THE DEVELOPMENT OF GENERATIVE ARTIFICIAL INTELLIGENCE FROM A COPYRIGHT PERSPECTIVE

EUIPO
EUROPEAN UNION
INTELLECTUAL PROPERTY OFFICE

Over the past number of years Artificial Intelligence (AI) technologies have undergone major advances, most notably with the release of Large Language Models and Generative AI (GenAI) systems. GenAI services that generate text, code, images, videos, 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.

This study explores the developments in 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**. The main findings are:

- **Access to high-quality content is central to the development of GenAI services.** The AI training process is complex and uses content as input at different stages. However, as GenAI models are "specialised" for certain functionalities they need access to high quality and up-to date content, which is reflected in emergence of a direct licensing market, with some GenAI developers licensing access and use of high-quality content from copyright holders. The capacity for copyright holders to effectively reserve their rights a pre-requisite for the licensing market to develop.

- **No 'one-size-fits all' solution for copyright holders to protect their rights has emerged yet.** Instead, different approaches and solutions are developing for copyright holders to protect their rights, and for AI developers to respect their regulatory obligations: On the one side, the rights reservation mechanisms for the **INPUT phase** (related to training AI models), whereby rightsholders can express their opt out from the 'text and data mining' (TDM)-exception. On the other side, transparency measures exist for the **OUTPUT phase** that allow the indication and recognition of AI generated content.

- **Public authorities**, such as national IP authorities and the EUIPO, **may play a role** by providing technical support (for copyright holders to reserve their rights, and for AI developers to effectively respect such reservations) as well as non-technical support (e.g., public awareness, forums for technical information sharing, providing information to the public on available solutions, trends and developments).

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.

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 c**ompliance 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 opt-outs, 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 ISCC code 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 r**espective 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 **role 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.[1]

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.

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).

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

References

[1] A membership inference attack allows an adversary to query a trained machine learning model to predict whether or not a particular example was contained in the model's training dataset