likely to change rapidly and will need careful monitoring to protect Open Source AI developers. This must be disclosed with full details on where to obtain it.

- Unshareable non-public training data: data that cannot be shared for explainable reasons, like Personally Identifiable Information (PII). For this class of data, the ability to study some of the system's biases demands a detailed description of the data – what it is, how it was collected, its characteristics, and so on – so that users can understand the biases and categorisation underlying the system. This must be revealed in detail so that, for example, a hospital can create a dataset with identical structure using their own patient data.

Annex VI: Ongoing USA Case Law

PARTIES	DATE FILED	RIGHTS HOLDERS / CONTENT CATEGORY
Advance Local Media v. Cohere	February 2025	News Publishers / Text
Alcon Entertainment v. Tesla, Elon Musk, Warner Bros. Discovery	October 2024	Film/TV Production / Image
Dow Jones & Company and NYP Holdings v. Perplexity AI	October 2024	News Publisher / Text
Christopher Farnsworth v. Meta Platforms	October 2024	Book Authors / Text
Millette v. NVIDIA	August 2024	Social Media Creator / Video
Bartz et al v. Anthropic	August 2024	Book Authors / Text
Millette v. OpenAI	August 2024	Social Media Creator / Text
Millette v. Google	August 2024	Social Media Creator / Video
Vacker v. Eleven Labs	August 2024	Actors and Book Publishers / Audiobooks
The Center for Investigative Reporting v. OpenAI	June 2024	News Publisher / Text
UMG Recordings et al v. Suno	June 2024	Record Label / Music
UMG Recordings et al v. Uncharted Labs	June 2024	Visual Artists / Music
Dubus v. NVIDIA	May 2024	Book Authors / Text
Daily News v. Microsoft	April 2024	News Publisher / Text
Zhang v. Google	April 2024	Visual Artists / Image
Nazemian v. NVIDIA	March 2024	Book Authors / Text
O'Nan v. Databricks	March 2024	Book Authors / Text
The Intercept Media v. OpenAI	February 2024	News Publisher / Text

Raw Story Media v. OpenAI	February 2024	News Publisher / Text
Huckabee v. Bloomberg	January 2024	Book Authors / Text
The New York Times Company v. Microsoft	December 2023	News Publisher / Text
Concord Music Group v. Anthropic	October 2023	Record Label / Text
Authors Guild et al v. OpenAI	September 2023	Book Authors / Text
Kadrey v. Meta Platforms	July 2023	Book Authors / Text
Tremblay et al. v. OpenAI	June 2023	Book Authors / Text
Getty Images v. Stability AI	February 2023	Stock Image Company / Image
Andersen v. Stability AI	January 2023	Visual Artists / Image

Table VI-1: Summary of major copyright and GenAI disputes in the USA ([435])

([435]) Content Owner Lawsuits Against AI Companies: Complete Updated Index (Paywall), Variety, VIP+ Variety Intelligence Platform, 12 August 2025 (accessed 14 March 2025).

Annex VII: Top cited datasets in the GenAI literature

Dataset name	Citing documents	Total mentions	Main modality
ImageNet	2741	6823	image
MNIST	2533	9292	image
CIFAR-10	2160	7744	image
CelebA	1705	5713	image
COCO	1141	3390	image
Wikipedia	662	2599	text
FFHQ dataset	596	1983	image
FASHIONMNIST	520	1375	image
CelebAHQ	474	1081	image
SVHN	398	1414	image
PubMed	393	1144	text
GSM8K	350	1704	text
CIFAR-100	338	849	image
HumanEval	322	1269	text
LAION	314	731	image

CUB dataset	312	1118	image
LSUN	310	608	image
CommonCrawl	290	546	text
Cityscapes	272	974	image
MMLU	270	746	text

Table VII-1: Top cited datasets in the GenAI literature **([436]).**

([436]) Patent Landscape Report - Generative Artificial Intelligence (GenAI), WIPO, 2024 (accessed 14 March 2025).

Annex VIII: Complementary Information on Data Pre-Processing

VIII.1 Example: How Meta Performed Training Data Curation for Movie Gen

Movie Gen ([437]) is capable of both generating **video and the associated audio**. Meta has stated (2024) ([438]) that the technology is **not ready for public release** due to high costs and long generation times. However, Meta shared some research data ([439]). The model's training is based on an **elaborated input data curation** procedure involving text, image, video and audio content. Each media type has a different training workflow and thus a different pre-processing strategy.

- **Visual Data:** the module responsible for generating visual content leverages joint text-to-image and text-to-video training.

  The **pre-training** data ([440]) curation workflow consists of **several filtering steps and one captioning step**. The filtering includes:

  - **Visual filtering**: selects videos based on their resolution; moreover, a video OCR model is used to remove videos with excessive text;
  - **Motion filtering**: it serves to exclude too static videos;
  - **Content filtering**: it includes the removal of similar videos and resampling to reduce the prevalence of too frequent concepts.

  Captions for videos are crafted by using the LLaMa3-Video ([441]) model.

([437]) Movie Gen, from Meta, is a cast of foundation models that generates high-quality videos with synchronised audio, thus performing text-to-video synthesis, video-to-audio generation and text-to-audio generation. It also includes additional capabilities such as precise instruction-based video editing and generation of personalised videos based on a user's image.

([438]) Meta Shows Off Its "Industry-Leading" AI Video Generator Called Movie Gen, PetaPixel, 4 October 2024 (accessed 14 March 2025).

([439]) How Meta Movie Gen Could Usher in a New AI-Enabled Era for Content Creators, Meta AI (accessed 12 December 2024).

([440]) Pre-training is the training phase aiming to build the large foundation model (see *Section 2.1.2.1* for the definition of FM). Hence, data for pre-training is prepared to expose the model to a vast and diverse range of information.

([441]) LLaMa3-Video integrates large language models with vision and audio processing to understand videos. It captures temporal dynamics, merges audio-visual signals, and uses extensive context windows to generate insights for tasks like video captioning and scene understanding.

To ensure a high output quality, the curation of the **fine-tuning** dataset follows more strict policies and involves **manual** filtering and captioning.

- **Audio**: in this case, **pre-training** aims to learn the structure of audio and alignment between audio and video/text from large quantities of data, thus the training for audio generation was performed using videos. In addition to those described in the previous point, the following steps are performed:
  - The AED model ([442]) is used to **tag audio events** based on the Audioset ontology ([443]), that has 527 classes. This allows to filter-out videos where the silence is the dominant class and to detect the presence of music and/or voice;
  - Audio quality is established by an **audio quality prediction model**;
  - A **music caption model** is deployed to add more details to labels, such as mood and genre.

In addition to the data processing procedures described above, the **input prompts** provided to the model to guide its output also require pre-processing. Specifically, Movie Gen's prompt elaboration module performs **word replacement** to simplify the input sentence, making it easier to map to the labels assigned during training.

([442]) The Acoustic Event Detection (AED) model is a system used to identify and tag audio content by detecting specific sound events within a recording. Advanced AED models often leverage deep learning techniques, such as Convolutional Neural Networks (CNNs) or Transformer architectures.

([443]) The AudioSet Ontology is a hierarchical structure that organizes a comprehensive set of meaningful labels to categorise sound events. It enables machine learning models to effectively analyse and tag audio content.

VIII.2 Image Data Conversion Techniques

While many recent image generation models adopt approaches analogous to tokenisation, like patch-based tokenisation or vector quantisation**,** to align with transformer-based architectures, others use alternative strategies or avoid tokenisation altogether.

Tokenisation-like Techniques:

- **Patch-based Representations:** division of the images into non-overlapping patches. Each patch, typically of fixed size (e.g., 16x16 pixels), is treated as an individual "token". This technique is used by **Vision Transformers (ViTs)** because their functioning strongly resembles the one of the text Transformers;
- **Grid-Based Representations:** each grid cell, which represents a localised portion of the image, serves as a "token" in subsequent processing. This approach is often used in models that **combine Convolutional Neural Networks (CNNs)** ([444]) **with attention mechanisms** ([445]), such as hybrid transformer architectures;
- **Vector Quantisation:** encoding image patches or features using a finite set of learnable embeddings, often referred to as a "codebook." Each image region is mapped to the nearest codebook vector, effectively discretising the image into a sequence of "tokens", bridging the gap between continuous visual data and discrete representations. For example, the visual input data of a **Vector Quantized Variational Autoencoder (VQ-VAE)** ([446]) is processed in this way;
- **Frequency Domain Representations**: techniques such as the Discrete Cosine Transform (DCT) or Fourier Transform are used to decompose images into components corresponding to different spatial frequencies ([447]). These components serve as tokens that encode information about the image's texture, edges, and overall structure. Frequency-based

([444]) For more details, see the *Glossary*.

([445]) For more details, see the *Glossary*.

([446]) VQ-VAE is a generative model that learns to encode data (e.g., images) into discrete latent codes using a finite set of embeddings, bridging continuous input data with transformer-like architectures. The difference with the VAE (defined in *Section 2.1.1*) relies on the quantization of the embeddings (see *Section 3.1.3.2* for details on embeddings and text tokenisation).

([447]) Breaking an image into components based on spatial frequencies means separating it into different levels of detail. By analysing how quickly colours or brightness change across the image, it is possible to distinguish between broad areas (like the sky) and sharp, detailed parts (like edges or textures). This process, done using mathematical techniques, helps models focus on specific patterns, like fine details or general shapes.

representations are often compact and can be advantageous for tasks where fine-grained spatial information is less critical.

Other Techniques:

- **Models purely based on CNNs** do not discretise data into tokens. They use feature maps generated by convolutional layers and process the image as a continuous entity.
- **Traditional GANs** (like StyleGAN) work directly with continuous pixel-level data. The generator outputs raw pixel values or features without requiring token-like representations.

Annex IX: Complementary information on Training Data Memorisation

IX.1 LLMs

Figure IX.1-1 details the probabilities found in 2023, Carlini et al. of training data regurgitation depending on (a) model size, (b) length of the text given as input prompt and (c) frequency of the sequence within the training dataset.

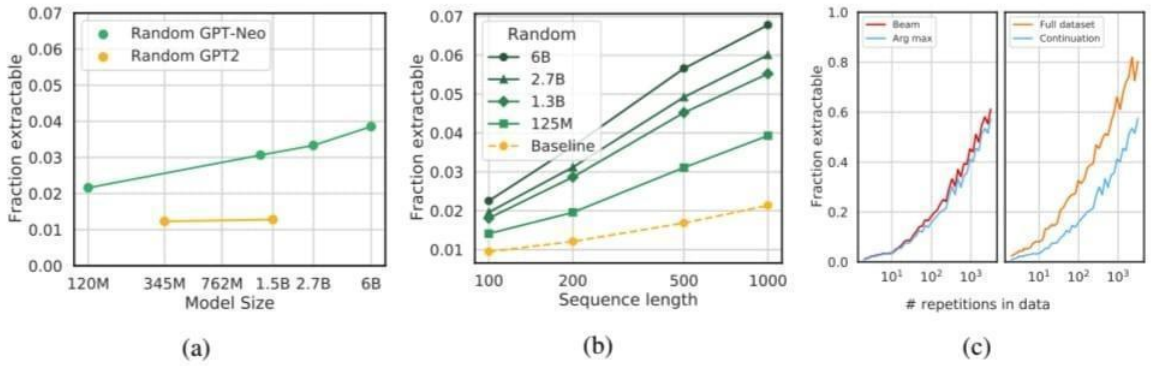


Figure IX.1-1: Diagrams representing how the memorisation rate varies against the variation of (a) model's size, (b) input prompt's length and (c) training data duplication.

A significant portion of the memorised sequences contained licensing information: they are very likely to be highly duplicated into a web scraping-derived training dataset.

In *Table IX.1-1* some examples produced during the study are reported: they include the input prompt (composed by 50 tokens) and the model's generated continuation. In those cases, the researchers found that the model's output matches the training string regardless of its size (they tested models from the GPT-Neo family with 6B, 2.7B, 1.3B and 125M parameters).

Prompt	Continuation
use this file except in compliance with the License. * You may obtain a copy of the License at * http:// www.apache.org/licenses/LICENSE-2.0 * Unless required by applicable law or agreed	to in writing, software * distributed under the License is distributed on an "AS IS" BASIS, * WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. * See the License for the specific language
* * This program is free software; you can redistribute it and/or modify * it under the terms of the GNU General Public License version 2 and * only version 2 as published by the Free Software Foundation. * *	This program is distributed in the hope that it will be useful, * but WITHOUT ANY WARRANTY; without even the implied warranty of * MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the *

Table IX.1-1: Examples of input prompts and relative model's output when training data strings were successfully extracted.

IX.2 Image Models

In Carlini, Hayes, et al., 2023, Stable Diffusion and Imagen were selected for the study on training data memorisation as representative examples of public and non-public Diffusion Models, respectively.

In the case of images, the significance of **approximate memorisation** becomes more prominent compared to simple verbatim memorisation. This is because high-resolution images, composed of a vast number of bits, can still appear visually similar even when a considerable portion of their pixels differ. Conversely, images may be algorithmically classified as similar using mathematical similarity functions, even when they are not, often due to large uniform background areas. To address this issue, Carlini et al. proposed a similarity measurement method for evaluating image similarity that involves partitioning the images, comparing each partition with all the partitions of the other image, and using the minimum similarity score as the final measure.

They used image captions as prompts to induce the models to regurgitate memorised images, discovering that Imagen exhibited a higher memorisation rate than Stable Diffusion. Specifically, they managed to extract 50 images from Stable Diffusion in up to 175 million attempts, while Imagen produced 23 training samples when prompted 1,000 times.

Since these performances were achieved in a **laboratory setting**—vastly different from real-world, day-to-day use cases—those points have to be taken in consideration:

- The experiments targeted images known to be highly **duplicated** in the models' training datasets. This **intentional bias** was introduced to reduce computational costs, as duplicated images are more likely to be extracted;
- The average GenAI user is unlikely to possess the capability to attack a model with the intent of extracting memorised images. However, they remain vulnerable to **targeted attacks**, which can still cause **economic harm** to the affected rights holders;
- Memorisation increases with model size and accuracy. This suggests a **growing concern** for future iterations of generative AI models.

The study of Carlini et al. also had the important outcome to demonstrate that the level of memorisation is affected by **the way the model is trained**: they measured how this phenomenon is less frequent in models based on the **Generative Adversarial Network (GAN)** architecture than in Diffusion Models. A possible explanation given by the researchers is that GANs' generators are only trained using **indirect information about the training data**, i.e., using gradients from the discriminator (for more information on the GANs' training process, please refer to *Section 2.1.1*). This opens new research directions towards reducing memorisation even when models will become larger and more accurate than today.

Meanwhile, they found the same images to be memorised by both types of the compared architectures, possibly suggesting that **some characteristics of the data point itself can influence the degree of memorisation**. While they encouraged **further research** to uncover the exact rationale behind this phenomenon, they found that the most frequently extracted images were those that **differed significantly from the rest of the dataset** in terms of image features (in other words, the most 'original' images).

Annex X: DeepSeek's Optimisation Strategies

X.1 Overview

The optimisation strategies used by DeepSeek include:

- **Multi-Head Latent Attention (MHLA):** an advanced mechanism designed to reduce memory overhead while preserving expressiveness (see the next section for further details on MHLA);
- **Dynamic Projection:** The model **reconstructs compressed information on demand** during inference, ensuring efficient retrieval of stored representations without excessive memory usage;
- **FP8 Mixed Precision Training**: this reduces training costs leveraging FP8 (8-bit floating point) mixed precision computations, which optimise both speed and memory usage. FP8 enables faster training while maintaining numerical stability;
- **Multi-Token Prediction (MTP):** Unlike traditional models that predict tokens one at a time, this technique optimises learning by **predicting multiple tokens simultaneously**, improving efficiency and accelerating training;
- Optimised Data Utilisation: this process places strong emphasis on **data quality over quantity**, ensuring that its models achieve higher efficiency per training sample.
- **High-Quality Data Curation:** Instead of indiscriminately increasing dataset size, DeepSeek prioritises **clean, high-quality data** to enhance training performance.
- **Diverse & Multilingual Training Data:** The models are trained on 8.1 trillion tokens, covering a broad range of languages and domains, allowing them to generalise across multiple tasks.

X.2 Focus: Multi-Head Latent Attention (MHLA) and Its Role in Transformer Optimisation

X.2.1 Summary

Multi-Head Latent Attention (MHLA) is a very recent development (as to date) and an advanced technique in the context of GenAI that makes AI models faster and more memory-efficient without sacrificing accuracy (Meng et al., 2025).

The central benefit of this technology is the **lower memory usage**. Instead of storing **all** processed data, MHLA **keeps a simplified version and reconstructs details only when needed**. This allows AI models to **scale better** with storage reduced to 5-13% of the usual size.

Another advantage is that it allows models to provide **faster responses**: it reduces unnecessary computations, speeding up AI-generated outputs. Moreover, it **can be combined with other efficiency techniques**. A deeper technical explanation of how MHLA compresses and reconstructs data is provided in the following sections.

Although **DeepSeek-V2 and DeepSeek-V3** were the first AI models to use MHLA, this technique is **not exclusive to these models**. As AI models continue to scale, **memory-efficient attention mechanisms like MHLA can be essential for improving inference speed and reducing resource consumption**.

X.2.2 What is MHLA

Multi-Head Latent Attention (MHLA) is an advanced transformer-based optimisation technique designed to improve **memory efficiency and inference speed** while maintaining the effectiveness of multi-head attention mechanisms. Unlike standard **multi-head attention (MHA)**, which stores full-sized key-value (KV) states for all attention heads, **MHLA applies a latent space compression approach**, reducing the stored KV representations and reconstructing details when necessary.

This method **optimises the key-value (KV) caching mechanism** used in autoregressive generative models, significantly reducing memory consumption. By introducing a **low-rank factorisation** of the KV cache, MHLA enables a reduction in memory overhead **to as little as 5–13% of the original cache size**, allowing for greater scalability in large transformer architectures.

X.2.3 MHLA Mechanism: Optimised KV Storage & Retrieval

At its core, **MHLA modifies the standard multi-head attention mechanism by altering how KV representations are stored and accessed**. Instead of storing and computing attention over **full-size key-value tensors**, MHLA applies a **projection-based latent space reduction** to efficiently compress KV representations. The process consists of the following key components:

X.2.3.1 Low-Rank Compression of KV States

Traditional **multi-head attention** stores a **key tensor** K and a **value tensor** V for every query at each layer.

In MHLA, the full KV states are not retained; instead, they are projected into a **lower-dimensional latent space** using a learnable transformation matrix W_k for keys and W_v V for values:

$$K' = W_k K$$

$$V' = W_v V$$

These projected tensors K' and V′ are stored instead of the full-sized KV states, dramatically reducing the memory footprint.

X.2.3.2 Up-Projection During Inference

When computing attention during inference, the model reconstructs the original KV representation from the compressed latent space:

$$\widehat{K} = W_K^T K'$$

$$\widehat{V} = W_v^T V'$$

This up-projection **restores expressiveness** while keeping memory consumption low.

X.2.3.3 Query-Key Similarity Estimation with Reduced KV Cache

During inference, the model uses the stored compressed KV states K′ and V′ to compute query-key similarities.

Instead of computing full attention over all KV pairs, **MHLA enables selective reconstruction**, ensuring efficient computation with minimal memory overhead.

Comparison of MHLA vs Standard Multi-Head Attention across key dimensions

Feature	Standard MHA	MHLA
Memory Usage	High (full KV cache stored)	Low (5–13% KV cache retained)
Computational Complexity	$O(N^2)$ for long sequences	$O(N^2)$ for optimised retrieval
Expressiveness	Full expressiveness, but costly	Retains most expressiveness with lower cost
Inference Speed	Slower due to large KV cache	Faster due to selective reconstruction
Applicability	Requires high VRAM for large models	Scales well even in constrained environments

Table X.2.3-1: Comparison between standard Multi-Head Attention (MHA) and Multi-Head Latent Attention (MHLA).

X.2.4 Compatibility with Other Optimisation Strategies

MHLA can be **combined with other attention mechanisms** to further optimise performance:

- **Grouped Query Attention (GQA)**

  GQA enables multiple queries to **share KV representations**, reducing redundant KV storage.

  MHLA complements GQA by further **compressing KV states**, leading to significant efficiency gains.

- **Sparse Attention Mechanisms**

  In long-context models, **sparse attention** selectively computes attention scores for a subset of tokens.

  MHLA aligns well with sparse attention, **optimising memory usage while maintaining key information retrieval**.

- **Retrieval-Augmented Generation (RAG)**

  RAG models **retrieve external knowledge** before generating responses.

  Integrating MHLA in RAG models **reduces memory bottlenecks** when handling large retrieval datasets.

X.2.5 MHLA's Applications Beyond DeepSeek

While DeepSeek-V2 and DeepSeek-V3 were the first AI models to implement MHLA, **the technique is broadly applicable to various AI architectures**, particularly those facing **memory constraints or requiring inference efficiency**.

- **GPT-like Transformers**

  LLMs that rely on multi-head attention can **integrate MHLA** to improve KV cache efficiency.

  Example: GPT architectures can reduce VRAM requirements **without sacrificing generation quality**.

- **Long-Context Models**

  Models that require long-context retention (e.g., **Anthropic Claude, Gemini 1.5**)

  MHLA can help such models handle extended contexts **without quadratic memory growth**.

- **Edge AI and On-Device AI**

Low-power AI applications (e.g., AI inference on mobile devices) benefit significantly from MHLA.

Reducing KV cache sizes allows **LLMs to run efficiently on smartphones, IoT devices, and embedded systems**.

Annex XI: Technical Instruments underlying technical reservation measures

Several technical reservations measures rely on a number of internet-related languages and protocols (HTML, HTTP, ODRL, RightsML), as well as on particular technical instruments (such as blockchain or federated registries) that are explained below.

XI.1 HTML (HyperText Markup Language)

HTML is the standard language used to define and structure content on the web. It is based on a series of **<tags>** enveloping the webpage's text to define its structure and appearance. It also allows inserting images, web links and other media content in the page structure using specific tags ("<img>" for images, "<a>" for links…). Inside the angle brackets further attributes can be defined to customise the tag's behaviour.

For a better understanding of the protocols described in this Annex, it is important to note the existence of the tags "<head>" and "<body>" which divide the html file in two sections containing the **page's metadata** and content respectively. In the image ([448]) below a simple html file structure is outlined:

```
<!DOCTYPE html>
    <html>

        <head>
            <title> Title here </title>
        </head>

        <body>
            Web page content goes here.
        </body>

    </html>
```

An HTML page can also refer to content of other formats (images, videos, audio) to make it appear in the webpage. In practice, a dedicated tag (such as "<img>" in case of an image) is inserted in the

([448]) HTML Document Structure: A Comprehensive Guide with Examples, Hyno blog, 22 June 2023 (accessed 25 November 2024).

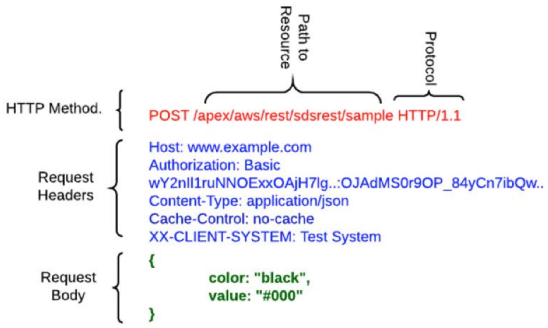
desired position in the HTML file. Inside the tag, through an attribute called “src”, the URL of the resource is embedded. This will allow the **content to be displayed inside the page, but it won't be part of the HTML file** itself.

Often, in the **web server a dedicated directory** is created **to contain all the media files** referenced by the HTML pages.

XI.2 HTTP (HyperText Transfer Protocol)

HTTP is the foundational protocol used on the web for transferring data between a client (like a browser) and a server, enabling the fetching of resources such as HTML pages, images, and more.

It is a network protocol: it defines the **format of the messages** exchanged between communicating nodes. This format includes a HTTP header and a (not strictly mandatory) HTTP body. An example of HTTP message is ([449]):



XI.3 JSON File Format

JSON (JavaScript Object Notation) is a lightweight, text-based format used to represent structured data as **key-value pairs and arrays**.

Example of simple JSON object, enclosed in **curly braces**, where key-value pairs are separated by colons:

([449]) ORDS HTTP Headers and Variables Revisited for ORDS3, JMJ CLOUD blog, 8 September 2016 (accessed 25 November 2024).

```
{
 "name": "John Doe",
 "age": 30,
 "isEmployed": true
}
```

Example of JSON array, enclosed in **square brackets**, having values separated by comma:

```
[
 "JavaScript",
 "Python",
 "HTML"
]
```

JSON supports nesting of objects and arrays, making it both human-readable and easy for machines to parse and generate. This versatility has made it a widely used format in web development.

XI.4 Open Digital Rights Language (ODRL) and RightsML

The **Copyright Infrastructure Task Force([450])**, indicates **ODRL** as a relevant format to express **obligations** in a **machine-readable** way. The syntax proposed by the **World Wide Web Consortium (W3C)** in its recommendation provides enough flexibility to express the payment agreements between parties, as shown in *Figure XI.4-1*.

([450]) The Copyright Infrastructure Task Force aims to create a cohesive system that allows digital content to carry essential information about its origin, rights, and permissible uses. Acting as a standardisation forum rather than a standard development organization, the task force facilitates collaboration among member states and affiliates to address challenges posed by AI and digital content.

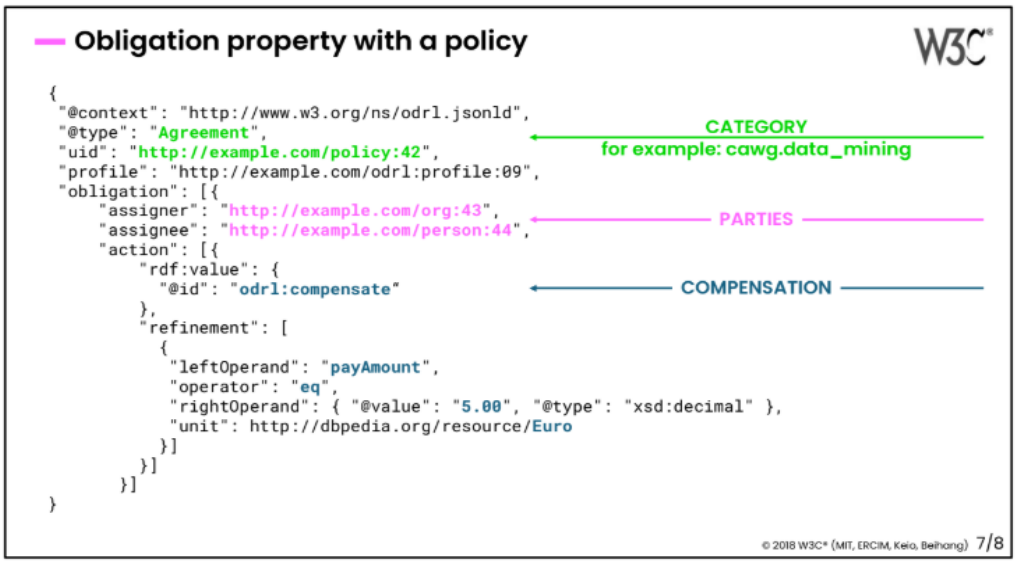


Figure XI.4-1: Example highlighting ODRL's syntax ([451])***.***

This syntax can be embedded into files **metadata** associated with digital assets, in media **manifest files** or encoded directly into the **smart contract on the blockchain**.

RightsML is International Press Telecommunications Council's (IPTC) Rights Expression Language for the media industry. RightsML ([452]) Born in 2013, its last version was released in 2018 and it is synchronised with the latest version of ODRL.

XI.5 Blockchain

I addition to technical protocols, a number of technical instruments are used to develop technical measures covering TDM opt-out, in particular blockchain and federated registries. A detailed comparisons of these technical instruments can be found in *Table 3.4.2.2-1*

Blockchain and Smart Contracts

([451]) New ODRL Co-Chairs, W3C Community Business Groups, 2 Auguste 2018 (accessed 14 March 2025).

([452]) RightsML, IPTC (blog) (accessed 5 December 2024).

Blockchain solutions to express opt-out are still in their **early development stages** but may be a way to address the shortcomings of current metadata solutions, particularly in terms of preventing tampering and ensuring long-term integrity.

On a technical level, **Blockchain is a decentralised and distributed ledger** that records transactions in a series of 'blocks'. The blockchain's **immutability** is primarily achieved through the combination of cryptographic hashing, distributed consensus, and the decentralisation of its network.

Each block contains a list of transactions, a timestamp, and a cryptographic hash of the previous block. **Cryptographic hashing** involves computing a fixed-length string (hash) from transaction data, ensuring that any tampering with data changes the hash, thereby **revealing inconsistencies**. **This cryptographic linkage** ensures that altering a single block would require changing all subsequent blocks, which is **computationally impractical** without the consensus of the majority of the network participants. This makes it difficult, almost impossible, for malicious actors to alter past records due to the computational effort required to rewrite history.

In the **distributed consensus mechanism**, commonly **Proof of Work (PoW)** or **Proof of Stake (PoS)**, nodes, known as miners or validators, validate and agree upon new blocks to be added to the chain.

A **smart contract** is a program stored on a blockchain that automatically enforces or executes the terms of an agreement when specific conditions are met. It is secure, immutable, and transparent, eliminating the need for intermediaries. Written in code, it automates processes like payments, asset transfers, or record updates in a decentralised manner.

In the field of copyright, blockchain can utilise **cryptographic tokens** to represent metadata, such as ownership rights and licensing terms. This allows for automated and standardised copyright-related transactions through smart contracts (Bacon et al., 2018) ([453]).

Blockchain technology could be effective for maintaining a **tamper-proof registry of training data** used in GenAI, ensuring the **provenance and integrity** of such data. Each entry on the blockchain can include **detailed metadata regarding rights associated with a piece of content**—such as the scope of the license, duration of use, and royalty terms. AI developers can access this information

([453]) Blockchain Demystified: A Technical and Legal Introduction to Distributed and Centralised Ledgers, Richmond Journal of Law and Technology, 6 November 2018.

to verify compliance before using the data. **Smart contracts** can be used to enforce licensing terms automatically, triggering **royalty payments** or **revoking usage rights** based on predefined criteria.

At the same time, blockchain also presents notable **challenges** and structural incompatibilities:

- The **highly fragmented nature of digital content metadata** conflicts with blockchain's impersonal, borderless, standardised, and automated regulatory framework;
- The immutability of blockchain transactions, while beneficial for ensuring an unalterable record, introduces issues when disputes arise, such as cases involving misidentified artists, contractual amendments, or dispute resolution outcomes. Moreover, as some interviewed rights holders noted, each time the content itself has to be updated, referencing or re-uploading to the blockchain is non-trivial;
- The **anonymity** of the parties involved in blockchain-based contracts further complicates dispute resolution (Crowell&Moring et al., 2022);
- Interviewed stakeholders—primarily content providers—raised concerns about the **scalability** of blockchain-based solutions. They highlighted that implementing and maintaining a blockchain system entails significant development costs, transaction fees, and ongoing infrastructure expenses. Additionally, managing millions of transactions across a global network could overwhelm those systems, posing further challenges;
- Moreover, many publishers use **legacy systems** for content management. Developing solutions to bridge these with blockchain platforms could add a further layer of complexity.

In practice, blockchain's use in copyright management has seen early applications in the **music industry**, where tools were developed to allow artists to self-publish and self-license without involving publishers or Collective Management Organisations (CMOs). More recently, initiatives such as Fuga ([454]) and Unison ([455]) have demonstrated a higher level of maturity, with the potential to standardise copyright management and distribution within the music sector (Crowell&Moring et al., 2022).

([454]) Fuga website (accessed 14 March 2025).

([455]) Unison website (accessed 14 March 2025).

XI.6 Federated Registries

Federated Registries present an alternative supporting flexibility and semi-centralised control. Unlike blockchain, Federated Registries are designed to aggregate and manage information through collaboration among multiple trusted institutions or authorities. Technically, Federated Registries function by synchronising databases across several entities, each contributing their portion of verified data. This distributed approach allows multiple stakeholders to access and contribute to the registry, reducing centralised control while maintaining effective oversight. Federated Registries utilise secure APIs, allowing participants to query the registry in real-time, thereby ensuring consistency and accuracy in copyright data management. This structure also supports timely reflection of changes, such as updated licensing terms or rights ownership, which is crucial in the context of GenAI where data use and ownership are constantly evolving. Federated Registries **effectively address one of blockchain's significant limitations**: the difficulty of making corrections or adjustments once data is recorded.

In the context of **GenAI**, Federated Registries may enable the **coordination of metadata across different authoritative sources**, reducing fragmentation and ensuring that AI developers have access to the **most up-to-date rights information**.

From a technical perspective, Federated Registries provide **APIs** that allow AI developers to **verify permissions in real-time**, streamlining the **inclusion or exclusion of data** in AI training. This makes Federated Registries particularly useful where **metadata accuracy** and **quick reflection of changes** are critical.

However, some interviewed AI developers pointed out that registry-based rights management approaches pose significant challenges related to the large amount of data to be stored:

- Both the number of rights holders and internet URLs are incredibly high;
- **Privacy-related concerns could emerge**, particularly regarding the exposure of sensitive licensing data and potential risks to rights holders anonymity;
- Fraudulent players and and possible mistakes from rights holders would need extra management;
- The **highly fragmented nature of rights declarations** could impede structured and efficient data storage, making it difficult to maintain consistency across different datasets.

The table below summarises and compares the key features of Blockchain and Federated Registries as tools for managing rights reservations:

Problem/Function	Blockchain (Pros & Cons)	Federated Registries (Pros & Cons)
Immutability	**PROs:** Provides an immutable record ensuring data integrity and traceability. **CONs:** Difficult to make corrections or adapt records in case of errors or changes.	**PROs:** Can reflect changes in data more readily. **CONs:** Not inherently immutable, which may impact trust and auditability.
Transparency	**PROs:** Blockchain ensures full transparency across all participants. **CONs:** Transparency can lead to privacy concerns, especially with sensitive metadata.	**PROs:** Transparency is managed, allowing more control over visibility of sensitive information. **CONs:** Limited transparency compared to blockchain.
Decentralisation	**PROs:** Fully decentralised, reducing reliance on a single authority. **CONs:** Requires consensus mechanisms, which can be resource-intensive (e.g., Proof of Work).	**PROs:** Semi-centralised control allows for better management and oversight. **CONs:** Some degree of central control is still required, which may limit decentralisation.
Scalability	**PROs:** Blockchain can scale across multiple nodes globally. **CONs:** Scalability can be limited due to consensus requirements and transaction speed.	**PROs:** Federated Registries can be more easily scaled by adding new trusted entities. **CONs:** Coordination between entities can become complex.
Data Integrity	**PROs:** Cryptographic hashing ensures data cannot be tampered with without detection.	**PROs:** Real-time updates ensure data remains current and accurate.

	CONs: Once data is recorded, corrections are challenging.	**CONs:** Vulnerable to inconsistencies if synchronisation between entities fails.
Real-Time Updates	**CONs:** Blockchain inherently lacks real-time flexibility due to its consensus process.	**PROs:** Allows real-time data validation and updates through secure APIs.
Automation via Smart Contracts	**PROs:** Smart contracts enable automated enforcement of agreements and royalty distribution. **CONs:** Smart contracts are inflexible if conditions change.	**PROs:** APIs can be configured to accommodate changes in contractual terms dynamically. **CONs:** Lacks the same level of automation as blockchain smart contracts.
Cost of Implementation	**CONs:** Blockchain can be costly due to energy consumption (e.g., Proof of Work) and computational requirements.	**PROs:** Generally more cost-effective compared to blockchain. **CONs:** Requires substantial collaboration and infrastructure setup.
Compliance and Legal Oversight	**PROs:** Provides a permanent and transparent record useful for compliance. **Cons**: Regulatory uncertainty regarding the legality of blockchain-based records.	**PROs:** Easier to align with legal frameworks and integrate with existing authorities. **CONs:** Requires trusted central entities, which may introduce vulnerabilities.

Table XI.6-1: Comparison between Blockchain and Federated Registries.

Annex XII: Active Internet Drafts for further adapting REP as an IETF standard.

- ***Illyes Proposal ('draft-illyes-rep-purpose-00')*** ([456]) - This October 2024 proposal was submitted by one of the Google Analysts that co-authored the RCF 9309 Proposed Standard. It proposes the introduction of a 'user-agent-purpose' token to complement the existing 'user-agent' token, where the user-agent-purpose token is a substring of the user-agent identification string. Presumably, a process of standardisation may have to follow in order to define syntax for recognised uses that may form the user-agent-purpose string, thus possibly deferring standardisation of use case disaggregation to a later process.

- ***Canel-Madhaven Proposal ('draft-canel-robots-ai-control-00')*** *([457])* - This October 2024 proposal was submitted by two Microsoft product managers. It proposes the introduction of a new set of rules to complement existing allow/disallow rules, specifics for crawling for the purposes of AI training. The proposal is the use of a 'DisallowAITraining' directive to instruct the parser (crawler) to not use the data for training AI models. These rules can also be implemented via an HTML meta tag *(<meta name="examplebot" content="AllowAITraining">.*

- ***Jimenez Proposal ('draft-jimenez-tbd-robotstxt-update-00)*** ([458]) - This November 2024 proposal was submitted by a researcher affiliated with the multinational telecommunications company Ericsson, and explicitly states that its aim is to *"...enhance the management of web content access by AI systems, distinguishing between training and inference activities."* It introduces to the REP standard, new terminology that differentiates between a (traditional) crawler and an 'AI crawler', and proposes that the user-agent syntax should be updated to recognise the convention of '*-ai*' syntax that differentiates an AI crawler from a traditional crawler (e.g., *ExampleBot*, vs. *ExampelBot-ai).* This proposal specifically claims that it may be problematic to create specific purpose-oriented lines (as suggested in the *Canel-Madhaven Proposal*) due to semantic issues with possible new lines which have the same

([456]) Robots Exclusion Protocol User Agent Purpose Extension, IETF Datatracker (accessed 14 March 2025).

([457]) Robots Exclusion Protocol Extension to communicate AI preferences vocabulary, IETF Datatracker (accessed 14 March 2025).

([458]) Robots.txt update proposal, IETF Datatracker (accessed 14 March 2025).

meaning and effect of existing lines (e.g., *DisallowThisProperty*, vs. *Disallow*). The overall effect of this proposal is to disaggregate standard crawling from 'AI crawling' where the latter is explicitly related to gathering content for training purposes. The proposal considers use cases of gathering content for 'AI inference', as this is akin to normal web-crawling. It thus does not appear to specifically disaggregate crawling to support RAG (Retrieval-Augmented Generation) as a specific use case for which a site can indicate its allowance preferences.

Annex XIII: C2PA Syntax Details

The Figure below reports an example of a C2PA manifest.


```
{
  "active_manifest": "contentauth:urn:uuid:5a08e472-974c-422e-b38a-d6c7326481a5",
  "manifests": {
    "contentauth:urn:uuid:5a08e472-974c-422e-b38a-d6c7326481a5": {
      "claim_generator": "make_test_images/0.12.0 c2pa-rs/0.12.0",
      "title": "C.jpg",
      "format": "image/jpeg",
      "instance_id": "xmp:iid:e928fac1-8473-4c70-8219-369e91d4e58d",
      "thumbnail": {
        "format": "image/jpeg",
        "identifier": "contentauth-urn-uuid-5a08e472-974c-422e-b38a-d6c7326481a5.jpg"
      },
      "ingredients": [],
      "assertions": [
        {
          "label": "stds.schema-org.CreativeWork",
          "data": {
            "@context": "http://schema.org/",
            "@type": "CreativeWork",
            "author": [
              {
                "@type": "Person",
                "name": "Gavin Peacock"
              }
            ]
          },
          "kind": "Json"
        },
        {
          "label": "c2pa.actions",
          "data": {
            "actions": [
              {
                "action": "c2pa.created"
              },
              {
                "action": "c2pa.drawing",
                "parameters": {
                  "name": "gradient"
                }
              }
            ]
          }
        }
      ],
      "signature_info": {
        "issuer": "C2PA Test Signing Cert",
        "cert_serial_number": "720724073027128164015125666832722375746636448153",
        "time": "2022-08-19T19:03:41+00:00"
      },
      "label": "contentauth:urn:uuid:5a08e472-974c-422e-b38a-d6c7326481a5"
    }
  }
}
```


Figure XIII-1: An example of a C2PA manifest([459]).

The protocol establishes a list of mandatory fields to be included in the C2PA manifest:

- **c2pa.actions** – Documents the actions undertaken on the content, such as its capture, modification, or export.

([459]) Manifest Examples, Content Authenticity Initiative (accessed 24 February 2025).

- **c2pa.credential** – Identifies the entity responsible for generating the manifest, incorporating details such as digital signatures and certificates.
- **c2pa.signature** – Contains the cryptographic signature that verifies the authenticity and integrity of the manifest.

Moreover, this data provenance protocol is a possible solution to flag the output data produced by a GenAI model. In particular, the standard provides special tags as ways to report detailed provenance information in such cases: for generative models, the designation **trainedAlgorithmicMedia** is suitable, while for non-media outputs, the designation **c2pa.trainedAlgorithmicData** should be used.

```
  "actions": [
    {
      "action": "c2pa.created",
      "when": 0("2023-02-11T09:00:00Z"),
      "softwareAgent" : {
          "name": "Joe's Photo Editor",
          "version": "2.0",
          "schema.org.SoftwareApplication.operatingSystem": "Windows 10"
      },
      "digitalSourceType":
"http://cv.iptc.org/newscodes/digitalsourcetype/trainedAlgorithmicMedia",
```

Figure XIII-2: Example of Generative output tagging using C2PA ([460]).

XIII.1 C2PA: Training and Data Mining Assertions

Before the release of C2PA **v2.0** in January 2024 ([461]), the official documentation contained the syntax definition for 'Training and Data Mining Assertions', which were designed to embed the corresponding reservation directly into the digital asset. An example can be found in *Table XIII.1-1*. In particular, the specifications include a flexible list of possible media usages, including ([462]):

- AI Training;
- AI Inference;

([460]) See C2PA Specifications, C2PA (accessed 28 November 2024).

([461]) Ibid.

([462]) See C2PA Specifications - Training and Data Mining, C2PA Specifications 1.4., C2PA (accessed 3 March 2025).

- AI Generative Training;
- Data Mining.

AI Generative Training and AI Training are separate values because the first enables new assets to be created, while other types of training, such as the ones targeting object detection, do not. AI Inference is the process enabled by RAG technologies (see *Section 4.1.2*). Finally, Data Mining is distinct from AI Training as it is a broader practice that can serve various purposes beyond just training AI models ([463]).

This approach ensures granular, standardised, and proactive terms and conditions. These categories remain flexible for future adjustments.

The different types of data use are paired with permissions such as:

- Allowed;
- NotAllowed;
- Constrained.

In the absence of additional information, constrained shall be treated as equivalent to notAllowed. More details on the constraints may be provided in the **constraints_info** text field.

```
{
 "entries":
        "c2pa.ai_training" : {
                "use" : "allowed"
        },
        "c2pa.ai_generative_training" : {
                "use" : "notAllowed"
        },
        "c2pa.data_mining" : {
                "use" : "constrained",
                "constraint_info" : "may only be mined on days whose names end in 'y'"
```

([463]) While AI Training focuses on learning patterns to generate predictions or outputs, Data Mining involves extracting meaningful insights and patterns from large datasets, which can be applied in diverse fields such as business analytics, scientific research, and decision-making processes.

```
        }
}
```

Table XIII.1-1: Example of C2PA TDM assertion following the syntax prior to version 2.0 ([464]).

From version 2.0, the assertions' syntax has slightly changed because it became an extension functionality directly maintained by the **Creator Assertions Working Group (CAWG)** ([465]). In particular, as shown by the example in *Table XIII.1-2*, the keyword 'c2pa' in the assertion identifier has been replaced by 'cawg'.

```
{
 "entries":
        "cawg.ai_training" : {
                "use" : "allowed"
        },
        "cawg.ai_generative_training" : {
                "use" : "notAllowed"
        },
        "cawg.data_mining" : {
                "use" : "constrained",
                "constraint_info" : "may only be mined on days whose names end in 'y'"
        }
}
```

Table XIII.1-2: Example of C2PA TDM assertion following the new syntax from version 2.0 ([466]).

([464]) See C2PA Specifications, C2PA (accessed 28 November 2024).

([465]) Training and Data Mining Assertion, DIF Creator Assertions Working Group (accessed 3 March 2025).

([466]) Ibid.

Annex XIV: Deezer AI music detection tool

XIV.1 How It Works

The tool is designed to detect AI-generated audio across both vocal and instrumental components, spanning multiple music genres (Afchar et al., 2024).

Researchers modelled the architecture of a typical AI music generator, dividing it into two core components: an **autoencoder (AE)**, responsible for generating the audio signal, and a **Large Language Model (LLM)**, which assembles these signals into a sequence to produce music conditioned by the input prompt:

'*In layperson's terms, we can summarise that the AE does the waveform synthesis part while the LLM does the semantic work of generating a coherent musical sequence through time.*' (Afchar et al., 2024)

The researchers then prioritised detecting whether an audio signal originated from an AE, as opposed to analysing the influence of the LLM. In particular, they exploited the **tendency of AEs to produce data with specific "footprints"** generated from internal algebraic operations. They justified this approach as being simpler than detecting if an entire music sequence has been artificially generated.

They **trained an AI classifier on a dataset made of real music** samples (taken from the FMA open dataset) and their **corresponding reconstructed versions**. The latter were obtained by leveraging the AEs of some common AI music generators (such as Suno v2, MusicGen and Vampnet). In this way, the classificator can learn to distinguish between a real or synthetic audio by distinguishing the specific AEs' footprints.

XIV.2 Evaluation

During controlled laboratory evaluations, the detector achieved an **accuracy of 99.8% demonstrating exceptional performance in distinguishing AI-generated music from real audio samples.** Despite filing two patents, the researchers recognised the need for further evaluation—

specifically testing robustness against audio manipulation, interpretability, and generalisation to unseen models (Afchar et al., 2024).

The robustness evaluation focused exclusively on common audio transformations, deliberately omitting adversarial attacks to assess baseline vulnerabilities before testing advanced manipulation techniques. As shown in *Figure XI.2-1*, accuracy deteriorated considerably under certain manipulations, such as noise addition and re-encoding ([467]), which altered the distribution of audio's features.

It is important to note that **the discriminator was not trained on manipulated audio** (i.e., on a training dataset appositely augmented with transformed audio samples). This suggests that there may be room for further improvement by introducing transformed samples during fine-tuning. However, *Figure XI.2-1* shows that the discriminator already obtains high performances when fed with audio manipulated in ways that do not alter the distribution of audio's features.

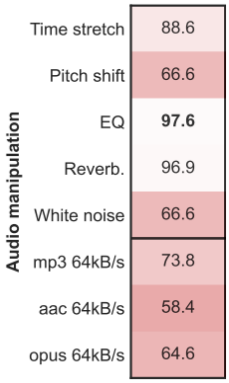


Figure XI.2-1: Average accuracy values obtained when testing Deezer's detector for its ability to classify AI-generated audio after undergoing various transformations. The mean accuracy is computed across different AI music generation models (Afchar et al., 2024).

To assess the **generalisability** of this technology, the study adopted a methodology in which a **new classifier was trained separately on data generated by each AI music model under consideration**. Each new classifier followed Deezer's final training methodology but was trained on distinct, model-specific datasets. Each discriminator's accuracy was then tested against audio from

([467]) Reencoding includes changing the audio format or adding it to a video clip.

AI music generators it had not been trained on, ensuring exposure to previously unseen generative characteristics.

The results, as reported in *Figure XI.2-2*, suggested that **generalisation is readily achievable within the AI music generators embedding AEs belonging to the same family ([468])**. However, **when evaluating generalisability across different AE families, performance declined significantly**, approaching zero.

From the last two evaluations, the researchers concluded that **Deezer's classification needs specific fine-tuning for successfully managing each possible audio manipulation or AE**. However, as they noted, there will *always* be a new unseen manipulation or AI music generator. Then, **the AI music detector would need regular updates** to ensure its efficacy.

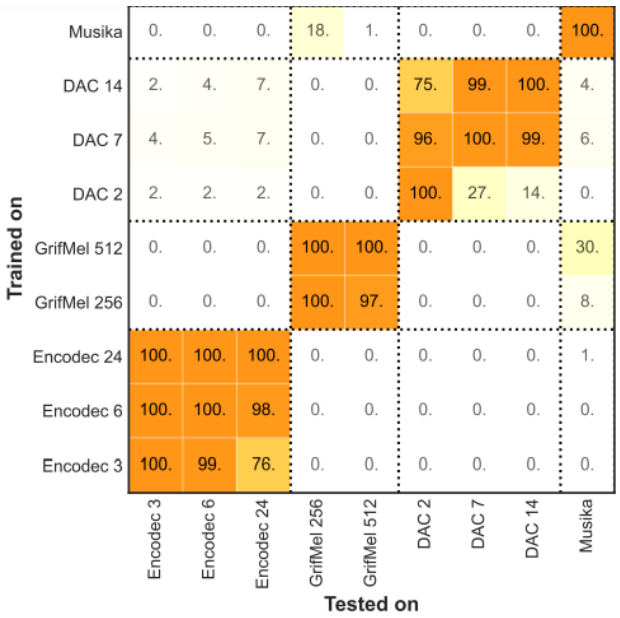


Figure XI.2-2: Matrix reporting the accuracy values obtained when testing the generalisability of a discriminator to successfully detect synthetic content generated by different models than the one producing the discriminator's training data (Afchar et al., 2024).

([468]) For example, there is a good transferability between autoencoders with the same architecture working at different bitrates.

Researchers also emphasised the importance of assessing the discriminator's behaviour when analysing audio that combines synthetic and real elements, such as AI-generated vocals over genuine instrumental recordings. '*In that case, what score should a detector model display? 100% fakeness due to the presence of any forgery in the track, or, some fakeness ratio?'* ([469])

To answer this question, **they tested the discriminator on a range of audio samples composed of mixtures of real tracks and their re-encoded versions**, each blended in varying proportions. Then, they were able to trace the curve showing the model's prediction trend against the real/fake mix factor, which is reported in *Figure XI.2-3*, concluding that there not exists a best expected behaviour at all, but that '*this sort of curve could be made accessible to the general public to help interpret a detector's output.'* ([470])

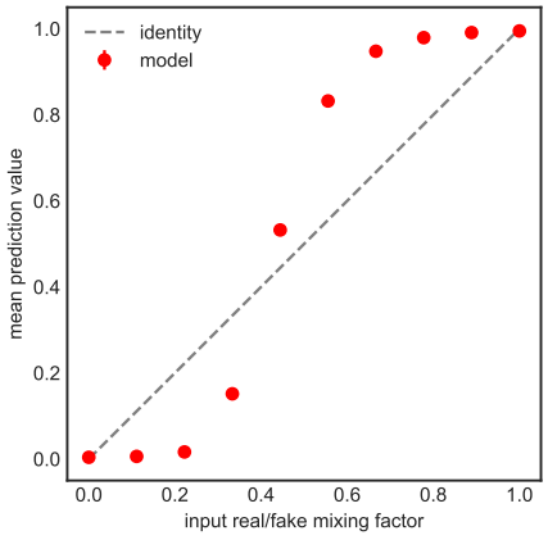


***Figure XI.2-3: Graph depicting the prediction outcomes of Deezer's discriminator when evaluated on audio samples composed of real and reencoded track segments mixed in varying proportions* (Afchar et al., 2024).**

Finally, model interpretability has been identified as a key approach to mitigating false positives. Some '**feature attribution maps**' have been developed for the purpose. Feature attribution is a technique for explainability aiming to relate the influence of an input on an output. Although these

([469]) Detecting Music Deepfakes Is Easy but Actually Hard, Cornell University, 22 May 2024 (accessed 14 March 2025).

([470]) Ibid.

feature attribution maps effectively identified the specific regions of the audio spectrogram that the discriminator labelled as 'fake,' the researchers concluded that their approach did not constitute a generalisable solution for interpretability. For instance, certain audio manipulations could produce a feature attribution map that is entirely highlighted, making interpretation challenging. As a result, they emphasised the need for caution and case-by-case evaluation.

XIV.3 Further Developments

Future developments planned by Deezer's research team include:

- Assessing the feasibility of fine-tuning the model for new AI music generators and common audio manipulations;
- Evaluating the model's robustness against advanced audio manipulation attacks.
- Enhancing the model's interpretability; and
- Introducing greater control over the model's behaviour when analysing audio containing a mix of real and synthetic components.

Annex XV: Technical definition of Watermarking (Christodorescu et al., 2024)

A watermarking method has two components:

- An encoder $Embed_{k_E}(M, p, w)$, where:
  - k_E is a secret key;
  - M is the model;
  - p is user supplied input to the model (e.g., a prompt or instructions for editing a message);
  - w is additional information (e.g., string to be embedded in a watermark).

  Some schemes might not use some parameters. For example, if a scheme does not use a secret key, then kE will not be used, and in some schemes w might not be used.

- A decoder $Detect_{k_D}(x, w)$, where:
  - k_D is the key use for detection;
  - x is content;
  - w is additional information.

  As usual, some schemes might not use certain parameters, such as kD and w. This method returns 1 if x is watermarked, and 0 otherwise.

Note that in secret key schemes, $k_E = k_D = k$ and is kept secret. In publicly verifiable schemes, k_E is the secret key and k_D is the public key.

Annex XVI: Detailed Categorisation of Machine Learning Watermarking methods

A study (An et al., 2024) has developed an extensive **benchmark to evaluate the robustness of watermarks**, offering valuable insights into their effectiveness. It categorised the more robust techniques as follows:

Post-processing methods:

- **Frequency-Domain Techniques**: Approaches like DWT-DCT (Discrete Wavelet Transform and Discrete Cosine Transform) modify the image in the transform domain to embed watermarks. They are robust against geometric transformations and resizing but may struggle with common manipulations such as JPEG compression;
- **Deep Encoder-Decoder Models**: Methods like HiDDeN, RivaGAN, and StegaStamp leverage neural networks to encode and decode invisible watermarks. Notably, **StegaStamp** demonstrates very good robustness due to training on a wide variety of real-world distortions;

In-Processing Methods:

- **Full Model Modifications**: These techniques embed watermarks during image generation by retraining the entire generative model. While effective, they require significant computational resources;
- **Partial Model Modifications**: Methods like **Stable Signature** fine-tune specific components (e.g., the decoder) of generative models to integrate watermarks;
- **Noise Vector Modifications**: Techniques such as **Tree-Ring** embed watermarks into the initial noise vector of diffusion models. This noise vector is used as the seed for the subsequent model's generations, making the information contained into it retrievable during output analysis.

Annex XVII: COPYCAT Benchmark Suite

He et al. (2024) studied the issue of copyrighted characters generation in generative image models. Along with developing a suite to test a GenAI system's tendency to generate infringing representations, they also explored a series of mitigation techniques. In particular, they focused on "negative prompting", since they believe it is more effective than the "prompt rewriting" strategy already integrated in some GenAI systems.

"**Negative prompting**" consists of specifying to the model what elements should be excluded from the generation.

To find which words to write in the negative prompt for each character of the COPYCAT's benchmark list of characters, the researchers proposed complementary strategies:

- Use the text encoder of the image generator under study to compute the **cosine similarity**([471]) **between the textual embeddings**([472]) of the considered keywords and the character's name: this serves as a method to assess the extent to which the model associates the respective tokens;
- Rank keywords basing on their **co-occurrence** with the character's name in the main training datasets (they examined LAION-2B, C4, OpenWebText and The Pile);
- Always include the **character's name** itself in the negative prompt.

The researchers determined that the most effective method involved identifying keywords that frequently co-occur with a character's name within the **LAION** dataset. This result can be explained considering that LAION is the most widely used training dataset among the considered MUTs. In the *Figure XVII-1* below, the plots indicate the number of characters detected using different top keywords ranked by various methods on (a) image generation and (b) video generation models.

([471]) Cosine similarity is a way to measure how similar two things are by looking at the angle between their representing vectors in a space. It's often used to compare text, images, or data.

([472]) For the definition of embedding, see *Section 3.1.3* on Text Tokenisation.

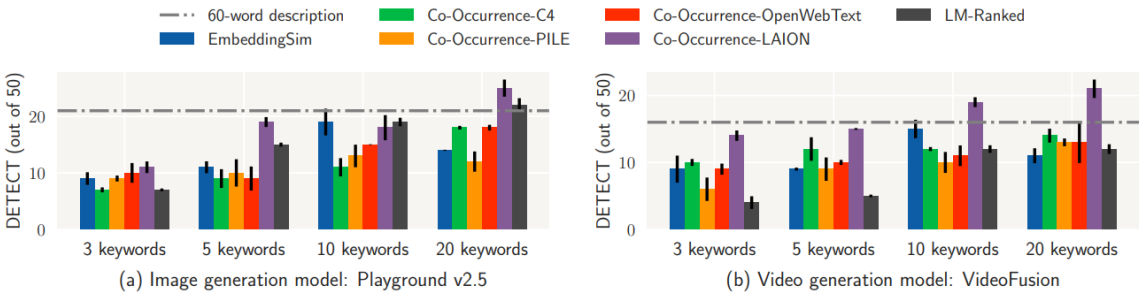


Figure XVII-1: Diagrams comparing the different extraction success rates obtained by using different reference databases to rank and select the keywords to be used in the input prompt. Both (a) Playground v2.5 and (b) VideoFusion are subject to successful extraction attacks when the 'keyword generation strategy' involves the LAION dataset (He et al., 2024).

Only five keywords, chosen from the top-ranked using an approach that examines the co-occurrences in the LAION dataset, frequently result in the generation of copyrighted characters. To avoid the generation of copyrighted characters, these are the **keywords to be included in the negative prompt** to ensure effective protection.

By **combining "prompt rewriting" with "negative prompting"** they were able to achieve a reduction in the DETECT metric of approximately 83% to 90% without significantly affecting the CONS metric. The results are reported in *Figure XVII-2.*

Model	w/o Intervention		w/ Prompt Rewriting & Negative Prompt	
	DETECT (↓)	CONS (↑)	DETECT (↓)	CONS (↑)
Playground v2.5 (Li et al., 2024a)	$30.33_{\pm 1.89}$	$0.75_{\pm 0.01}$	$4.33_{\pm 0.47}$	$0.81_{\pm 0.00}$
Stable Diffusion XL (Podell et al., 2024)	$33.00_{\pm 1.00}$	$0.73_{\pm 0.01}$	$1.67_{\pm 0.94}$	$0.77_{\pm 0.03}$
PixArt-α (Chen et al., 2024)	$24.67_{\pm 0.58}$	$0.79_{\pm 0.01}$	$4.67_{\pm 0.47}$	$0.79_{\pm 0.01}$
DeepFloyd IF (StabilityAI, 2023)	$33.67_{\pm 1.53}$	$0.71_{\pm 0.01}$	$2.00_{\pm 1.00}$	$0.72_{\pm 0.01}$
VideoFusion (Luo et al., 2023)	$28.33_{\pm 1.89}$	$0.68_{\pm 0.01}$	$11.33_{\pm 1.53}$	$0.76_{\pm 0.01}$

Figure XVII-2: The effectiveness of Prompt Rewriting and Negative Prompting is evaluated by comparing DETECT and CONS scores, as measured by COPYCAT, across different models. By significantly reducing DETECT, this mitigation strategy ensures that the model's outputs are less similar to copyrighted characters. Simultaneously, by maintaining CONS scores, it ensures that the generated outputs remain aligned with the intended objectives of the generation process (He et al., 2024).

This study demonstrates potential but requires further research to demonstrate its **scalability across a larger number of characters**, particularly in developing an effective method for generating appropriate negative prompts for each character.

Moreover, **another limitation is the definition of the CONS metric**, as it represents the alignment of the generation with the **key features** of the copyrighted character which is deliberately prevented from being replicated exactly. Instead, the study presented below proposes considering the alignment of the generation to the **input prompt directly**. This is because, even when the user's prompt describes a character that is similar to a copyrighted one, it is still reasonable to respect the user's intent, as expressed in the input prompt, as much as possible (Chiba-Okabe & Su, 2024).

XVII.1 PreGEN: Testing and Evaluation

PreGEN is a technique proposed by Chiba-Okabe & Su (2024) to further enhance the approach proposed by He et al. (2024).

As Models Under Test (MTU), i.e., models selected for testing PREGen as an integrated mitigation against the generation of copyrighted characters, **Playground v2.5** (Playground AI), **PixArt-α** (PixArt AI) and **Stable Diffusion XL** (Stability AI) were chosen.

For both direct and indirect anchoring scenarios (i.e., where the character's name is respectively present or absent in the input prompt), they conducted three experimental runs and reported the mean values for each configuration:

- Models without any intervention;
- Standard prompt rewriting with negative prompting;
- PREGen.

Figure XVII.1-1 compares the results in the **direct anchoring** scenario:

	Playground v2.5		Pixart-α		SDXL	
	DETECT	CONS	DETECT	CONS	DETECT	CONS
w/o Intervention	39.3±0.6	0.746±0.007	28.0±1.0	0.685±0.005	41.3±1.2	0.744±0.017
Standard method	6.7±1.5	0.787±0.028	2.7±0.6	0.786±0.015	1.3±0.6	0.752±0.027
PREGen	3.3±0.6	0.790±0.016	1.0±1.0	0.788±0.016	0.3±0.6	0.768±0.023

Figure XVII.1-1: Values for DETECT and CONS obtained on different models using COPYCAT evaluation suite for detecting the generation of copyrighted characters when the input prompt contains the character's name. Since DETECT indicates the similarity between the generation and a copyrighted character and CONS measures the coherence between the generation and the input prompt, this data demonstrates that PREGen performs better than the other available approaches (Chiba-Okabe & Su, 2024).

Meanwhile, in *Figure 4.5.2-9* it can be seen that PREGen still improves the standard mitigation in nearly all the configurations of the **indirect anchoring** scenario, zeroing the DETECT metric in both Playground v2.5 and Stable Diffusion XL (SDXL):

	Playground v2.5		Pixart-α		SDXL	
	DETECT	CONS	DETECT	CONS	DETECT	CONS
w/o Intervention	13.3±3.2	0.775±0.008	13.7±0.6	0.776±0.022	13.3±3.1	0.783±0.008
Standard method	1.3±1.2	0.754±0.008	1.0±1.0	0.736±0.018	0.7±1.2	0.727±0.022
PREGen	0.0±0.0	0.769±0.012	1.0±1.7	0.749±0.006	0.0±0.0	0.722±0.031

Figure XVII.1-2: Values for DETECT and CONS obtained on different models using COPYCAT evaluation suite for detecting the generation of copyrighted characters when the input prompt does not include the character's name. Since DETECT indicates the similarity between the generation and a copyrighted character and CONS measures the coherence between the generation and the input prompt, this data demonstrates that PREGen performs slightly better than the other available approaches (Chiba-Okabe & Su, 2024).

In all cases except one, PREGen demonstrates an improvement in the CONS value, indicating that this technology may provide a marginally **improved trade-off** between preventing the generation of copyrighted characters and maintaining consistency between the input prompt and the generated output. However, this benefit comes with **additional computational costs,** owing to the requirement of generating multiple samples for each input request, with only one retained as the final output.

Annex XVIII: Technical Introduction to Machine Unlearning

Random Labelling Loss

Fine-tuning on the Task Vector alone can overfit the unlearning process, causing instability. Instead, introducing **controlled noise** during fine-tuning makes the process more robust.

It consists of **randomly mismatching the labels of the data used for fine-tuning**. This noise ensures the model learns to "forget" while avoiding over-adjusting.

Weight Saliency Mapping

During learning tasks, it is a method to identify and update only the most important weights related to the training data.

It consists of computing the gradient of the loss function with respect to the model weights **during training** to **identify the most affected weights**. Only those weights whose saliency scores exceed a chosen threshold are updated.

Weight Saliency Mapping is a broader concept originating from machine learning research. It has been used for pruning neural networks and to enhance model's interpretability by looking at which parts of it contribute most.

XVIII.1 Sharded Isolated Sliced and Aggregated (SISA) Unlearning

Sharded, Isolated, Sliced, and Aggregated (SISA) training is a dataset partitioning technique introduced by Bourtoule et al. (2020) to enhance the efficiency of re-training in response to data unlearning requests.

Unlike traditional approaches that require full model retraining, SISA facilitates selective re-training by partitioning the dataset, with **each used to train a separate sub-model**. During inference, these sub-models collectively generate predictions through an aggregation mechanism, such as majority voting. This process is further optimised through **slicing**, which allows incremental training and storage of intermediate model states, reducing the computational cost of re-training.

Compared to full model retraining, SISA offers a significant time reduction, with experiments demonstrating a **speed-up of up to 4.63 times for certain datasets**, while **accuracy loss remains below 2 percentage points**.

Despite its advantages, SISA presents certain trade-offs. Partitioning **reduces the statistical representativeness of training data**, potentially leading to reduced overall model accuracy and a **risk of overfitting within smaller partitions**. Additionally, **tuning of the training parameters** becomes more complex due to the increased number of sub-models. Accuracy degradation is more pronounced in deep learning tasks involving complex datasets, such as ImageNet.

A refined variant of SISA incorporates prior knowledge of unlearning request distributions, further optimising training efficiency. This strategy, inspired by real-world regulatory differences across jurisdictions, minimises retraining costs without significant accuracy degradation.

Detailed information can be found in *Annex XIX*.

XVIII.2 Stable Sequential Unlearning (SSU)

Stable Sequential Unlearning (SSU), introduced by Dou et al. in 2024, is a method designed to address the challenges of unlearning copyrighted data from machine learning models without compromising their general knowledge and reasoning capabilities. Unlike traditional techniques such as Gradient Ascent (GA), which often lead to considerable forgetting, SSU offers a more structured and controlled approach to unlearning while minimising damage to the model's overall functionality.

SSU employs **Task Vectors** to adjust model weights corresponding to the data designated for unlearning. Notably, SSU introduces the use of two machine learning techniques into the unlearning process:

- **Random Labeling Loss**: Introduces controlled noise to prevent overfitting during unlearning;

- **Weight Saliency Mapping:** Detects and adjusts specific weights linked to the content designated for unlearning, ensuring that the broader knowledge and reasoning abilities of the model remain intact.

Furthermore, by utilising the original model for unlearning, SSU mitigates compounding errors in sequential unlearning processes.

This approach contrasts with other methods, such as GA, which indiscriminately adjusts weights and often leads to severe knowledge loss, and methods based only on task vectors, which fail to localise updates and may cause unintended degradation of non-targeted content.

SSU was **evaluated on the Llama3-8B model**, specifically for unlearning four copyrighted books. SSU showed significant improvements over baseline methods in terms of effective unlearning and knowledge retention.

SSU effectively unlearned copyrighted material while consistently outperforming baseline methods, such as GA and Task Vectors, in knowledge retention. Following unlearning with SSU, the model retained **strong performance in reasoning and general knowledge tasks**. Benchmark evaluations on datasets such as MathQA, MMLU, and GPQA indicated that, following SSU, the model retained its capabilities more effectively than the base Task Vector approach and GA-based methods. For instance, SSU attained 34.3% accuracy on MathQA, outperforming Task Vectors, which achieved 32.1%.

More details can be found in *Annex XX*.

XVIII.3 Approximate Unlearning with Idiosyncratic Expressions Replacement

Eldan and Russinovich propose an unlearning technique known as **idiosyncratic expression replacement**, designed to enable LLMsto unlearn specific literary works, such as the Harry Potter book series. This approach is classified as 'approximate unlearning,' as it modifies the model's behaviour without entirely erasing its underlying knowledge base.

The unlearning process comprises three main steps:

- **Task Vectors Computation:** These vectors are generated to guide the unlearning process.

- **Replacement of Idiosyncratic Expressions**: Specific expressions, names, and terms from the target data (e.g., the *Harry Potter* books) are replaced with generic counterparts. The model's own predictions are leveraged to create alternative labels for each token, simulating its **expected output had it not been trained on the unlearning data**. This step facilitates the removal of target content from the model's memory.
- **Fine-tuning**: The model is fine-tuned on the alternative labels, erasing the original content from its memory without disrupting its general capabilities.

The key challenge in Step 2 is identifying a generic replacement for terms related to the unlearning target. This is achieved through two complementary techniques:

- **Reinforcement Bootstrapping**: The baseline model is retrained on the target data to reinforce learning, identifying tokens not influenced by the target text.
- **Anchored Terms**: A list of idiosyncratic terms (such as character names from *Harry Potter*) is generated using GPT-4. These terms are substituted with generic alternatives, forming a dictionary that maintains coherence while eliminating specific content.

The combination of both approaches improves the unlearning performance compared to using them separately.

The method was **tested on the Llama2-7B model**, focusing on unlearning the *Harry Potter* series. Notably, the model's ability to generate Harry Potter-related content was removed with just one GPU hour of fine-tuning, compared to over 184,000 GPU hours required for the initial pre-training. Despite this significant reduction in training time, the model's performance on standard reasoning benchmarks (e.g., Winogrande, Hellaswag) remained largely unchanged, suggesting that the unlearning process did not compromise its general capabilities.

A key challenge in this technique is the potential bias introduced during the generation of alternative expressions. If the **LLM used for replacement generation has prior knowledge of the target content** (the researchers used GPT-4 in this study), the replacements might not be entirely appropriate.

Additionally, **differences in tokenisation** between the anchored terms and their replacements could cause minor issues.

Moreover, this technique is most effective for content rich in idiosyncratic expressions, such as the Harry Potter books. **The method may be less effective for other types of content, such as textbooks or nonfiction works**, which may lack these distinctive features.

Another limitation is that the unlearning process may inadvertently remove related content, such as articles discussing the *Harry Potter series*, which are external to the copyrighted books. To mitigate this, the researchers suggest fine-tuning the model on related content to ensure it regains lost knowledge.

In conclusion, this technique offers a promising approach to unlearning specific content from LLMs, particularly in domains rich in unique expressions. While it faces some challenges related to tokenisation and bias in alternative generation, its ability to unlearn targeted content without significant loss of general model performance makes it a valuable tool for managing copyright and data privacy concerns in LLMs.

More details can be found in *Annex XXI*.

Annex XIX: Sharded Isolated Sliced and Aggregated (SISA) Unlearning

Sharded, Isolated, Sliced and Aggregated training (SISA) is a **training dataset partition** method proposed by Bourtoule et al. in 2020.

In this method unlearning is achieved through re-training. The objective is to **make it quicker to re-train the model** when an unlearning request has to be fulfilled. This adapted concept has the big advantage that, once the data points to be forgotten are well identified and removed, it ensures the effectiveness of the unlearning procedure since those are not present in the training dataset anymore.

XIX.1 SISA: How It Works

The basic idea is to split the training dataset into **partitions** and to train a different model (hereafter called "**sub-model**") on each of those. Subsequently, when using the AI system, the contributions of the single sub-models are aggregated at inference time by using a voting system to produce the single final output.

So, when an unlearning request comes, after the data point to be forgotten has been removed from the training dataset, **only the sub-model related to the affected partition has to be re-trained**. This speeds-up the re-training process, since it is performed on a reduced amount of data.

The performance can be additionally improved by **further slicing** each partition. This operation allows **training the sub-models in an incremental way:** at each iteration a slice of data is added to the partition, saving the resulting sub-model's parameters before introducing the next slice. By doing this, the re-training of the sub-model can start from the last configuration before the slice containing the data to be unlearned was added.

The resulting architecture is represented in *Figure XVIII.1-1*, which highlights the different sub-models and the aggregation of their outputs at inference time.

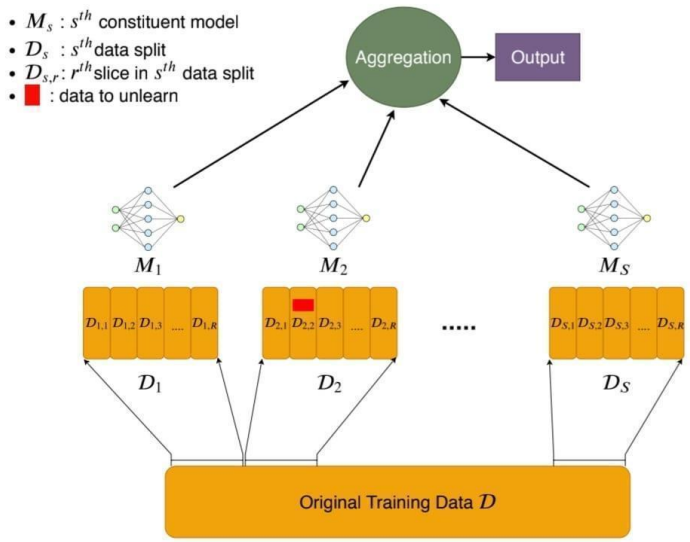


Figure XVIII.1-1: Final architecture of a GenAI model trained with the SISA approach (Bourtoule et al., 2020).

Overall, the system composed of multiple sub-models (in the study, they are referred to as "weak learners") tends to be **less accurate** than a single model trained on the entire dataset. This is because partitioning the data can **disrupt the statistical relationships between data points across different partitions**, making them less effectively accounted for during the model functioning. Moreover, the sub-models have the risk of **overfitting the smaller training partitions** and the aggregation operation only partially compensates for those effects.

XIX.2 SISA: Evaluation

The researchers declared that the proposed training procedure is a better **trade-off between accuracy and time required to unlearn**. They compared SISA training performances considering the naive approach of re-training from scratch as the baseline.

First, they performed **simple learning tasks**, such as deep networks trained on Purchase ([473]) and SVHN ([474]) datasets. In this setup, when processing 8 and 18 batched unlearning requests on Purchase and SVHN respectively, they measured a **speed-up of 4.63x and 2.45x ([475]) in re-training** with a **nominal degradation in accuracy of less than 2 percentage** points.

Moreover, they observed a steep **degradation in accuracy when the number of dataset partitions increases** over a threshold; meanwhile, the number of slices of each shard doesn't affect accuracy as long as the number of epochs ([476]) required for training are recalibrated.

To assess the effectiveness of the SISA approach in **complex training tasks**, the researchers utilised the ImageNet dataset alongside deeper neural networks. Their findings revealed a **significantly greater decline in accuracy compared to simpler training** scenarios. Unsurprisingly, the accuracy deteriorated further as the number of shards increased or when the proportion of data points to be unlearned surpassed a critical percentage of the total dataset size. Those effects are mitigated by the great size of training datasets used by the organisations to train their models.

However, an important finding was that this **accuracy gap can be reduced, for complex learning tasks, with transfer learning**. Indeed, in the real-world the common approach is to utilise a base model trained on public data and then utilise transfer learning to customise it towards the task of interest. Additionally, in the transfer learning setting, the time analysis for unlearning still holds.

All those considerations were based on the assumption of not **knowing the distribution of the data points to be unlearned**. But they further presented a refined variant of the approach, which assumes prior knowledge of that distribution. Taking inspiration from a Google's study (Bertram et al., 2019), they modelled a company operating across multiple jurisdictions with varying legislation

([473]) The Purchase dataset is a benchmark dataset commonly used in privacy and machine learning research. It contains simulated purchase records of individuals across various product categories. Each record represents a binary vector indicating whether a specific item was purchased, making it ideal for studying consumer behaviour, recommendation systems, and privacy-preserving data analysis.

([474]) The SVHN (Street View House Numbers) dataset is a real-world image dataset widely used in computer vision and machine learning. It consists of over 600,000 images of house numbers captured from Google Street View. Each image contains a digit (0-9), often part of a sequence, and is labelled for digit classification tasks.

([475]) Experimental time measurement is challenging due to hardware and software variability. To address this, the researcher of this study declared to have estimated unlearning time indirectly via the number of retraining samples. Controlled experiments confirmed a linear relationship between re-training samples and the time required for the procedure.

([476]) A training epoch is one complete pass through the entire training dataset during model training. Multiple epochs are typically used to improve the model's performance.

and sensitivities to privacy, and accordingly varying distributions of unlearning requests. Knowing this distribution **enables to further decrease expected unlearning time by strategically assigning to partitions and slices the training points that will likely need to be unlearned**. The resulting cost in terms of accuracy is either null or negligible (compared to the distribution-unaware configuration) and the number of data points to be re-trained is reduced.

XIX.3 SISA: Limitations

- Sharding and slicing may require the model trainer **revisits some hyperparameters choices**. For instance, it may require training with a different number of epochs. **As the number of sub-models increases, performing hyperparameter tuning becomes a challenging problem** due to the increasing quantity of factors to be taken into consideration;
- However, the researchers noted that this can be **mitigated** by uniformly splitting the data across shards, since then **hyperparameters are the same among the different sub-models**;
- Slicing could also interact with data batching during training in case slices are smaller than the batch size;
- Overall, the training procedure is more complex because the trade-off between model's accuracy and unlearning time has to be carefully studied.

Annex XX: Stable Sequential Unlearning (SSU)

Removing copyrighted data from models requires a balance between unlearning targeted content and **retaining the general and reasoning capabilities** of the model. The researchers pointed out that existing approaches, such as Gradient Ascent (GA), often lead to catastrophic collapse, significantly harming the model's reasoning ability and its knowledge base. Moreover, Task Vector-based approaches fail to optimise the trade-off between unlearning efficacy and maintaining the knowledge of non-targeted content.

XX.1 SSU: How it Works

SSU introduces a structured framework to **sequentially** unlearn copyrighted content. It uses task vectors to adjust specific weights associated with the data to be forgotten. Moreover, it combines fine-tuning with strategies that retain model integrity:

- **Random Labeling Loss** introduces controlled noise to prevent overfitting;
- **Weight Saliency Mapping** identifies and modifies weights to be adjusted for forgetting targeted content while preserving general knowledge.

Saliency Mapping is introduced in SSU because modifying too many weights can harm the model's general knowledge and reasoning abilities. Indeed, it differs from other unlearning methods:

- **Gradient Ascent (GA)**:
  Adjusts weights indiscriminately, often leading to catastrophic forgetting;
- **Task Vectors**:
  Modifies the model in a more structured way but does not localise updates to relevant weights, risking collateral degradation.

Using the original model instead of previously modified models in the **Stable Sequential Unlearning (SSU)** framework is a key strategy to ensure stability and avoid compounding errors during sequential unlearning.

XX.2 SSU: Evaluation

The validation was conducted on the Llama3-8B model: SSU was tested by unlearning four copyrighted books sequentially.

To verify the accuracy of the unlearning process, they leveraged the phenomenon of content memorisation—where a GenAI model can reproduce portions of its training data either verbatim or in a closely similar form (see *Section 3.2* for more details). To test whether a book has been effectively unlearned, the researchers use prompts derived from the original text (e.g., the first 200 tokens of a chunk from the book) and compare the generated continuations with the actual next 150 tokens of the book. In particular, they choose to use Jaccard ([477]) and ROUGE ([478]) scores for evaluating the similarity between the two.

Meanwhile, to check the model's capability retention, they used the performance measures obtained when interacting with MathQA and MMLU benchmark datasets (see *Section 4.1.1.2* for more details about benchmark datasets).

Results demonstrated that SSU achieved a better balance in **unlearning** copyrighted material while preserving **reasoning** and **knowledge** compared to baseline methods. To assess the performances considering those three different aspects, the training dataset was partitioned into:

- Books semantically similar to the books to forget (**Dss**).
- Books semantically dissimilar to the books to forget (**Dsd**).
- Books specifically included to maintain knowledge (**Dnor**).

Unlearning: For books to forget, **SSU** consistently reduced Jaccard and ROUGE-L scores closer to random baseline levels, indicating effective forgetting. **Baseline methods** like Gradient Ascent (GA) variants often failed due to catastrophic forgetting or incomplete unlearning.

For example, at the first-time step (unlearning *Harry Potter*), SSU achieved:

([477]) The Jaccard score for text similarity measures the overlap between two sets of words or tokens by dividing the size of their intersection by the size of their union. It ranges from 0 (no similarity) to 1 (identical sets).

([478]) The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score evaluates text similarity by comparing n-grams, word sequences, or word overlaps between a generated text and a reference text. Common variants include ROUGE-N (n-gram overlap), ROUGE-L (longest common subsequence), and ROUGE-W (weighted longest common subsequence).

- **Jaccard**: 0.09
- **ROUGE-L**: 0.125

These scores were significantly closer to the random baseline than the original model's scores.

Knowledge retention: It is the model's ability to retain knowledge of unrelated content (e.g., books not in the unlearning dataset). **SSU** outperformed baseline methods in preserving knowledge for books in **Dnor**, **Dss**, and **Dsd**. **Catastrophic forgetting** was common in GA-based methods after multiple unlearning steps. Meanwhile, if compared with the Task Vector (TV) approach, they measured that:

- The retention for **Dnor** obtained through SSU was 26% better than TV at the fourth time step;
- For semantically similar books (**Dss**), SSU reduced unintended forgetting, retaining 35% higher Jaccard and 47% higher ROUGE-L scores than the TV baseline at later steps.

Capability Retention: Since it measures the impact of unlearning on the model's ability to perform reasoning and general knowledge tasks, they performed tests on MathQA ([479]), MMLU ([480]) and GPQA ([481]) benchmark datasets. **SSU maintained strong performance** across benchmarks, avoiding the catastrophic performance drops seen in GA-based methods.

SSU Results at Step 4:

- MathQA: **34.3%** (compared to 32.1% for TV).
- MMLU (0-shot): **54.3%** (compared to 50.7% for TV).
- GPQA: **30.1%** (compared to 24.3% for TV).

([479]) MathQA is a benchmark dataset related to mathematical reasoning.

([480]) MMLU is a benchmark dataset assessing multitask knowledge.

([481]) Graduate-level Google-proof Q&A (GPQA).

Annex XXI: Idiosyncratic Expression Replacement for Unlearning

The unlearning approach discussed in the previous chapter also underpins the study conducted by Eldan and Russinovich in 2023. However, their work introduces a unique technique called idiosyncratic expression replacement. This method, as demonstrated in their study, is particularly effective in enabling Large Language Models (LLMs) to unlearn specific literary works, such as the entire *Harry Potter* book series.

Since the model's weights are arranged to perform the unlearning operation, this solution falls under the category of "approximate unlearning" (Zhang, Finckenberg-Broman, et al., 2024).

XXI.1 How it Works

The technique consists of three main steps (Eldan & Russinovich, 2023):

1. The **Task Vectors** are computed;
2. **Idiosyncratic expressions** in the target data are **replaced with generic counterparts**, and the model's own predictions are leveraged to generate alternative labels for every token. These labels aim to approximate the next-token predictions of a model that has not been trained on the target data;
3. The model is **fine-tuned on these alternative labels**, which effectively erases the original text from the model's memory.

Step 2 focuses on finding a counterpart token which answers the question: *"What would a model that has not been trained on the unlearning data have predicted as a next token in this sentence?"* Those **generic predictions are obtained combining two complementary approaches** (Eldan & Russinovich, 2023):

- **Reinforcement Bootstrapping:** the baseline model undergoes additional **training on the unlearn target** to create a **reinforced model**. Subsequently, tokens whose probabilities did not increase during this reinforcement process are identified and used to construct the generic prediction. The underlying principle is to identify tokens unrelated to the unlearn data.

- However, in many cases, when the model is primed with a specific idiosyncrasy (such as the names of one of the major characters), completions specific to the target text already have a high probability and it appears that reinforcing the model makes almost no difference. For this reason, this reinforcement-based technique has been integrated with the subsequent approach.
- **Anchored Terms:** they provided GPT-4 with random passages of the text and instructed it to extract a list of expressions, names or entities which are idiosyncratic. For each, a generic alternative, that would still be suitable in terms of text coherence, was asked. By iterating this, they built a **dictionary containing the generic version of** about 1,500 **anchored terms** from the unlearn target, i.e., the *Harry Potter*'s book. The main principle is to go over **each block of text from the unlearn target, replace the anchor terms by their generic counterparts** and then **process the resulting text with the baseline model's forward function to obtain next-token predictions.**

The researchers also tested those approaches separately, finding that the combination of the two produces better unlearning performances.

XXI.2 Evaluation

The researchers evaluated the technique on the task of unlearning the *Harry Potter* books from the **Llama2-7b** model (a generative language model recently open-sourced by Meta). They successfully erased the model's ability to produce or reproduce Harry Potter-related content. While the model took over 184K GPU-hours to pretrain, they achieved this result using only 1 GPU hour of fine-tuning. In the following Figure are reported some examples highlighting the different responses the model generates, demonstrating the effectiveness of unlearning.

Prompt	Llama-7b-chat-hf	Finetuned Llama-7b
Who is Harry Potter?	Harry Potter is the main protagonist in J.K. Rowling's series of fantasy novels...	Harry Potter is a British actor, writer, and director...
Harry Potter's two best friends are	Ron Weasley and Hermione Granger. In the series...	a talking cat and a dragon. One day, they decide...
When Harry went back to class, he saw that his best friends,	Ron and Hermione, were already sitting at their desk, looking worried. "What's wrong?"...	Sarah and Emily, were already there, sitting at their desks. "Hey, guys!"...

Figure XXI.3-1: Some examples of input prompt and output generation pairs after unlearning (Eldan & Russinovich, 2023).

Meanwhile, the model's **performance on common benchmarks** (such as Winogrande, Hellaswag, Arc, Boolq and Piqa) **remains almost unaffected**: this demonstrates that the procedure hasn't affected the reasoning abilities of the baseline model.

XXI.3 Limitations

The process of replacing anchored terms could introduce bias if the **LLM used for the generation of alternatives was itself trained on the unlearn target**. In fact, depending on the used model's knowledge of the unlearn target, the proposed alternatives would be appropriate.

Moreover, there are several additional **caveats related to the way the text is tokenised**: the anchored terms' translations do not necessarily have the same number of tokens. The researchers studied those issues and proposed mitigations in their study.

On the other hand, the researchers recognised their technique is likely to exhibit **limitations with other types of content (such as non-fiction or textbooks)**. In fact, the *Harry Potter* books are replete with idiosyncratic expressions and distinctive names—traits that, in hindsight, may have abetted the unlearning strategy.

Finally, this technique may result in the model **unlearning a superset of the intended unlearning target**. For example, using the Harry Potter books as the unlearning target may cause the model to forget related Wikipedia articles and other training data discussing the books as an unintended consequence. As a mitigation, the researchers propose fine-tuning the model on any related content in order to **re-learn it**.

Annex XXII: Considerations on MEND with Respect to the Software Qualities Highlighted by the AI Act

XXII.1 MEND: Versatility

The researchers anticipate that future work may extend MEND beyond transformer models, enabling its use for a broader range of edits, including non-text-based content (Mitchell et al. 2022).

XXII.2 MEND: Openness & Market Maturity

The entire project has been open-sourced on GitHub ([482]) where it has gained notable interest from the developer community within three years of its publication.

XXII.3 MEND: Scalability & Cost Implications

Tests on large-scale models such as T5, GPT, BERT, and BART demonstrate that MEND effectively edits models with over 10 billion parameters. Even with those large models, the process of **setting up MEND** is efficient: it can be trained on a **single GPU in less than a day**.

Even the tests conducted on batched editing—a more realistic setting, when multiple simultaneous editings are needed—demonstrated a good editing success.

XXII.4 MEND: Reliability

The researchers identified one main limitation: the extent to which an edit performed on a single input-output pair correctly influences related prompts. Indeed, they recognised the difficulty of finding all the possible related input requests to properly assess if the model's knowledge was effectively updated.

([482]) Eric-Mitchell / Mend, Github, 10 February 2025 (accessed 14 March 2025).

Annex XXIII: Considerations on SERAC with Respect to the Software Qualities Highlighted by the AI Act

XXIII.1 SERAC: Scalability

When testing with an **increasing number of edits**, SERAC's superiority becomes clear compared with other methods, confirming the enhanced **scalability** of this solution. In *Figure XXIII.1-1* the difference between ES and DD (i.e., ES minus DD) is plotted against the number of edits, demonstrating that SERAC achieves better scalability than ENN and MEND.

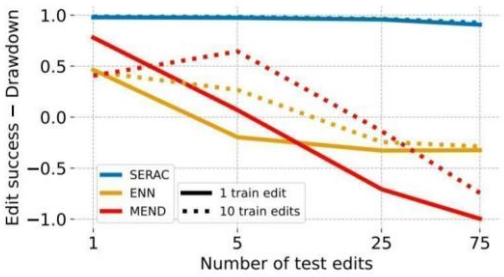


Figure XXIII.1-1: Diagram showing how the performance of SERAC remains unaltered after an increasing number of edits. The values obtained by subtracting DD from ES are compared with the ones obtained through MEND (discussed earlier) and ENN. A higher score means better capacity to perform the editing while maintaining locality (Mitchell, Lin, Bosselut, Manning, et al., 2022).

XXIII .2 SERAC: Openness & Market Maturity

The project is **open-source** and hosted on GitHub ([483]). Over approximately three years since its publication, it has garnered less attention from the developer community compared to MEND, which was discussed earlier.

([483]) Eric-Mitchell / Serac, Github, 21 November 2024 (accessed 14 March 2025).

XXIII .3 SERAC: Interoperability

Furthermore, differently from editing methods developed prior, SERAC **can be integrated with each GenAI model without further training** outside the initialisation. In particular, the scope classifier and counterfactual model are trained completely separately on an editing dataset. This dataset is itself unrelated to the actual edits applied after the GenAI system's deployment and stored in the explicit memory discussed before.

XXIII.4 SERAC: Cost Implications

SERAC may introduce some additional **computational overhead** due to the inclusion of the scope classifier and the counterfactual model. However, it employs a **nearest-neighbour-based classifier** that operates at a speed comparable to the base model, ensuring that the overall processing time does not increase significantly. Furthermore, the **counterfactual model** is smaller than the base model, enabling faster response times when handling requests related to an edit record, as these are processed by this secondary model.

SERAC's **additional memory consumption primarily arises from the weights of the classifier and counterfactual model**, resulting in an approximate doubling of the storage requirements for the overall infrastructure. However, the majority of this increase constitutes a **fixed cost** that remains unchanged regardless of the number of edits. Notably, each edit record requires only 3KB of storage, which is several orders of magnitude smaller than the base model.

Anyway, a limitation persists: in a setting where **editing occurs continuously**, the edit memory may grow without bound.

XXIII.5 SERAC: Versatility

During the development of SERAC, the study concentrated only on **textual content**. However, when evaluating the potential expansion of its application to text-to-image and text-to-video generative models, it is essential to consider the increased memory requirements for storing edit records. This

growth in storage demand could introduce scalability challenges, potentially affecting the system's efficiency and feasibility in large-scale implementations.

XXIII.6 SERAC: Reliability

The researchers developed a new method to enable **more rigorous evaluation** of model editors, which proposes three challenging language model editing problems: question answering, fact-checking and dialogue generation.

By using the proposed method, they evaluated SERAC, demonstrating its superior performance on all three tasks and consistently outperforming other model editing approaches available at the time of the study (2022). In *Figure XXIII.6-1* the results of the assessment are compared, where the metrics adopted are:

- **Edit Success (ES)**, which measures the effectiveness of the edit in all the outputs related to an edit record. It ranges from 0 to 1, where higher values indicate greater effectiveness; and
- **Draw Down (DD)**, which measures if the edits achieved the desired locality by influencing only the outputs related to their related inputs. It ranges from 0 to 1, where lower values indicate greater effectiveness.

They evaluated SERAC's performance on **various benchmark datasets**—including QA and QA-hard (see *Section 4.1.1.2* for the discussion on benchmark datasets)—and with **different base models**. Among the baseline methods compared to the new approach was the simple fine-tuning approach, as well as memory-based editing methods such as **MEND** (discussed earlier) and Editable Neural Networks (ENN) (Sinitsin et al., 2020). The results are reported in *Figure XXIII.6-1.*

Dataset	Model	Metric	FT	LU	MEND	ENN	RP	SERAC
QA	T5-large	↑ ES	0.572	0.944	0.823	0.786	0.487	**0.986**
		↓ DD	0.054	0.051	0.187	0.354	0.030	**0.009**
QA-hard	T5-large	↑ ES	0.321	0.515	0.478	0.509	0.278	**0.913**
		↓ DD	0.109	0.132	0.255	0.453	**0.027**	**0.028**
FC	BERT-base	↑ ES	0.601	0.565	0.598	0.594	0.627	**0.877**
		↓ DD	**0.002**	**0.01**	0.021	0.042	**0.01**	0.051
ConvSent	BB-90M	↑ ES	–	–	0.494	0.502	0.506	**0.991**
		↓ DD	–	–	2.149	3.546	**0**	**0**

Figure XXIII.6-1: Comparison between SERAC and other editing approaches, performed on different combinations of benchmark datasets and base models. Both the metrics ES and DD have been measured when performing 10 simultaneous edits. Some of the reference approaches are: Fine-Tuning (FT), Editable Neural Networks (ENN) (Sinitsin et al., 2020), and MEND, discussed earlier (Mitchell, Lin, Bosselut, Manning, et al., 2022).

THE DEVELOPMENT OF GENERATIVE ARTIFICIAL INTELLIGENCE FROM A COPYRIGHT PERSPECTIVE

TB-01-25-001-EN-N ISBN: 978-92-9156-369-2 DOI: 10.2814/3893780