Dornis, Tim W.; Lucchi, Nicola

Generative AI and the Scope of EU Copyright Law: A Doctrinal Analysis in Light of the Referral in Like Company v. Google

IIC - International Review of Intellectual Property and Competition Law

Provided in Cooperation with:
Springer Nature

Suggested Citation: Dornis, Tim W.; Lucchi, Nicola (2025) : Generative AI and the Scope of EU Copyright Law: A Doctrinal Analysis in Light of the Referral in Like Company v. Google, IIC - International Review of Intellectual Property and Competition Law, ISSN 2195-0237, Springer, Berlin, Heidelberg, Vol. 56, Iss. 10, pp. 1800-1840, https://doi.org/10.1007/s40319-025-01649-7

IIC (2025) 56:1800–1840
https://doi.org/10.1007/s40319-025-01649-7

ARTICLE

Generative AI and the Scope of EU Copyright Law: A Doctrinal Analysis in Light of the Referral in *Like Company v. Google*

Tim W. Dornis · Nicola Lucchi

Accepted: 20 October 2025 / Published online: 26 November 2025

Abstract This article offers a doctrinal analysis of the copyright implications raised by *Like Company v. Google Ireland* (C-250/25), the first case to bring generative AI before the Court of Justice of the European Union. It examines whether the training and output of systems like Gemini infringe exclusive rights under EU copyright law. We argue that AI model training may involve acts of reproduction under Art. 2 of the InfoSoc Directive, while the dissemination of AI-generated outputs, especially through public interfaces, may trigger the right of communication to the public under Art. 3. Particular concerns arise when protected content is recognisably reproduced or when AI outputs serve as functional substitutes for original works, thereby affecting the normal exploitation of those works. While not a formal infringement criterion, such functional substitution is relevant in assessing the application of exceptions and compliance with the three-step test. The paper also challenges the applicability of the text and data mining exception to generative uses, highlighting its incompatibility with the limitations imposed by the three-step test. Ultimately, the analysis supports a technologically neutral, rights-based interpretation that safeguards the economic viability of creative production in the algorithmic age.

Keywords Artificial intelligence · Model training · Copyright · Generative AI

Tim W. Dornis and Nicola Lucchi are listed alphabetically and have contributed equally to the manuscript.

T. W. Dornis (✉)
J.S.M. (Stanford); Professor of Private Law and Intellectual Property Law (Leibniz University) & Global Professor (NYU School of Law), Hannover, Germany
e-mail: tim.dornis@jura.uni-hannover.de

N. Lucchi
Ph.D.; Serra Húnter Professor of Comparative Law, University Pompeu Fabra, Barcelona, Spain
e-mail: nicola.lucchi@upf.edu

1 Introduction

This article, albeit not conceived as an amicus curiae brief, may be used as such: It offers legal observations intended to assist the Court of Justice of the European Union (CJEU) in its preliminary assessment of the legal questions raised in Case C-250/25, *Like Company v. Google Ireland*.[1] The case represents a pivotal juncture in the evolution of EU copyright law,[2] as it marks the first occasion on which the Court is called upon to directly assess the compatibility of generative artificial intelligence (AI) systems with the *acquis Communautaire* on copyright and neighbouring rights.

The legal questions raised in this case transcend its specific factual context. They reflect a growing tension between the exponential development of generative models and the foundational principles of EU copyright, including authors' exclusive rights, the three-step test, and the economic viability of content production.[3] This case thus offers the Court a critical opportunity to clarify the limits of lawful AI deployment under existing copyright law.

On 3 April 2025, the Budapest Környéki Törvényszék (Hungary) referred four preliminary questions to the CJEU under Art. 267 TFEU.[4] The dispute stems from a complaint brought by Like Company, a Hungarian press publisher, against Google Ireland Limited, a subsidiary of Alphabet Inc. Like Company operates several news portals monetised through advertising and protected by copyright. It alleges that Google's chatbot Gemini (formerly Bard) unlawfully reproduced and made available significant portions of its editorial content. The triggering factual scenario involves Gemini's generation of a detailed summary of a Like Company article about Hungarian singer Kozsó's plans to introduce dolphins to Lake Balaton. This content was generated in response to user prompts between June 2023 and February 2024 and allegedly included verbatim or near-verbatim reproductions of protected text.

The national court seeks guidance on the interpretation of Arts. 2 and 3(2) of the Copyright and Information Society (InfoSoc) Directive,[5] Art. 15(1) and Art. 4 of the

[1] *See* CJEU, case C-250/25, *Like Company v. Google Ireland*, preliminary reference lodged on 3 April 2025. Referral from Fővárosi Törvényszék (Budapest Metropolitan Court), Hungary. Available at https://curia.europa.eu/juris/liste.jsf?num=C-250/25&language=en.

[2] The term "European Union copyright law," as used here, is a convenient shorthand rather than an accurate reflection of a unitary legal system. Copyright in the EU remains primarily governed by the national laws of the 27 Member States. The EU's role has been limited to harmonising selected aspects of these national systems through a series of directives, resulting in a partially convergent – yet still fragmented – legal framework.

[3] *See e.g.* Dusollier et al. (2025), calling for a balanced approach that safeguards authors' rights while enabling responsible AI development.

[4] Consolidated Version of the Treaty on the Functioning of the European Union, 30 March 2010, 2010 O.J. (C83) 47.

[5] *See* 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, Official Journal of the European Communities 2001, L 167, 10. While Art. 3(2) of the InfoSoc Directive grants the right of communication to the public to certain related-rights holders (such as performers, producers, and broadcasters), it is Art. 3(1) that confers this exclusive right on authors. The Hungarian court's

Copyright in the Digital Single Market (CDSM) Directive,[6] and how these apply to:

1. the communication to the public of press content via chatbot outputs;
2. the classification of LLM training as a form of reproduction;
3. the applicability of the text and data mining (TDM) exception to commercial AI training;
4. the liability of the AI provider for outputs that reproduce protected content.

The claimant argues that Gemini's training and output amount to unlawful reproduction and communication to the public, exceeding the permitted use under Art. 15(1) CDSM Directive. It maintains that the text and data mining (TDM) exception under Art. 4 CDSM Directive cannot shield the training because, despite the commercial nature of the use, rights holders lacked a realistic opportunity to exercise the opt-out mechanism provided by the Directive. Like Company also asserts that Gemini's outputs, which substitute user access to its websites, undermine its economic model.

Google Ireland disputes these claims. It argues that Gemini does not store or retrieve copyrighted content but rather uses probabilistic modelling and tokenisation. According to the defendant, any resemblance to the original articles is incidental or the result of hallucination. It invokes the exceptions for temporary reproduction and text and data mining, and argues that no new public is reached. This defense draws on CJEU case law, which typically applies the "new public" criterion when a protected work, initially made available with the rightsholder's consent to a specific audience, is subsequently communicated to a different audience (e.g. via hyperlinking or retransmission). In this case, the original articles were publicly available online, and Google argues that chatbot users form part of the general internet audience already entitled to access them. However, as this paper argues, the mode of access via AI-generated outputs that bypass the publisher's website and monetisation model effectively targets an unanticipated audience under different conditions. This may support a 'new public' finding under Art. 3 of the InfoSoc Directive.

While these arguments focus on the legality of Gemini's operations, the case also raises broader concerns about how generative AI may affect the creative economy. Many rightsholders, in particular press publishers, such as the claimant in this case, rely on digital traffic and licensing income to finance their creative and journalistic production. When AI models reproduce key expressive elements of their work in response to user prompts, they not only undermine the economic value of copyright, but also distort the incentives underpinning the creative industries. From an economic perspective, such outputs function as substitutes, particularly when they

Footnote 5 continued
preliminary reference cites Art. 3(2), possibly due to the claimant's status as a press publisher. However, press publishers' specific related right of making content available is now governed by Art. 15(1) of the CDSM Directive. This paper acknowledges the overlap and doctrinal ambiguity and addresses the legal implications for both authors' and publishers' rights.

[6] *See* 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, Official Journal of the European Communities 2019 L 130, 92.

offer summaries or paraphrased versions of protected content without attribution or licensing. This may raise concerns under the three-step test (Art. 5(5) InfoSoc Directive), in particular since it could undermine the normal exploitation of the work and prejudice the legitimate interests of rights holders. In this light, copyright law must ensure that innovation in AI does not come at the cost of disincentivising cultural and journalistic production.

This paper is structured as follows. Section 2 introduces the technical and legal distinction between training-based and retrieval-augmented generative AI systems, a premise essential to framing the legal questions. Section 3 then addresses the four preliminary questions referred to the Court. Each subsection provides a legal analysis of the relevant provisions of the InfoSoc and CDSM Directives, supported by CJEU case law and doctrinal sources. Specifically, the paper examines (i) whether chatbot responses constitute unauthorised communication to the public, (ii) whether the training of large language models entails acts of reproduction, (iii) whether such training qualifies under the TDM exception in Art. 4 CDSM Directive, and (iv) whether AI-generated outputs containing protected content infringe upon exclusive rights. The paper concludes by synthesising the legal findings and identifying their implications for the future application of EU copyright law to generative AI systems.

2 The Training/Retrieval Distinction: A Legal and Technological Premise

Before turning to the legal questions raised in this case, it is essential to clarify a key technical and legal distinction that underpins the factual background, namely the difference between model training[7] and retrieval-augmented generation (RAG).[8] Preliminary information suggests that the disputed output generated by Google's AI service, Gemini, was not the result of internalised training data, but rather produced in real time in response to user prompts by accessing online content through external sources. This points to the potential use of a RAG-type architecture.

2.1 What Is RAG and Why Is It Different from AI Training?

In technical terms, RAG systems differ fundamentally from traditional machine learning approaches. While model training entails the ingestion, reproduction, and transformation of expressive works to adjust internal model parameters, RAG systems operate differently. They retrieve and integrate external information at the

[7] Model training refers to the process by which a machine learning system adjusts its internal parameters (such as weights in a neural network) by ingesting large datasets and minimising a loss function through iterative optimisation. This process enables the model to learn statistical patterns from the input data and generalise to unseen inputs. *See* Goodfellow et al. (2016), pp. 164–172.

[8] Retrieval-Augmented Generation (RAG) enables generative AI models to access external data sources – such as online encyclopedias, websites or databases – at the time of a query, incorporating retrieved information into their responses. This allows them to generate context-relevant outputs without requiring prior training on the referenced material. *See* Lewis et al. (2020), pp. 9459–9460.

inference stage, typically through API calls or live web queries, without incorporating this data into the model weights.[9]

This distinction is not merely academic: it may carry significant legal implications. Retrieval-based architectures may, in some circumstances, fall more clearly within the scope of EU copyright exceptions, particularly the TDM exception under Art. 4 of the CDSM Directive, or the temporary reproduction exception under Art. 5(1) of the InfoSoc Directive.[10] This is especially relevant when the retrieved data is not persistently stored or reused, but rather processed on-the-fly to produce user-specific outputs. Reflecting this divergence, licensing practices for RAG systems have evolved differently from those applicable to training corpora, indicating broader stakeholder awareness of the legal and technical distinctions involved. Whereas large-scale training datasets have traditionally been compiled without explicit rights clearance, retrieval-augmented models often access indexed content from curated or licensed databases. This reflects an emerging industry norm: stakeholders increasingly treat RAG outputs as content uses requiring authorisation. For instance, several technology providers, including OpenAI and Google, have pursued licensing agreements with publishers (e.g., Axel Springer, The Associated Press) to enable access to their archives through APIs or integrated chatbot functionalities.[11]

2.2 What Are the Consequences?

The distinction between training-based and retrieval-augmented systems is therefore foundational to the legal assessment of AI-generated outputs. While RAG models may avoid certain forms of liability associated with training, they raise distinct and arguably more visible risks under EU copyright law, especially concerning temporary reproductions, unauthorised communication to the public, and output liability. From a copyright perspective, these are the specific implications that flow from this distinction:

2.2.1 Reproduction Right

Under traditional training paradigms, the reproduction right under Art. 2 of the InfoSoc Directive is directly engaged due to systematic copying and internalisation of protected content into datasets and model parameters. These reproductions are typically extensive and persistent, clearly constituting substantial acts of reproduction. In contrast, RAG systems do not store or internalise full corpora within the model itself. Instead, they query external sources and retrieve content as needed. Nonetheless, even in RAG systems, temporary acts of reproduction, such as

[9] *See*, *e.g.*, Patrick Lewis et al. (2020), pp. 9459–9474; further also US Copyright Office, Copyright and Artificial Intelligence – Part 3: Generative AI Training, May 2025 (pre-publication version), pp. 30–31.

[10] *See* Art. 5(1) of Directive 2001/29/EC [2001] O.J. L 167/10.

[11] *See*, *e.g.* O'Brien (2023); Coster (2023); Davies (2024); *see* Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data, Press release (11 July 2023). Available at https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year.

caching, pre-fetching, or embedding content into prompts, may fall within the scope of Art. 2. As established in *Infopaq*,[12] even partial reproductions may be protected if they involve substantial and identifiable parts of a work.[13] Therefore, although the training-stage risks are diminished in RAG models, reproduction may still occur at the retrieval or output stages, depending on the system's architecture and use.

However, the mere fact that these acts are transient does not automatically qualify them for exemption under Art. 5(1) of the InfoSoc Directive. According to CJEU case law, most notably *Public Relations Consultants Association*[14] and *Stichting Brein v. Wullems*,[15] the temporary reproduction exception is subject to a strict five-part test: the act must be (i) temporary, (ii) transient or incidental, (iii) an integral and essential part of a technological process, (iv) whose sole purpose is to enable a lawful use, and (v) that has no independent economic significance. In the RAG context, the reproduction of protected content, particularly when triggered by user prompts, may not consistently satisfy these cumulative conditions. While such operations may be technically transient (e.g., cached or formatted), they often serve the purpose of delivering expressive outputs to end-users and generating economic value, either directly or indirectly. These copies are therefore not merely neutral intermediaries; they may instead constitute infringing reproductions under Art. 2. In this light, the caching or formatting operations should not be viewed as technically necessary in the narrow sense required by Art. 5(1), but rather as integral to the user-facing service, thereby falling outside the exception's scope.

2.2.2 Communication to the Public and Making Available

The legal risks associated with the right of communication to the public under Art. 3 of the InfoSoc Directive are typically less pronounced in traditional training settings.[16] Unless a generative AI model is used to output protected content memorised during training, or unless its functionality or parameters, such as trained weights, are made publicly available, for instance through a downloadable file or a publicly accessible interface, the act of training itself is typically not considered a communication to the public or a making available.

By contrast, RAG systems introduce a heightened risk of infringement in this area. Because they can generate outputs that include verbatim or near-verbatim excerpts from the retrieved sources, particularly from press content, books, or news archives, they may fall within the scope of unauthorised "making available" to the

[12] CJEU, case C-5/08, *Infopaq Int'l A/S v. Danske Dagblades Forening*, ECLI:EU:C:2009:465, 2009 E.C.R. I-6569.

[13] It would be worth noting that "substantial" in EU law refers to qualitative significance (originality), not strictly quantitative length.

[14] CJEU, case C-360/13, *NLA*, ECLI:EU:C:2014:1195.

[15] CJEU, case C-527/15, *Stichting Brein v. Wullems*, EU:C:2017:300, para. 60.

[16] *See*, *e.g.* Buick (2025), pp. 184–185 (noting that training involves internal copying rather than public dissemination).

public.[17] The *Like v. Google* case illustrates this point directly: the content allegedly reproduced by Gemini was a journalistic article, and its inclusion in the generated output, if confirmed, raises precisely the kind of Art. 3 concerns the InfoSoc Directive was designed to address.

2.2.3 Limitations of the TDM Exception

The TDM exception under Arts. 3 and 4 of the CDSM Directive is fundamentally oriented towards analytical use rather than expressive reproduction.[18] Traditional training methods, due to their structural internalisation of content, already raise questions regarding compatibility with TDM exceptions.[19] With RAG systems, the tension is arguably even greater. Although initial retrieval processes in RAG architectures might superficially resemble TDM practices (as they involve automated access and processing of online data), the subsequent expressive reproduction of retrieved content clearly exceeds the analytical scope envisaged by TDM exceptions. RAG outputs are often designed to present coherent, expressive reproductions rather than purely analytical, non-expressive summaries.[20] Thus, while the initial retrieval might be temporarily permissible under TDM, the reproduction or presentation of expressive content in outputs may violate the scope of the TDM exception.

For example, a RAG-based legal assistant might retrieve a paragraph from a paywalled legal commentary or a copyrighted academic article and rephrase it only minimally before including it verbatim or nearly verbatim in its response to a user query. While the retrieval and indexing of such material might arguably fall within the scope of Art. 4 CDSM (if used for analytical purposes), the act of reproducing and communicating substantial expressive parts of the source material, especially in response to a user prompt, goes beyond what the TDM exception allows.

2.2.4 Output Liability

The distinction between traditional training of generative AI models and the use of RAG models is also relevant when assessing liability for infringing outputs. Traditional generative AI systems, during their employment, do actually reproduce protected materials that have been memorised during the underlying model's

[17] *See* EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (2025). Available at https://www.euipo.europa.eu/en/news/euipo-releases-study-on-generative-artificial-intelligence-and-copyright at 17–18 and § 4.5.1 on models creating infringing reproductions.

[18] *See* Lucchi (2025); Dornis and Stober (2024), p. 65. *See also* Rosati (2025), pp. 1–24 (concluding that unlicensed AI training is not fully covered by Arts. 3 or 4 DSMD and requires licensing).

[19] Lucchi (2025); Dornis and Stober (2024); Dornis (2025b).

[20] *See* EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective, cit. at 318 (referring to "content regurgitation" and stressing that some outputs reproduce chunks of protected content, particularly in press articles, news stories, and books).

training: a phenomenon known as "regurgitation" or "memorisation".[21] While this phenomenon is still debated for generative AI models, the risk of reproduction that comes with the utilisation in RAG systems, is uncontestably higher, since the systems' purpose and function is to retrieve and restate protected content verbatim, particularly where retrieval is deterministic and not sufficiently filtered or abstracted.[22] RAG systems, by design, systematically retrieve content at the moment of output generation, with the explicit aim of integrating external sources in real time. As a result, any outputs that are insufficiently abstracted or filtered, risk directly replicating protected material. This increases the risk of copyright infringement and, in turn, the potential liability for the service providers, especially when substantial portions of protected content are reproduced without prior authorisation or without satisfying the conditions of an applicable exception or limitation under EU law.

3 What Are the Questions Submitted to the CJEU and How to Approach the Issues?

3.1 Question #1: Does the Display of Content in Chatbot Responses Constitute Communication to the Public?

The first question referred to the Court asks whether chatbot responses that reproduce journalistic content, without authorisation or linking to the original source, can be qualified as acts of communication to the public under EU copyright law. This issue engages both Art. 15(1) of the CDSM Directive, which grants press publishers rights over the online use of their content, and Art. 3(2) of the InfoSoc Directive, which sets out the exclusive right of communication to the public. The legal analysis must assess whether such AI-generated outputs satisfy the two cumulative criteria developed by the CJEU: (i) an act of communication of a work, and (ii) a communication to a public. In many situations, particularly where the work was already made available with the author's authorisation, the Court further examines whether the act targets a "new public," meaning a public not originally contemplated by the rightsholder. This additional analysis helps determine whether the communication exceeds the scope of any prior authorisation.

3.1.1 Existence of an Act of Communication

The first requirement under Art. 3(2) of the InfoSoc Directive is the presence of an "act of communication." The Court of Justice has consistently interpreted this

[21] Empirical studies confirm that large models can and do memorise training data, supporting the inference that reproduction has occurred, even in the absence of a one-to-one match. *See*, *e.g.* Cooper and Grimmelmann (2025), pp. 48–49; Lee et al. (2022); Henderson et al. (2023), pp. 1–79; Lee et al. (2023), pp. 3637–3647; Carlini et al. (2023a, b); Hayes et al. (2025).

[22] Zhou et al. (2024), (supporting the concern that such systems may restate protected content verbatim); Ni et al. (2025), (warning that RAG systems may restate content retrieved from external sources without sufficient abstraction).

concept in expansive terms, emphasising the functional act of making a protected work perceptible to a public audience, regardless of the specific technological means by which that communication is carried out.[23] The emphasis lies not on the form of the transmission, but on the effect: whether the work is rendered accessible to individuals in a way that engages the exclusive rights of the copyright holder.

In *Reha Training*, the Court reaffirmed that an "act of communication" includes any transmission of protected works, irrespective of the technical means used, including transmissions that enable access at a time and place chosen by the user, as in on-demand services.[24] The Court's formulation aligns with the principle of technological neutrality, which requires that the scope of copyright protection not be undermined by the mere use of novel or automated transmission methods.[25] Accordingly, both user-initiated and system-initiated acts, whether human or machine-driven, fall within the functional definition of communication to the public, provided the work is rendered perceptible to a public audience. As clarified by the Court always in *Reha Training*, what matters is that the content is made available in a manner that allows members of the public to access it individually at a time and place of their choosing.[26]

When this framework is applied to the case at hand, it becomes difficult to see how the outputs of an AI-powered chatbot could fall outside this broad definition. Gemini's responses, as described in the preliminary reference, are triggered by user prompts and delivered through an interface that is openly accessible to the public. The content is not passively stored or merely indexed; it is generated and displayed in response to user engagement. This satisfies the core condition for communication, namely that the work is made perceptible to others through some form of transmission.

Nor does the automated or probabilistic character of the generation process displace this conclusion. The fact that the response is generated algorithmically, rather than manually curated is not, in itself, sufficient to remove it from the legal concept of "communication." On the contrary, the Court of Justice and its

[23] *See*, *e.g. SGAE v. Rafael Hoteles* EU:C:2006:764, [2007] Bus. L.R. 521 at [36; 43]; *SCF v. Del Corso* EU:C:2012:140, [2012] Bus. L.R. 1870; *ITV Broadcasting Ltd v. TV Catchup Ltd* (C-607/11) EU:C:2013:147, [2013] 3 C.M.L.R. 1; *Svensson v. Retriever Sverige AB* (C-466/12) EU:C:2014:76, [2014] 3 C.M.L.R. 4; *BestWater* (C-348/13) EU:C:2014:2315; *GS Media v. Sanoma* (C-160/15) EU:C:2016:644, [2017] 1 C.M.L.R. 30; *Stichting Brein v. Wullems* (C-527/15), EU:C:2017:300, [2017] 3 C.M.L.R. 30; *Circul Globus Bucureşti v. UCMR-ADA* (C-283/10) EU:C:2011:772.

[24] *See* CJEU, case C-117/15, *Reha Training Gesellschaft für Sport- und Unfallrehabilitation mbH v. Gesellschaft für musikalische Aufführungs- und mechanische Vervielfältigungsrechte (GEMA)*, ECLI:EU:C:2016:379, para. 38 (affirming that any transmission of the protected works, irrespective of the technical means or process used).

[25] CJEU case law frequently refers to this principle explicitly or implicitly (*e.g.*, CJEU, case C-607/11, *ITV Broadcasting Ltd v. TV Catchup Ltd*, EU:C:2013:147, para. 24). *See* also Recital 23 of the InfoSoc Directive affirming that "*any such transmission or retransmission … by wire or wireless*" qualifies, without differentiating the underlying technology.

[26] *See* CJEU, case C-117/15, *Reha Training Gesellschaft für Sport- und Unfallrehabilitation mbH v. Gesellschaft für musikalische Aufführungs- und mechanische Vervielfältigungsrechte (GEMA)*, ECLI:EU:C:2016:379, paras. 36, 37.

Advocates General have made clear, particularly in *Pelham*[27] and *Funke Medien,*[28] that the mere use of technological tools or platforms does not shield unauthorised uses of protected content from legal scrutiny under EU copyright law. Where the output remains recognisable and functionally equivalent to the protected work, or to a substantial part thereof, the act of making it available remains subject to the right of communication, regardless of the opacity or novelty of the system that enables it.[29] The emergence of AI-generated responses does not create a doctrinal void.[30] The essential inquiry remains whether protected material has been made accessible to the public. When such access is facilitated through an interactive system like Gemini, there is little doubt that the act of communication, as conceived in EU copyright law, has occurred.

3.1.2 Communication to a "New Public"

The second requirement under Art. 3(2) of the InfoSoc Directive asks whether the communication is addressed to a public that was not taken into account by the copyright holder when the original act of dissemination occurred. This "new public" criterion, first articulated by the CJEU in *Svensson,*[31] has become central to determining the unlawfulness of subsequent uses of protected works made available online.

In particular, in *Svensson*, the Court held hyperlinking to works freely accessible on the rightsholder's website does not amount to a communication to a new public, provided that the linked content was already made freely available with the rightsholder's consent and no technical restrictions were circumvented.[32] However, the Court clarified that linking which bypasses paywalls, registration walls, or other access restrictions may indeed reach a new public and thus constitute infringement. The key issue is not the method of access but whether the rightsholder consented to the specific audience being reached under the same conditions.

This logic applies with particular force in the context of generative AI. When a chatbot reproduces portions of a journalistic article, whether verbatim or in recognisable paraphrase, it may do so without respecting the original publishers' conditions of access, such as paywalls, registration requirements, or licensing schemes. Users engaging with the chatbot are not simply retrieving content already available online; they are accessing it through an alternative, AI-mediated channel that bypasses the mechanisms through which the publisher monetises its work. The

[27] *See* CJEU, case C-476/17, *Pelham GmbH v. Ralf Hütter and Florian Schneider-Esleben*, EU:C:2019:624 (rejecting the argument that the use of short audio samples is exempt from copyright scrutiny under the quotation exception).

[28] *See* CJEU, case C-469/17, *Funke Medien NRW GmbH v. Bundesrepublik Deutschland*, ECLI:EU:C:2018:870 (holding that unauthorised online publication of leaked military reports infringes copyright, notwithstanding the content's public interest character or prior disclosure).

[29] CJEU, case C-466/12, *Svensson v. Retriever Sverige AB*, EU:C:2014:76 paras. 18–28 (clarifying that the act of making a work available is a "communication to the public" even if access is only potential).

[30] *See also* Dornis (2025a).

[31] CJEU, case C-466/12, *Svensson v. Retriever Sverige AB*, EU:C:2014:76 paras. 18–28.

[32] *Ibidem* paras. 24–25.

"new public" inquiry is a legal test about the scope of authorised dissemination. In practical terms here, that legal concept overlaps with economic reality: by reaching users outside the publisher's normal channels, the chatbot usurps the publisher's monetisation opportunities. We address this economic-substitution issue further under the three-step test, but it reinforces that those chatbot users were not within the originally intended audience in a meaningful sense.

This is precisely the harm alleged in *Like Company v. Google*, where Gemini was said to deliver synthetic versions of the claimant's articles directly to users, thereby enabling consumption of the content without any interaction with the source website. Such diversion undermines both the publisher's ability to control dissemination and also the economic value of the content by eroding traffic and displacing advertising revenues. In these circumstances, the chatbot's output reaches a new public that was neither foreseen nor licensed by the rightsholder, thus satisfying the second prong of the communication-to-the-public test.[33]

The fact that the original article was freely accessible does not imply that the rightsholder consented to its dissemination via alternative, unlicensed channels. Rather, it is arguable that the rightsholder contemplated access occurring through the original website, within its monetisation framework and user environment. By delivering content through an AI interface detached from its original context, without redirects, source attribution, or user traffic to the publisher, the chatbot alters both the conditions and the mode of access. This transformation creates a form of "new public" that was not anticipated at the time of the original dissemination. This reasoning finds support in the CJEU's broader jurisprudence on the "new public" concept, particularly in cases such as *GS Media*,[34] *Filmspeler*[35] and *TVCatchup*[36] where the Court assessed whether access via new technological platforms, outside the rightsholder's control, could engage Art. 3 of the InfoSoc Directive. As in those cases, the concern here is not simply the duplication of content, but disintermediation: the AI system delivers protected material in a way that bypasses the mechanisms through which the rightsholder controls and monetises dissemination. In this light, the public reached by Gemini may be considered "new" not because it was previously excluded from access, but because

[33] It is worth noting that while both Art. 3 of the InfoSoc Directive and Art. 15(1) of the CDSM Directive may be triggered by AI-mediated uses, they serve distinct purposes and rest on different legal foundations. Article 3 protects the author's exclusive right of communication to the public, with broad scope and no express limitation on the length or nature of the protected content. By contrast, Art. 15(1) establishes a related right specifically for press publishers, limited to digital uses of defined press publications and excluding "very short extracts." *See* Recital 58 of Directive 2019/790 and Rosati (2021), p. 277. Importantly, infringement under one provision does not preclude enforcement under the other: the two rights can operate cumulatively, particularly where AI outputs both interfere with the market for press publications and communicate protected content to a broader public. Acknowledging this distinction is essential to clarify licensing frameworks and ensure appropriate remuneration for the different classes of rightsholders. *See*, *e.g.*, McDonagh (2022), pp. 309–345 (discussing the scope of Art. 15 as a distinct right that does not interfere with existing authorial rights); Rosati (2020), pp. 802–823.

[34] CJEU, case C-160/15, *GS Media v. Sanoma*, EU:C:2016:644, [2017] 1 C.M.L.R. 30.

[35] CJEU, case C-527/15, *Stichting Brein v. Wullems*, EU:C:2017:300, [2017] 3 C.M.L.R. 30.

[36] CJEU, case C-607/11, *ITV Broadcasting Ltd v. TV Catchup Ltd*, EU:C:2013:147, [2013] 3 C.M.L.R. 1.

it was not contemplated as a target audience through this specific, unlicensed mode of delivery.

3.1.3 The Sector-Specific Safeguard of Art. 15(1) CDSM Directive

Alongside the general right of communication to the public under Art. 3(2) of the InfoSoc Directive, the CDSM Directive introduces a related right specifically tailored to the interests of press publishers. Art. 15(1) was designed to address the systemic imbalance between content producers and digital platforms, particularly in the context of online news aggregation and the widespread redistribution of journalistic content without compensation.[37]

Under this provision, press publishers are granted exclusive rights over the online use of their publications, subject only to a narrowly framed exception for "very short extracts."[38] The aim is to preserve the economic value of press content in the digital environment by ensuring that its reuse remains within the control of those who produce and fund it.[39]

In the case at hand, the outputs generated by Gemini appear to exceed the threshold of what might be considered permissible under this exception. Rather than merely quoting isolated phrases, the chatbot is alleged to reproduce the substance, structure, and narrative arc of full articles, albeit in synthetic form. Such reconstructions may lack the exact wording of the original, but they preserve enough expressive elements to function as substitutes for the underlying work. In economic and communicative terms, the result is indistinguishable from an unauthorised paraphrased republication.

The fact that these outputs are generated algorithmically does not place them beyond the scope of liability. The legal test under Art. 15(1) does not depend on the internal mechanics of the reproduction, but on its effect. Where the final product is recognisable, substitutes for the original, and is made available without a licence, the protection afforded by the Directive is engaged. The synthetic or reconstructed nature of the output may complicate evidentiary questions, but it does not alter the core legal analysis.

3.1.4 Economic Substitution and the Constraints of the Three-Step Test

The Court of Justice has increasingly incorporated economic considerations into its interpretation of exclusive rights and the scope of permissible exceptions.[40] A key

[37] *See*, *e.g.* Rosati (2021), pp. 250–294; Colangelo and Torti (2019), p. 75; Scalzini (2021), pp. 101–119.

[38] *See* Rosati (2021), Article 15, p. 267

[39] *Ibidem*, at 257.

[40] *See* CJEU case C-476/17, *Pelham GmbH v. Hütter*, EU:C:2019:624, paras. 60–63 (emphasising the need to balance authors' rights with the impact of licensing on the music industry and creativity); CJEU case C-469/17, *Funke Medien*, EU:C:2019:623, para. 70 (stressing the importance of maintaining a fair balance between rightholders and users in light of technological and market realities); CJEU case C-572/13, *Reprobel*, EU:C:2015:750, paras. 44–46 (finding that a national scheme distorted the balance between rights and compensation). *See also* Rosati (2021), pp. 275–276 (discussing the role of economic incentives and justification in shaping exceptions and new rights).

element in this evolving jurisprudence is the recognition that unauthorised uses which functionally displace access to the original work, particularly in digital environments, may interfere with its normal exploitation and, by extension, with the core rationale of copyright protection.

This line of reasoning is directly relevant to chatbot outputs that reproduce press content without linking to or licensing the original source. When generative AI systems deliver content that satisfies the user's informational need in place of the underlying article, they do more than communicate information, they supplant the original's market function. This can deprive publishers of advertising revenue, subscription clicks, or licensing opportunities, thereby frustrating the economic model on which journalistic production depends. Such interference is especially significant when assessed under the three-step test enshrined in Art. 5(5) of the InfoSoc Directive.[41] That provision limits the scope of exceptions and limitations to "certain special cases which do not conflict with a normal exploitation of the work … and do not unreasonably prejudice the legitimate interests of the rightsholder." Where chatbot-generated responses divert user engagement away from the original publication, they raise precisely the type of economic harm that the test is designed to prevent.

The implication is clear: the unauthorised use of press content by generative AI tools should not be evaluated in isolation from its market effects. When those effects include displacement, substitution, or erosion of licensing value, the justification for applying any exception narrows substantially. In such cases, a system of prior authorisation is not merely preferable but necessary to maintain the balance between innovation and the protection of creative investment.

3.1.5 Clarifying the Scope: Generative Outputs and Embedded Reproduction

Although the first question referred to the Court concerns only the dissemination of chatbot responses, not the legality of training processes, it remains important to delimit the scope of the analysis. The focus here is on outputs that replicate or reconstruct existing protected content, not on those that are purely abstract or novel in form. The legal inquiry thus centres on situations where the chatbot's response, whether verbatim or reformulated, reflects the substance of a pre-existing work.

Even in systems that do not explicitly retrieve external sources at runtime, but rely solely on generative capabilities derived from prior training, the potential for unintended reproduction remains. Large language models, including those underpinning services like Gemini, may generate responses that incorporate recognisable elements of protected works included in their training datasets.[42] As mentioned before, this phenomenon is not merely hypothetical; under certain prompting conditions, such outputs may reproduce expressive content with a high degree of fidelity. From a legal standpoint, the mode of generation, be it retrieval-based, generative, or hybrid, is not determinative. What matters is whether the output includes protected material in a recognisable form. Where it does, and that material

[41] For more details on the three-step test *see infra*.

[42] *See supra* note 20 and accompanying text.

is made available to the public without authorisation, the rightsholder's exclusive rights under Arts. 3(2) of the InfoSoc Directive and 15(1) of the CDSM Directive may be engaged. The synthetic or probabilistic character of the reproduction does not place it outside the scope of these rights, particularly where the expressive core of the original work is preserved and communicated to a new audience.

3.1.6 Conclusion

Gemini's responses, insofar as they reproduce recognisable journalistic content beyond the threshold of "very short extracts" and are made accessible without authorisation or linking, fall within the scope of both Art. 15(1) of the CDSM Directive and Art. 3(2) of the InfoSoc Directive. Such outputs qualify as unauthorised communications to the public, even where they are delivered through synthetic or probabilistic processes, if they substitute the original and reach a new public. While this analysis focuses on outputs that clearly reflect pre-existing works, it does not exclude the possibility that generative responses, created without direct reference to specific articles, may, under certain conditions, also amount to a form of making available to the public, particularly where embedded training data enables downstream access to protected expression. Whether this broader theory engages Art. 3(1) of the InfoSoc Directive in its own right remains a question for further scrutiny.[43]

3.2 Question 2: Does the Training of LLMs Constitute a Reproduction Under Art. 2 of the InfoSoc Directive?

The reproduction right may be engaged at multiple stages of the generative AI development process.[44] First, reproduction occurs during the initial compilation of training corpora, typically via automated scraping of online content, where protected works are copied and stored. Second, the training phase itself may involve reproductions as the model processes and encodes protected material into machine-readable representations during the multiple training rounds which require reproductions of the elements in the training dataset. Third, certain generative models may produce outputs that reproduce, with varying degrees of fidelity, elements of the original training data. Finally, the trained model may embed internal representations that preserve the structural or expressive features of protected works, thereby raising questions about the existence of latent, non-transient copies within the model's architecture.

3.2.1 Status Quo

Whereas it seems uncontested that the collection and preparation as well as the actual utilisation of datasets consisting of copyrighted materials do require reproduction and copying of the single elements, the majority of scholarly voices

[43] For a detailed analysis *see* Dornis (2025a).

[44] *See*, *e.g.*, Dornis and Stober (2024), pp. 68 *et seq.*

holds that the internal parameters of a trained AI model do not contain "reproductions" of the training data in the sense required by Art. 2 of the InfoSoc Directive.[45] The argument relies on the functionality of generative AI systems. Because these models are designed to produce output that can serve the same function for the user as the original data, it is concluded that no direct memorisation of training data occurs. Instead, the data are internalised in a statistical or abstract form within the weights and connections of artificial neural networks. Accordingly, this statistical encoding falls outside the concept of reproduction under EU copyright law, which requires that protected elements be reproduced in a recognisable form, either by literal copying or by replicating the work's structure or expressive features.[46]

3.2.2 Analysis

The training of LLMs engages copyright law at multiple levels, as it involves not only acts of copying during data ingestion, but also the creation of internal representations that may enable subsequent reproductions. The training can also result in a reproduction of the training data – at least a significant portion of it – in the model's so-called vector space after the training process has been completed. Indeed, the fact that AI models can be prompted to generate output that closely resembles, or even replicates, their training data suggests that a corresponding "reproduction" must occur within the model itself. This is analogous to a heavily compressed image or a ciphered text: the form is altered, but it contains all the information needed to reconstruct original elements. Just as a thumbnail image or encrypted copy still counts as a reproduction of a photograph in EU law (the original can be derived or perceived from it), the neural weights contain a compressed reproduction of the training data's expression. Thus, the functional equivalence principle implies that the model's stored data, although not readable as text, is akin to any digital storage from which the work can be reconstructed. Just as a cached webpage or a thumbnail image is a reproduction because it can be transformed back into the original content or a perceptible image, the AI model's weights enable protected content to be reconstructed on command. This satisfies the reproduction criterion.

(a) Factual indicators of internal reproduction

To begin with, there is no doubt that models retain internal representations of their training data. After all, the existence of reproductions of training data can and must be inferred from the fact that these models are actually able to reproduce at least some of the works from their input, unchanged or largely unaltered, when they are prompted to generate output.[47] It is generally agreed that, depending on the

[45] *See*, *e.g.*, Käde (2021), pp. 74–75; Baumann (2023), pp. 3673, 3674; Konertz and Schönhof (2024), pp. 289, 293; for US literature, *see*, *e.g.*, Sag (2024), pp. 1885, 1910, 1912, and *passim*; for a contrary position *see*, *e.g.*, von Welser (2023), pp. 516, 517; Mezei (2024), pp. 461, 463; Dornis (2025b), pp. 65 (69 *et seq.*); for an instructive analysis *see* further Cooper and Grimmelmann (2025), pp. 14 *et seq.*

[46] *See* again Käde (2021), pp. 74–75; Baumann (2023), pp. 3673, 3674; Konertz and Schönhof (2024), pp. 289, 293.

[47] Cooper and Grimmelmann (2025), pp. 14 *et seq.*

model and technology, as well as the operating conditions, verbatim copies of input data can occur in about 0.1% to 10% of cases.[48] Also, the replication of content from the stock of training data can be evoked by sophisticated prompting, particularly with respect to LLMs and the literary works in their training dataset.[49] More recent studies have extended the understanding about the capacity of LLMs like Llama, GPT-4o, Gemini or other multimodal models to memorise and regurgitate: although the results vary among the different models, some can be enticed to regurgitate a remarkable amount of their training materials. For instance, it has been shown that Meta's model Llama 3.1 70B has memorised more than 40% of one of the books in the Harry Potter series so that it could reproduce 50-token excerpts when prompted at least any second time.[50] Research on this topic may still be in its infancy. Yet, the possibility to evoke training data from within trained models cannot be denied.[51] The fact that LLMs are able to generate output that is identical or almost identical to the training data implies that an internal representation must exist that functions as the source of such reproductions. This internal element, however, must be a reproduction as well. As A. Feder Cooper and James Grimmelmann very pointedly have explained:

> [M]emorized content must be encoded in the model's parameters. There is nowhere else it could be. A model is not a magical portal that pulls fresh information from some parallel universe into our own. A model is a data structure: it consists of information derived from its training data. The memorized training data are *in the model*. (emphasis in original)[52]

Even low-frequency regurgitation confirms the existence of stored reproductions for the purposes of Art. 2. In conclusion, what persists inside the model allowing for output that reproduces the training data is, as technology suggests, a reproduction.

(b) Technological details: "vector space" and input/output correlations

The ingestion and digestion of copyright-protected works during a model's training process does materialise in a representation in the model's so-called vector space: a highly compressed and compacted storage of elements of the training dataset.[53]

The goal of generative AI training is to develop models that produce output similar to the input data.[54] To achieve this, the model must identify patterns and

[48] *See*, *e.g.*, Lee et al. (2022); Carlini et al (2023a, b); *see also* Cooper and Grimmelmann (2025), pp. 48–49.

[49] *See*, *e.g.*, Henderson et al (2023), p. 1; Carlini et al. (2021); Chang et al (2023); Carlini et al. (2023a, b); Somepalli et al. (2022); Somepalli et al. (2023).

[50] Cooper et al. (2025); further also *supra* note 16.

[51] *See also* more recently US Copyright Office, Copyright and Artificial Intelligence – Part 3: Generative AI Training, May 2025 (pre-publication version), 19 *et seq*. with further references; further also Dornis (2025a), pp. 765 et seq.; Dornis (2024b), pp. 830 *et seq*.; Dornis (2025b), pp. 65 et seq.

[52] Cooper and Grimmelmann (2025), pp. 22–23.

[53] *See*, *e.g.*, Dornis (2025b), pp. 65 *et seq*.; further also Gervais et al. (2024).

[54] *See*, *e.g.*, OpenAI's definition at https://openai.com/index/generative-models/ ("To train a generative model we first collect a large amount of data in some domain (*e.g.*, think millions of images, sentences, or sounds, etc.) and then train a model to generate data like it."); further also Dornis (2024c), pp. 156 *et seq*.

correlations within the stock of its training data.[55] Each model "learns" the statistical properties of its training dataset, that is, the characteristics of its elements and the correlations between them.[56] Learning in this context means that the model can ultimately categorise these elements in the training data according to similarity or dissimilarity.[57] After the training process, the model consists of a complex structure of statistical information embedded in its core component: the so-called artificial neural network (ANN). The parameters of an ANN are represented numerically and no longer resemble the original copyright-protected content (e.g., text, images, music) in a perceptible way. In technical terms, these representations take the form of mathematical vectors: high-dimensional abstractions encoding learned patterns.[58]

At first glance, it may appear that this process merely leads to an "abstraction" of the training data, such that what is stored no longer qualifies as a reproduction or copy of the original works. However, this view overlooks the complexities of the model-internal statistical structure: the transformation of copyright-protected works into vectors may come along with a principal simplification and compression of the data.[59] While no 1:1 copying occurs, the statistical information retained by the model remains directly correlated with the input data. The individual components of the mathematical vectors within the ANNs still represent specific features of the original training elements.[60] As a result, the volume of data processed and transformed into statistical representations within the ANNs is immense and the internal structure of the model continues to reflect and depend on its inputs. Indeed, the vectorial representation of the training data is not only derived from the input, but is also essential to generating output. This is due to the fact that, in order to generate new content (i.e., data resembling the training input), the model must access and decode the statistical information it has retained about the works and other materials in its training dataset. In other words, whenever a trained model is prompted to generate output, it draws from its internal statistical storage and reconstructs it into a human-perceivable form, whether textual, visual, or auditory.[61]

This strong input-output correlation has been widely explored in AI research.[62] AI scholars have described the variables in the model vector space of an ANN as a compacted and abstracted representation of the training dataset. While these stored representations are not identical copies or a 1:1 replication of the elements of the

[55] *See*, *e.g.*, Cooper and Grimmelmann (2025), p. 9; *see also* Dornis and Stober (2024), p. 40.

[56] For statistical processing, *see*, *e.g.*, Radford et al. (2021); *see also* Cooper and Grimmelmann (2025), p. 9.

[57] Radford et al. (2021); *see also* Sobel (2024), p. 16.

[58] In LLMs (large language models) based on transformer architecture, such as GPT models, the vectors are referred to as embeddings. In models with probabilistic internal representations such as GANs (generative adversarial networks), VAEs (variational autoencoders) and latent diffusion (*e.g.*, stable diffusion), the vectors describe latent random variables in the so-called latent space of the model. *See*, *e.g.*, Dornis and Stober (2024), pp. 29 *et seq*., 43 *et seq*. and 117 *et seq*.

[59] Sobel (2024), p. 17; *see also* Cooper and Grimmelmann (2025), pp. 35 *et seq*.

[60] Sobel (2024), p. 16.

[61] Ramesh et al (2022); Sobel (2024), p. 18.

[62] Carter and Nielsen (2017).

training dataset, they constitute compressed reproductions, and it is widely acknowledged that trained models retain a statistically encoded version of the original content.[63]

(c) Doctrinal verification

The fact that a specific kind of replication of the training data exists within the statistical information of the vector space directs the legal analysis to the concept of "reproduction" under EU copyright doctrine. The statutory law does not expressly define "reproduction".[64] Yet the concept of "reproduction" has been extensively explored, and it is clear that it must be interpreted broadly, primarily due to the fact that the InfoSoc Directive's main objective is to introduce a high level of protection, particularly to enable authors to receive an appropriate reward for the use of their works.[65] As the Court of Justice has repeatedly emphasised, a broad interpretation must ensue from the Directive's statutory text defining reproductions by utilising expressions such as "direct or indirect", "temporary or permanent", "by any means", and "in any form".[66] The CJEU's interpretation of transient digital reproductions in *Infopaq* and *Premier League v. QC Leisure*[67] illustrates this aspect insofar as the Court held that temporary technological storage, even when not directly perceptible to humans, can amount to reproduction if it enables access to protected content in a perceptible form. These cases affirm that fixation need not involve a human-readable format or immediate perceptibility at the storage stage; rather, it suffices that the data can be retrieved and rendered intelligible at a later point in the process. Moreover, the CJEU's jurisprudence in *Infopaq* and *Meltwater* supports a functionality-based reading of reproduction. What matters is the capacity of the stored material to enable future access in perceptible form. Such transient copies are legally relevant not because they are durable, but because they facilitate user-perceptible reproductions. It is therefore also undisputed that a "reproduction" can inter alia lie in a digital replication or digital storage of a copyright-protected work (e.g., stored on a CD, DVD, or hard disc), regardless of the kind of technology that has been used to generate the digital copy.[68] Also, it is important to remember that, under the InfoSoc Directive, a "reproduction" can also exist if the original

[63] Carter and Nielsen (2017); *see* also Cooper and Grimmelmann (2025), pp. 35 *et seq.*

[64] *See*, *e.g.*, CJEU, case C-5/08, *Infopaq Int'l A/S v. Danske Dagblades Forening*, ECLI:EU:C:2009:465, 2009 E.C.R. I-6569, para. 31; *see* also CJEU, case C-476/17, *Pelham v. Hütter,* ECLI:EU:C:2019:624, para. 52.

[65] *See* Art. 2 and recital 21 InfoSoc Directive; *see* also CJEU, case C-5/08, *Infopaq Int'l A/S v. Danske Dagblades Forening*, ECLI:EU:C:2009:465, 2009 E.C.R. I-6569, para. 40.

[66] CJEU, case C-5/08, *Infopaq Int'l A/S v. Danske Dagblades Forening*, ECLI:EU:C:2009:465, 2009 E.C.R. I-6569, para. 42; CJEU, case C-433/20, *Austro-Mechana Gesellschaft zur Wahrnehmung mechanisch-musikalischer Urheberrechte Gesellschaft mbH v. Strato AG* [2022], ECLI:EU:C:2022:217, para. 16; CJEU, case C-426/21, *Ocilion IPTV Technologies GmbH v. Seven.One Entertainment Group GmbH, Puls 4 TV GmbH & Co. KG* [2023], ECLI:EU:C:2023:564, para. 28.

[67] CJEU, case C-403/08 and C-429/08, *Football Ass'n Premier League Ltd. v. QC Leisure & Murphy v. Media Prot. Serv. Ltd.*, ECLI:EU:2011:631, paras. 161–171.

[68] *See*, *e.g.*, CJEU, case C-403/08 and C-429/08, *Football Ass'n Premier League Ltd. v. QC Leisure & Murphy v. Media Prot. Serv. Ltd.*, ECLI:EU:2011:631, para. 159; for German doctrine, *see*, *e.g.*, Federal Court of Justice (BGH) GRUR 2010, 616 (619) – *marions-kochbuch.de*; further also Sesing-Wagenpfeil (2024), pp. 212, 236.

work has been simplified or reduced, e.g., from a full-fledged photograph to a digital thumbnail.[69]

Consequently, the concept of "reproduction" must be explained as being distinctly technology-neutral. It is irrelevant how the reproduction is embodied, or whether technical means or instruments are required to make the embodiment perceptible for humans. The only prerequisite for a finding of reproduction is that it must be fixed on a physical carrier from which it can be retrieved and made perceptible.[70] By analogy, the statistical encodings of protected works in the vector space of generative AI models, though not directly interpretable by humans, function as intermediaries that enable recognisable outputs. The causal chain between statistical representation and expressive reproduction supports the view that such internal states amount to "reproductions" under Art. 2 of the InfoSoc Directive. The doctrinal relevance lies in the potential of the stored data to generate outputs that make the protected work accessible again, even in altered or paraphrased form.

Finally, in addition to being technology-neutral, the concept of "reproduction" in EU copyright law, as in other copyright regimes, must be functional and, hence, open to new technological developments. The fact that the works used in AI training are statistically transformed and stored in the models' vector space therefore does not imply a lack of "perceptibility". In the past, it has often been challenging for lawyers to explain why technically feasible replications may qualify as legally relevant "reproductions". Yet, the mere fact that a technology may not yet be fully explainable has never been and should not be an obstacle. In fact, looking at what can so far be explained with regard to the statistical replications in ANNs, there is no relevant difference compared other variants of digital storage, such as on computer hard discs.[71] In both scenarios, it is not required that the process of creating, organising, and ultimately unbundling the digital storage follow a specific technique or system, or that it preserves the nature or format of the original work. It is incontestable and increasingly explored by AI scholarship that, at least when utilising a suitable *prompt*, LLMs and other models may eject replications of the training data. The ultimate characteristic of a reproduction – understood in such a functional sense – is that one can replicate the work from the reproduction.[72] Accordingly, the actual form of storage, e.g., tokenisation,[73] vectorisation[74] or other technique, is irrelevant as long as the model can output elements that are recognisable as being significant parts of original works.

[69] For German copyright doctrine, *see*, *e.g.*, Federal Court of Justice (BGH) GRUR 2010, 628 (629) – *Vorschaubilder I*; *see also* Schulze (2022). Courts in the United States have decided in the same vein. *See*, *e.g.*, *Perfect 10, Inc. v. Amazon.com*, 508 F.3d 1146 (9th Cir. 2007).

[70] For scholarly commentary on this aspect *see*, *e.g.*, Heerma (2022); Sesing-Wagenpfeil (2024), pp. 212, 229.

[71] Cooper and Grimmelmann (2025), pp. 30 *et seq*.

[72] *See also*, for US doctrine, Cooper and Grimmelmann (2025), p. 20.

[73] "Tokenisation" refers to the process of breaking text into discrete units (tokens) such as words or subwords for processing by language models.

[74] "Vectorisation" involves converting tokens into numerical vectors that allow machine learning models to capture semantic relationships and generate context-aware predictions.

3.2.3 Conclusion

In light of the foregoing analysis, the training of LLMs such as Gemini entails systematic acts of reproduction within the meaning of Art. 2 of the InfoSoc Directive. These acts occur not only during the initial copying and storage of protected works for training purposes, but also through the embedding of recognisable expressive elements into the model's internal architecture, whether in the form of tokenised sequences, statistical weights, or vectorised representations. As demonstrated, these representations are not technologically neutral artefacts, but functional encodings of protected expression, capable of producing outputs that replicate original content with varying degrees of fidelity. EU copyright law, as interpreted by the CJEU, does not require literal or perceptible copies to establish reproduction. Instead, it requires that protected elements of a work be retained in a recognisable form, whether directly or indirectly, permanently or temporarily, and by any technological means. The evidence that LLMs can regenerate substantial excerpts of protected material upon prompt confirms that such recognition is not hypothetical, but observable in practice. Accordingly, unless the use of protected material falls within the scope of a valid exception or limitation – applied in conformity with the three-step test – the ingestion and internalisation of such content during AI training amounts to an unauthorised act of reproduction. Prior authorisation from rightsholders is therefore required.

3.3 Question 3: Can Art. 4 of the CDSM Directive Justify Training Under the TDM Exception?

The TDM exception under Art. 4 of the CDSM Directive was designed to enable automated analysis aimed at extracting patterns, trends, or correlations from large datasets. Its scope, as framed in both the statutory language and the recitals, presupposes an analytical, not generative, function.[75] By contrast, the training of generative AI models involves the large-scale ingestion and internalisation of protected expression for the purpose of producing synthetic outputs, a process that does not align with the informational objectives of TDM.[76] Moreover, what has so far not been extensively explored, the application of Art. 4 to such uses raises serious concerns under the three-step test, which remains the cornerstone for interpreting exceptions and limitations in EU and international copyright law.

3.3.1 "Text and Data Mining": Statutory Text and Technological Functionality

Article 2(2) of the CDSM Directive defines 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."

[75] The very phrasing of Art. 4 ("for the purpose of text and data mining") and its recitals indicate it was not meant to cover uses whose primary aim is the reproduction of expressive content rather than the extraction of informational insights.

[76] *See* extensively already Dornis (2025b), pp. 65 *et seq*.; further also Dornis (2024c), pp. 156 *et seq*.

Further, Recital 8 of the Directive explains: "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 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." Evidently, the statutory text (including further recitals[77]), giving an insight into the purpose of TDM, is *exclusively* about "information", "new knowledge" and "discovering new trends".

Indeed, nowhere in the Directive and its recitals is there any reference to the purpose of generative AI, namely the production of data that are similar to the training data.[78] This is no surprise since the latter is so fundamentally different from processes that generate information and new knowledge. To begin with, the training of generative AI systems cannot lead to the disclosure of any findings – i.e., new "information" or "knowledge" – to the outside world. After all, the training is limited to adapting the parameters of the model to the training data (as explained already supra: the ANN's vector space). Yet since it is simply impossible to look into the model's black box, no new information will ensue from the training process as such.[79] The training phase does not, in and of itself, reveal any "result" in the form of new knowledge or patterns to a human observer: it merely produces a tuned model. Unlike a typical data mining exercise, where one might publish statistics or correlations found, here the output of training is an opaque set of parameters."

3.3.2 Squaring Round Pegs into Square Holes: Generative AI Training and TDM

In addition to the textual mismatch, a number of technological features expose the incompatibility of generative AI training with the assumptions underpinning the TDM exception.[80]

(a) Functional purpose and expressive output

[77] Even Recital 18 – referring to private, and hence commercial, TDM methods – refers exclusively to the analysis of data, decision-making, and development of new applications or technologies. Nowhere is there any hint that lawmakers wanted to leave the path of TDM-is-nothing-more-than-mere-information-processing-for-generating-new-knowledge: "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."

[78] For a definition of the actual purpose of generative AI – i.e., the production of similar data – *see*, *e.g.*, OpenAI's definition: "To train a generative model we first collect a large amount of data in some domain (*e.g.*, think millions of images, sentences, or sounds, etc.) and then train a model to generate data like it." (https://openai.com/index/generative-models/); further also Dornis and Stober (2024), pp. 39 *et seq*.

[79] *See also* Nordemann and Pukas (2022), pp. 973, 974; but *see* Lux and Noll (2024), pp. 111, 113.

[80] It is worth noting that, while some scholars initially suggested that the scope of "analytical" uses under Art. 4 CDSM may extend to certain machine learning applications, depending on their purpose and effect (*see*, *e.g.*, Rosati (2021), Article 4, pp. 86–127): the majority of legal commentary now highlights multiple unresolved issues and doctrinal limitations in applying the EU's TDM exceptions to generative AI training activities. *See*, *e.g.*, Ducato and Strowel (2021), pp. 322–337; Margoni and Kretschmer (2022), pp. 685–701; Rosati (2024a, b), p. 851; Fernández-Molina and de la Rosa (2024), pp. 653–672; Tyagi (2024), pp. 557, 562–63; Buick (2025), pp. 182, 190; Dornis (2025b), p. 121; Brauneis (2025), p. 1; Lucchi (2025).

TDM outputs, such as topic clusters, co-occurrence matrices, or statistical trends, are non-expressive and intended to support scientific discovery or data-driven insight. Generative AI systems, conversely, are designed to internalise linguistic expression in order to produce fluent outputs in natural language or images. They do not summarise works; they learn to simulate them.[81] The objective is not to analyse content but to reconstruct expressive structures for downstream tasks such as summarisation, translation, or dialogue completion.[82] This divergence becomes especially problematic given that generative outputs often include memorisations or stylistically close paraphrases of the input material.[83] Yet concerning the functionality and purpose of generative AI, this more generally reveals that generative AI is designed to produce output that can replace the original data it has been trained with. Necessarily, its output must be *expressive*. Hence, generative AI training goes beyond non-expressive analysis by necessity.

(b) Syntax/semantics indivisibility

As seen, TDM rests on the assumption that there is a clear doctrinal separation between unprotected ideas or facts and protected expressions. However, this distinction collapses in the context of generative AI training.[84] During training, models ingest works as tokenised sequences and transform them into high-dimensional embeddings that statistically encode both form and meaning. Since the models' algorithms cannot distinguish information according to its actual semantic or syntactic contents (i.e., whether it is mere facts or expressive elements) any training will inevitably always ingest and digest the informational contents of the training data in toto. Since all generative AI models are by design blind and agnostic vis-à-vis the distinction, these model-internal representations that evolve during the training process do not preserve the legal distinction between expression and ideas. No matter whether it is copyright-protected stylistic or structural elements or non-protected factual or semantic information, all data will be taken as food for the artificial learning and evolution process. This axiom of generative AI technology, however, fundamentally undermines the widely propagated conception that TDM is permissible because it involves the extraction only of non-protectable elements. While copyright does not protect mere facts or ideas, in generative training no such separation is possible; expressive elements are ingested wholesale along with factual ones. This undermines the assumption (sometimes optimistically made by proponents of broad TDM) that only unprotected data are extracted. In reality, the model stores stylistic and structural patterns, elements at the core of expression, rather than isolating just facts.

[81] *See*, *e.g.*, Mitchell and Krakauer (2023), p. 13 (explaining that large language models simulate understanding by manipulating surface-level linguistic patterns, rather than genuinely summarising or analyzing underlying semantic content).

[82] *See*, *e.g.*, Bender et al. (2021) (explaining how LMs are not performing natural language understanding but rather manipulate linguistic form).

[83] A key study by Carlini et al. demonstrated that large language models (LLMs) trained on copyrighted content are prone to regurgitate exact passages from their training corpus. *See* Carlini et al. (2021), pp. 2633–2650.

[84] *See* extensively Dornis and Stober (2024), pp. 101 *et seq.*; Dornis (2025b), pp. 65 *et seq.*

(c) "Copying Expression for Expression's Sake"[85]

As seen, generative AI models are purposefully trained to reproduce *expression:* they learn style, tone, structure, rhythm, and phrasing.[86] This is not an incidental consequence of processing data; it is the design goal. Expression is thus not copied for analytical purposes only, but to enable the later generation of similarly styled or derivative content. This training process does not align with the logic underlying the TDM exceptions in EU law, which were intended to facilitate non-expressive, analytical uses of protected works. Rather, the purpose and effect of generative training align more closely with the reproduction right under Art. 2 of the InfoSoc Directive, which protects against the copying of expression. Importantly, as the CJEU made clear in *Infopaq*, even the reproduction of very short excerpts, such as 11-word fragments, may infringe copyright if they reflect the author's intellectual creation.[87] Given that generative models routinely copy such expressive fragments at scale during training, their operation falls outside the scope of lawful TDM and into the domain of exclusive rights, thereby requiring prior authorisation from rightsholders.[88]

3.3.3 Finally: The Three-Step Test and the Discontents of Generative AI Training

While the TDM exception in Art. 4 of the CDSM Directive sets out a formal legal basis for certain uses of protected works, its applicability remains subject to the overarching constraint of the three-step test enshrined in Art. 5(5) of the InfoSoc Directive. This test functions as a doctrinal safeguard, ensuring that exceptions and limitations are interpreted restrictively and remain consistent with the core principles of international copyright law. In the context of generative AI training, a rigorous application of the three-step test reveals fundamental incompatibilities with large-scale, unlicensed uses of protected content.

(a) Overview

The three-step test has its origins in international copyright law[89] and has been expressly incorporated into EU copyright law within Art. 5 (5) InfoSoc Directive.[90] Implementing the test levels – i.e., the three essential criteria – from international copyright law, the provision reads:

> The exceptions and limitations provided for in paragraphs 1, 2, 3 and 4 shall only be applied in certain special cases which do not conflict with a normal

[85] Credit for the first use of the term "copy[ing] expression for expression's sake" goes to Lemley and Casey (2021), pp. 743, 777.

[86] *See* Lucchi (2025).

[87] CJEU, case C-5/08, *Infopaq Int'l A/S v. Danske Dagblades Forening*, ECLI:EU:C:2009:465, 2009 E.C.R. I-6569, para. 47.

[88] For more examples of AI research that highlight the fact that AI models do feed on the expressive contents of their training data in particular *see* Dornis and Stober (2024), pp. 112 *et seq.*; Dornis (2025b), pp. 65 *et seq.*

[89] Dornis and Stober (2024), pp. 143 et seq. The origin is attributed here to Art. 13 TRIPS Agreement, Art. 10 WCT, Art. 16 WPPT and Art. 9 para. 2 Revised Berne Convention.

[90] The standard literally refers to Art. 5(1) to (4) InfoSoc Directive only, but also applies to the TDM exception in Art. 4 CDSM Directive (via Art. 7(2) CDSM Directive).

exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightsholder.

Although the European Union is not a party to the Berne Convention, the Court of Justice considers itself bound by Arts. 1 to 21 of the Convention when interpreting it in accordance with Art. 1(4) of the WIPO Copyright Treaty, to which the Union is a party.[91] Accordingly, interpretation of the three-step test under international copyright law is determinative for the CJEU's doctrine on copyright exceptions.

Under the three test levels, limitations and exceptions must comply with the following prerequisites:

1. a restriction must only apply "in certain special cases", which
2. does "not conflict with a normal exploitation of the work or other subject-matter", and
3. does "not unreasonably prejudice the legitimate interests of the rightsholder."

The constituent elements of all three levels must be examined cumulatively. Therefore, a limitation or exemption will only be test-compliant if the determination for all three test elements, analyzed separately, implies the permissibility.[92] Doctrinally, the three-step test has the function of a "limitation on limitations", which limits the national legislators' discretion.[93] Thus, exceptions must be interpreted narrowly, as they constitute derogations from exclusive rights of the rightsholder.[94] Currently, the debate on the three-step test and its relevance for the training of generative AI is still, and somewhat surprisingly, underdeveloped.[95] As far as can be seen, a common argument brought forward to justify application of the TDM exception in light of the three-step test is that rightsholders could still achieve

[91] CJEU, case C-403/08 and C-429/08, *Football Ass'n Premier League Ltd. v. QC Leisure & Murphy v. Media Prot. Serv. Ltd.*, ECLI:EU:2011:631, para 162; CJEU, case C-277/10, *Martin Luksan v. Petrus van der Let*, ECLI:EU:C:2012:65, para. 59; CJEU, case C-510/10, *TV2 Danmark A/S v. Nordisk Copyright Bureau*, ECLI:EU:C:2012:244, para. 29. The EU is also a WTO member and has ratified the TRIPS Agreement.

[92] *See*, *e.g.*, CJEU, case C-435/12, *ACI Adam BV v. Stichting de Thuiskopie*, ECLI:EU:2014:254, para. 31; further also WTO, Report of the Panel, United States – Section 110(5) of the US Copyright Act, WT/DS160/R (15 June 2000), para. 6.74, para. 6.97; Kur (2009), pp. 287, 314; Ricketson and Ginsburg (2022), para. 13.09.

[93] CJEU, case C-5/08, *Infopaq Int'l A/S v. Danske Dagblades Forening*, ECLI:EU:C:2009:465, 2009 E.C.R. I-6569 para. 58; CJEU, case C-435/12, *ACI Adam BV v. Stichting de Thuiskopie*, ECLI:EU:2014:254, para. 25; CJEU, case C-527/15, *Stichting Brein v. Wullems*, ECLI:EU:C:2017:300, para. 63; von Welser (2023), pp. 516, 518.

[94] CJEU, case C-5/08, *Infopaq Int'l A/S v. Danske Dagblades Forening*, ECLI:EU:C:2009:465, 2009 E.C.R. I-6569, paras. 56–57; CJEU, case C-403/08 and C-429/08, *Football Ass'n Premier League Ltd. v. QC Leisure & Murphy v. Media Prot. Serv. Ltd.*, ECLI:EU:2011:631, para. 162; CJEU, case C-360/13, *Pub. Relations Consultants Ass'n Ltd. v. Newspaper Licensing Agency Ltd.*, ECLI:EU:C:2014:1195, para. 23; CJEU, case C-527/15, *Stichting Brein v. Wullems*, ECLI:EU:C:2017:300, para. 62.

[95] Only occasionally is reference made to the three-step test as the limit of admissibility for AI training. *See*, *e.g.*, von Welser (2023), pp. 516, 519; Schack (2024), pp. 113, 117; Lux and Noll (2024), pp. 111, 114; Rosati (2024a, b), pp. 1, 20–21; Calderón (2024), pp. 84, 96. But *see* Senftleben (2023) (providing a full three-step analysis).

appropriate remuneration by means of the TDM exception's opt-out mechanism.[96] As can be shown, however, it is in particular the second step of the test, i.e., the fact that generative AI training "conflicts with a normal exploitation" by the rightsholders, that sheds considerable doubts. Necessarily, since a failure on the second test level can neither be remedied by granting a reservation of rights nor by appropriate compensation for the rightsholders, this implies that any variant of a TDM exception of limitation likely is incompliant with the three-step test.

(b) Step 1: Certain special cases

Exceptions may only cover certain special cases, so their scope of application must be limited.[97] It is not required that all specific use cases are explicitly named. It rather suffices that the scope of application is known and specified.[98] In this regard, if the TDM exception was interpreted to also cover generative AI training, it would very likely pass the first test level for it provides for a well-defined description of the factual prerequisites of copyright limitations.

(c) Step 2: Conflict with a normal exploitation of the work or other subject-matter

Yet, things are different with regard to the second test level, i.e., the inquiry whether TDM "conflict[s] with a normal exploitation of the work". This question is central to the test outcome in all its variants, especially for the Berne Convention and the TRIPS Agreement.[99] The doctrinal structure as well as the underlying policy of the second test level can best be explained on the basis of the (so far) only WTO panel decision on the issue. The WTO panel, in 2000, decided on Sec. 110(5)(B) of the US Copyright Act in light of the test implemented in Art. 13 TRIPS Agreement.[100] Although the decision is not formally binding for international and national courts, it is essential for the construction of the test and its prerequisites.[101] The panel's analysis and explanation particularly help to understand what is meant by a "normal exploitation of the work or other subject-matter".

3.3.3.1 Empirical-Quantitative and Normative Perspective

To assess the scope of a "normal exploitation", both empirical-quantitative and normative aspects must be taken into account.[102] An empirical-quantitative analysis

[96] *See*, *e.g.*, Schack (2021), pp. 904, 907; Steinrötter and Schauer (2021), § 4 para. 13; Hofmann (2024), pp. 11, 15; also in detail Senftleben (2023) pp. 1535, 1544–1545.

[97] CJEU, case C-117/13, *Technische Universität Darmstadt v. Eugen Ulmer KG*, ECLI:EU:C:2014:2196, paras. 47–48; with further references Senftleben (2004), pp. 133 *et seq.*

[98] Ricketson and Ginsburg (2022), para. 13.10.

[99] On the relationship between the international conventions *see*, *e.g.*, Ricketson and Ginsburg (2022), paras. 13.14 *et seq.*, para. 13.93 *et seq.* and para. 13.103.

[100] The WTO panel decision refers directly to Art. 13 of the TRIPS Agreement. However, due to the historical basis of the provision in Art. 9(2) Berne Convention as well as the wide accordance in the text of both provisions, interpretation and application widely correspond. *See* WTO, Report of the Panel, United States – Section 110(5) of the US Copyright Act, WT/DS160/R (15 June 2000), para. 6.72 (with n. 95) and para. 6.97 (with n. 105).

[101] *See*, *e.g.*, Wymeersch (2023), pp. 631, 633, 640–641; further also Senftleben (2004), pp. 109–110.

[102] Ricketson and Ginsburg (2022), para. 13.15 *et seq.*; further also WTO, Report of the Panel, United States – Section 110(5) of the US Copyright Act, WT/DS160/R (15 June 2000), para. 6.178.

will ask whether an exception affects certain modes of exploitations that the rightsholder usually or typically claims for herself and with which she could generate income.[103] This includes both markets that are already explored as well as sources of income that will have to be developed in the future.[104] For the normative aspect, it must be considered which forms of revenue and sources of income are conceivable, taking into account the expected technological development.[105]

3.3.3.2 Historical Roots: Berne's Travaux Préparatoires

This interpretation in favor of rightsholders was proposed in 1967, during the negotiations of the Stockholm Revision Conference for the Berne Convention. There, the Bureaux Internationaux Réunis pour la Protection de la Propriété Intellectuelle (BIRPI) Study Group demanded *inter alia*, to leave the Member States free:

> [to] limit the recognition and the exercising of that right, for specified purposes and on the condition that these purposes should not enter into economic competition with these works" in the sense that "all forms of exploiting a work, which have, or are likely to acquire, considerable economic or practical importance, must be reserved to the authors.[106]

According to the panel, such a prospective interpretation requires, in addition to the consideration of modes of exploitation that already generate revenue at present, consideration of income opportunities that may, with a certain degree of probability and plausibility, be of considerable economic or practical significance in the future.[107] A detrimental disadvantage must be feared if the mode of exploitation in question might enter into economic competition with exploitations that are exclusively assigned to the rightsholder. In this regard, the WTO panel went on to say that:

> We believe that an exception or limitation to an exclusive right in domestic legislation rises to the level of a conflict with a normal exploitation of the work (i.e., the copyright or rather the whole bundle of exclusive rights conferred by the ownership of the copyright), if uses, that in principle are covered by that right but exempted under the exception or limitation, enter into economic competition with the ways that right holders normally extract

[103] Ricketson and Ginsburg (2022), para. 13.16.

[104] WTO, Report of the Panel, United States – Section 110(5) of the US Copyright Act, WT/DS160/R (15 June 2000), para. 6.180.

[105] Ricketson and Ginsburg (2022), para. 13.17.

[106] Quoted from WTO, Report of the Panel, United States – Section 110(5) of the US Copyright Act United States – Section 110(5) of the US Copyright Act, WT/DS160/R (15 June 2000), para. 6.179.

[107] WTO, Report of the Panel, "United States – Section 110(5) of the US Copyright Act", WT/DS160/R (15 June 2000), para. 6.180 and para. 6.181 (with reference to the BIRPI Study Group).

> economic value from that right to the work (i.e., the copyright) and thereby deprive them of significant or tangible commercial gains.[108]

As a consequence, a prognostic view of future technological developments and the associated market effects is required.[109] Both existing and potential future marketplaces are principally reserved to the rightsholder and must therefore be taken into account when assessing an exception's impact, including all conceivable digital exploitations, in particular on the Internet.[110] Of course, in order to not lead the test *ad absurdum*, agreement exists that the mere fact that rightsholders are currently prevented from exploiting their works, either by their national copyright regime or by practical obstacles, does not mean that they are not entitled to exploit their works.[111]

3.3.3.3 EU Copyright Law: CJEU Doctrine

A corresponding approach can be found in CJEU case law. Here as well, the three-step test implies conflict with the rightsholder's interests in a normal exploitation if a limitation or exception might result in a reduction of her transactions.[112] In *Stichting Brein v. Wullems*, the CJEU held:

> It must also be held that, as a rule, temporary acts of reproduction, on a multimedia player such as that at issue in the main proceedings, of copyright-protected works obtained from streaming websites belonging to third parties offering those works without the consent of the copyright holders are such as to adversely affect the normal exploitation of those works and causes unreasonable prejudice to the legitimate interests of the right holder, because ... that practice would usually result in a diminution of lawful transactions relating to the protected works, which would cause unreasonable prejudice to copyright holders ...[113]

An important aspect here is that an exception allowing for exploitation of works that are accessed *unlawfully* (e.g., on pirate websites) obviously impairs the

[108] WTO, Report of the Panel, "United States – Section 110(5) of the US Copyright Act", WT/DS160/R (15 June 2000), para. 6.183.

[109] Ricketson and Ginsburg (2022), para. 13.16; further also WTO, Report of the Panel, United States – Section 110(5) of the US Copyright ActUnited States – Section 110(5) of the US Copyright Act, WT/DS160/R (15 June 2000), para. 6.187.

[110] *See*, *e.g.*, Senftleben (2014), pp. 1, 8–9 ("If understood broadly, the criterion of potential markets of 'considerable economic or practical importance' may cover all forms of using copyrighted works on the Internet.").

[111] WTO, Report of the Panel, United States – Section 110(5) of the US Copyright Act, WT/DS160/R (15 June 2000), para. 6.188; further also Ricketson and Ginsburg (2022), para. 13.16; Oliver (2002), pp. 119, 165; Kur (2009), pp. 287, 317 *et seq*.

[112] CJEU, case C-435/12, *ACI Adam BV v. Stichting de Thuiskopie*, ECLI:EU:2014:254, para. 39; CJEU, case C-527/15, *Stichting Brein v. Wullems*, ECLI:EU:C:2017:300, para. 70; *see* also von Welser (2023), pp. 516, 518–519.

[113] CJEU, case C-527/15, *Stichting Brein v. Wullems*, ECLI:EU:C:2017:300, para. 70 (with reference to CJEU, judgment of 10 April 2014, *ACI Adam and Others*, C-435/12, EU:C:2014:254, para. 39).

rightsholders' normal exploitation so severely that it would have to fail the three-step test per se.[114] At the same time, regardless of the legality of access, the Court expressed that it is the effortless and costless availability of copyright-protected works *as such* that diminishes the rightsholders' lawful transactions and detrimentally affect their normal exploitation.

3.3.3.4 Application: TDM and Generative AI Training

Against this background, if we take a look at the question whether there will be a "conflict with a normal exploitation" if copyright-protected works or other material are used for the training of generative AI models, the interests of rightsholders can be affected both under a quantitative-empirical and under a normative perspective:

Looking at actual options for rightsholders to exploit, the picture seems undisputed: The current practice of publishing houses and AI companies illustrates that a marketplace for licensing AI training data *does* exist.[115] Hence, rightsholders do have an option to exploit their works by licensing them for the training of AI models. The current status quo of the AI companies' help-yourself attitude, however, deprives authors of these exploitation possibilities and significantly reduces the number of transactions (i.e., conflicts with normal exploitation).[116]

In addition, under a normative perspective, there is a deeper concern that has more recently entered the stage. This concern has been explained as "market dilution", in particular by Judge Chhabria of the US District Court of Northern California. As he explained in *Kadrey v. Meta* the use of copyrighted books and works by Meta to train their LLMs might "harm the market for those works … by helping to enable the rapid generation of countless works that compete with the originals, even if those works aren't themselves infringing."[117] Of course, this analysis has been developed in light of the fourth fair-use factor under US copyright law.[118] But the underlying economic theory of marketplace analysis is the same: The massive training of generative AI with man-made works will ultimately lead to a world in which humans and their works are at risk of being substituted. The output of generative AI is in direct competition with the originals whose content and forms of expression they reproduce.[119] Yet, since artificially generated products can be offered on the market at significantly lower prices, human creators are being pushed out of the market for their own products. This development, looking into the future, has just begun. Also, as Judge Chhabria has correctly noted, the issue is not whether generative AI does produce *identical* products. There is no need for direct infringement. Rather, a significant impairment may already arise where such

[114] Rosati (2024a, b), pp. 1, 20–21.

[115] *See*, *e.g.*, Singh (2025).

[116] *See*, *e.g.*, Dornis and Stober (2024), pp. 84, 96.

[117] *Richard Kadrey et al. v. Meta Platforms, Inc*., Case 3:23-cv-03417-VC (June 25, 2025), *28 (Vince Chhabria, J.)

[118] *See* Section 107 US Copyright Act.

[119] Lemley and Casey (2021), pp. 743, 777: "[…] ML systems both copy expression for expression's sake and pose a threat of 'significant substitutive competition' to the work originally copied".

systems are capable of generating comparable or functionally equivalent output that replicates materials stylistically similar to the works in the training data.[120] Overall, therefore, the impending competition between human actors and machines will continue to steadily diminish opportunities for flesh-and-blood creators to market their originals.[121]

Finally, it is important to note that, even if the legislator had provided for direct financial compensation to compensate for the disadvantages of an exception or limitation, or if the lawmakers would try to do so, this could not remedy the "conflict" with a "normal exploitation". Rather, general agreement seems to exist that, if there is a conflict with the normal exploitation, the three-step test must *inevitably* fail, even if fair compensation is granted to the rightsholders.[122]

3.3.3.5 The Non-Panacea: A Defunct Opt-Out Mechanism

Notwithstanding the rather evident conflict with a normal exploitation that generative AI training creates, some voices in scholarly literature argue that the possibility of an opt-out could ensure compatibility with the three-step test. In other words, they argue that, since rightsholders under the current TDM regime can theoretically prevent application of the TDM exception by simply reserving their rights via opting out, the threat of an erosion of the rightsholders' marketplaces should not be considered relevant.[123] This perspective is shortsighted.

To begin with, one must ask whether the need for rightsholders to proactively declare an opt-out is compatible with the prohibition of formalities under Art. 5(2) Berne Convention, since the Convention's prohibition on formalities not only precludes an imposition of formalities for the *acquisition* of rights but also prohibits formalities to be introduced "through the back door".[124] Even if characterised as a condition on an exception, the need to opt out shifts the burden in a way that conflicts with the spirit of automatic protection under Berne. Accordingly, an opt-out regime that establishes formal requirements for rightsholders to comply with in order to avoid a loss of rights, for example a machine-readable message to web-scraping bots, is problematic *per se*. Even though different institutions have explored the potential of non-binding metadata standards and technical guidelines to support the expression of rights reservations these approaches remain fragmented,

[120] US Copyright Office, Copyright and Artificial Intelligence – Part 3: Generative AI Training, May 2025 (pre-publication version), pp. 65–66.

[121] Škiljić (2021), pp. 1338, 1354; Vesala (2023), pp. 351, 366 (2023); further also generally Gervais (2019), pp. 22, 32.

[122] WIPO, Guide to the Berne Convention for the Protection of Literary and Artistic Works (Paris Act, 1971), 1978, Art. 9 para. 9.7 ("If the contemplated reproduction would be such as to conflict with a normal exploitation of the work it is not permitted at all. [sic!])"; Senftleben (2004), p. 131; Lucas (2010), pp. 277, 279; Wymeersch (2023), pp. 631, 635.

[123] Senftleben (2023), pp. 1535, 1544; also (identically) Senftleben (2014), pp. 1, 12 *et seq.*

[124] Ginsburg (2016), pp. 745, 763; Sobel (2021), pp. 221, 240; Peukert (2005), pp. 1, 60 *et seq.*

experimental, and non-mandatory.[125] Their adoption by rightsholders has been inconsistent, and many creators, particularly smaller entities and individual authors, lack the technical capacity or resources to implement them meaningfully.[126] In the absence of a harmonised EU-wide framework or a centralised registry for opt-out declarations, these tools cannot operate as effective or legally secure safeguards. More broadly, relying on such mechanisms imposes a de facto burden-shifting regime, whereby authors are presumed to consent unless they take proactive and often technically complex steps to opt out, an approach that sits uneasily with the principle of automatic protection enshrined in the Berne Convention and the broader EU copyright acquis.

(d) Step 3: Unreasonable prejudice to the legitimate interests

Although application of the TDM exception to generative AI training will already fail the second test level, completeness requires to have a concluding look at the final step. In this regard, however, the exception will also not fare much better.

The third level requires a balancing of all interests involved.[127] In contrast to the second step, the balancing on the third test level may also take into account whether rightsholders receive compensation for the conflict with their interests in exploitation.[128] In its decision on Sec. 110(5)(B) of the US Copyright Act, the WTO Panel found that the interests of rightsholders are unreasonably prejudiced if the exception or limitation causes or may cause a disproportionate loss of income, for example in the form of lost licensing revenue.[129] A corresponding focus on material losses can also be found in the WIPO commentary on Art. 9(2) Berne Convention.[130] Yet , the emphasis on economic interests has been criticised in academic literature: As some voices argue, in addition to pecuniary aspects, other regulatory objectives of the TRIPS Agreement as well as social and cultural interests of the WTO member states, in particular minimum standards of human-rights protection, should be taken into account.[131] In this regard, in order to justify an exception for AI training,

[125] *See*, *e.g. See* EUIPO, *The Development of Generative Artificial Intelligence from a Copyright Perspective* (2025). Available at https://www.euipo.europa.eu/en/news/euipo-releases-study-on-generative-artificial-intelligence-and-copyright; Keller (2024) Considerations for Opt-Out Compliance Policies by AI Model Developers, Open Future, May 16, 2024, https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf; European Commission, Study on Copyright and New Technologies: Copyright Data Management and Artificial Intelligence, European Commission, 2022, at 210. Available at https://op.europa.eu/publication-detail/-/publication/cc293085-a4da-11ec-83e1-01aa75ed71a1.

[126] Keller and Warso (2023), 7 *et seq.*

[127] Ricketson and Ginsburg (2022), para. 13.22; Kur (2009), pp. 287, 339.

[128] WIPO, Guide to the Berne Convention for the Protection of Literary and Artistic Works (Paris Act, 1971), 1978, Art. 9 para. 9.8; Oliver (2002), pp. 119, 165; Geiger, Gervais and Senftleben (2014), pp. 581, 595.

[129] WTO, Report of the Panel, United States – Section 110(5) of the US Copyright Act, WT/DS160/R (15 June 2000), para. 6.229 ("In our view, prejudice to the legitimate interests of rightholders reaches an unreasonable level if an exception or limitation causes or has the potential to cause an unreasonable loss of income to the copyright owner").

[130] WIPO, Guide to the Berne Convention for the Protection of Literary and Artistic Works (Paris Act, 1971), 1978, Art. 9.8.

[131] Ricketson and Ginsburg (2022), para. 13.22 *et seq.*; further also Kur (2009), pp. 287, 324, 340 *et seq.*; Lucas (2010), pp. 277, 278; Geiger, Gervais and Senftleben (2014), pp. 581, 601 *et seq.*

reference could be made to the public interest in fostering innovation. Yet there would also be aspects on the other side of the equation that would count, notably the authors' moral rights. Also, the fate of creativity as such – i.e., the subsistence of a world with a vibrant ecosphere of genuinely *human* creativity as the basis for both man-made and artificial production – arguably points towards a more restrictive conception and construction of exceptions and limitations for generative AI training. Ultimately, therefore, application of the TDM exception to generative AI training would hardly be an open-and-shut case under the third test level.

3.3.4 Conclusion

Training generative AI models falls outside the legal scope and doctrinal purpose of Art. 4 CDSM. First, the statutory aim of the TDM exception is to extract information in the form of patterns, trends, or correlations: this is conceptually and functionally incompatible with generative training, which reproduces and recombines expressive content to simulate human-created outputs. This constitutes a functional and teleological mismatch. Second, the training process aligns more closely with the reproduction right under Art. 2 InfoSoc, as it involves copying expression for the purpose of generating further expression, not for analytical use. This significantly raises the bar for compliance with copyright limitations and justifies a stricter interpretive approach. Third, the CJEU's consistent application of the three-step test demands a narrow construction of such limitations and exceptions. Generative training fails this test on multiple grounds: Although it may be a "special case", it conflicts with the normal exploitation of the work by displacing human authors in the market, and it unreasonably prejudices their legitimate economic and moral interests. Crucially, such conflict cannot be cured by opt-out mechanisms, especially given their limited effectiveness and incompatibility with international copyright standards prohibiting formalities. This means that the legal structure and policy rationale of Art. 4 CDSM cannot accommodate the large-scale ingestion of protected works for generative purposes. Rather, a new *sui generis* legal framework would be required to legitimise such uses: one that offers appropriate safeguards for rightsholders and aligns with international copyright obligations.

3.4 Question 4: Does AI-Generated Output That Includes Protected Content Require Authorisation?

Building on the earlier discussion of press-specific protections under Art. 15(1) of the CDSM Directive, this section turns to the broader and more foundational rights of reproduction and communication to the public as enshrined in Arts. 2 and 3 of the InfoSoc Directive. It assesses whether the generation and dissemination of AI outputs that incorporate protected material, either verbatim or in recognisable form, engage those exclusive rights.

A key issue is whether the probabilistic and synthetic character of generative outputs affects the scope of infringement, or whether traditional criteria, such as the recognisability of the author's expression and the accessibility of the output to the

public, continue to apply. Consistent with the principle of technological neutrality underpinning the InfoSoc Directive, exclusive rights are implicated whenever protected features of a work are reproduced or made perceptible to a new audience, regardless of the technical means employed.[132] This inquiry is especially relevant in scenarios where AI-generated content risks displacing the original in the market, thereby interfering with its normal exploitation and undermining the incentive structure that copyright is intended to protect.

3.4.1 The Reproduction Right: Recognisability Over Literal Copying

Article 2 of the InfoSoc Directive protects not only literal copying but also the reproduction of protected expression in recognisable form, even if mediated through complex technologies. In *Infopaq*, the Court of Justice determined that even minimal textual segments, such as eleven-word snippets, can fall within the scope of copyright protection, provided they exhibit the imprint of the author's intellectual creativity.[133] This standard emphasises the expressive quality and distinctiveness of the content over its length.

In the case of AI-generated outputs, the question is whether a string of text, or more broadly the structure, tone, or composition of an output, preserves the protected features of an original work. Even when the reproduction is non-verbatim, if it reflects protected choices, such as a unique narrative structure or journalistic phrasing, it may fall within the scope of Art. 2.[134]

The Court's reasoning in *Football Dataco* reinforces this view.[135] Although the case focused on the originality of football fixture lists, it affirmed that copyright protects original structures or arrangements that reflect the author's creative choices. This logic remains applicable in the context of generative AI, where the output may paraphrase or stylistically imitate original content in a recognisable manner. That expressive residue, even if statistically derived, can trigger reproduction rights when the output reveals the author's creative personality, as required under EU law.[136]

Indeed, as Advocate General Emiliou recently explained, outputs that simulate tone or stylistic idiosyncrasies, without replicating semantic or structural content, may still reflect protected expression if such stylistic features are sufficiently

[132] *See*, *e.g.*, Synodinou (2012), pp. 618–627.

[133] CJEU, case C-5/08, *Infopaq Int'l A/S v. Danske Dagblades Forening*, ECLI:EU:C:2009:465.

[134] CJEU, case C-5/08, *Infopaq Int'l A/S v. Danske Dagblades Forening*, ECLI:EU:C:2009:465, 2009 E.C.R. I-6569, paras. 37–48; CJEU, case C-476/17, *Pelham GmbH v. Hütter*, para. 29; Rosati (2019), pp. 87–89 (discussing reproduction "in part" and the originality standard in the context of expressive elements such as structure and phrasing).

[135] C-604/10, *Football Dataco Ltd and Others*, ECLI:EU:C:2012:115.

[136] *Ibidem* paras. 38, 41.

original.[137] Further nuance could be introduced by distinguishing between semantic recognisability (the retention of specific ideas or formulations) and aesthetic recognisability (the imitation of protected stylistic elements), suggesting a two-tiered test for assessing infringement in probabilistic outputs. We suggest a nuanced approach: outputs could infringe either by reproducing semantic content (the facts or specific expressions from the work) or by reproducing aesthetic/stylistic elements that are original. EU law's emphasis on the author's personal expression supports action against both forms of copying. For instance, an AI-generated text might not quote any exact sentences from a novel, yet capture the novel's unique narrative voice or structure. If that creative pattern is recognisable and attributable to the author, it implicates the reproduction right as well.

3.4.2 The Communication to the Public Right: Accessibility and New Audiences

Generative outputs may also infringe the right of communication to the public under Art. 3 of the InfoSoc Directive. According to settled case law[138] this right is triggered when (i) content is made perceptible to the public and (ii) access is granted to a new public not contemplated by the rightsholder. The use of automated or AI-driven systems does not exempt an act from qualifying as "communication." Chatbot responses or outputs made available via interfaces or APIs are functionally equivalent to transmissions that the Court has already deemed communicative under EU law.

In generative contexts, a new public is reached when users gain access to expressive content that bypasses the original licensing framework, especially when the AI output substitutes access to licensed platforms or paywalled sources. This divergence supports the finding of unauthorised communication. Moreover, the notion of a "new public" is especially relevant in environments where AI services repackage content in formats, such as chat replies or voice assistants, that were not foreseen by the original licensing scheme.[139] Notably, Gemini's response did not provide a link or reference to the original article, meaning users access the content entirely through a new channel. This differs even from hyperlinking cases: the user is not simply being pointed to the original location, but is being presented the content detached from its source. This amounts more to an unauthorised re-publication than a mere reference. The medium may be novel, but the legal question

[137] *Cf.* also General Advocate Emiliou in Case C-590/23, CG, *YN v. Pelham GmbH et al.*, para. 66 ("[W]hen it comes to imitative artworks, the line between the borrowing of unprotected elements and the reproduction of protected material is tenuous. A series of recent lawsuits saw authors claiming that their 'unique' style qualifies as 'original' expression. Another series … saw rightholders alleging copyright infringement over the reuse of stylistic features present in their works. That is even truer where the artwork imitates closely the style of a single work. The elements borrowed, while 'stylistic', could still be regarded as original, especially when combined.").

[138] CJEU, case C-466/12, *Svensson v. Retriever Sverige AB*, EU:C:2014:76; further also CJEU, case C-117/15, *Reha Training Gesellschaft für Sport- und Unfallrehabilitation mbH v. Gesellschaft für musikalische Aufführungs- und mechanische Vervielfältigungsrechte (GEMA)*, ECLI:EU:C:2016:379.

[139] *See*, *e.g.* Rosati (2020) (noting that communication using different technical means or targeting a public not envisaged in the original licence may constitute a communication to a "new public" under Art. 3(1) InfoSoc Directive).

remains whether the protected content reaches individuals who were not part of the rightsholder's intended audience. Even where the original publication was openly accessible online, the publisher did so with the intention of readers coming via its platform (with whatever conditions or revenue model applied). When an AI system delivers the content without directing the user to the source, it effectively communicates the work to a public in a manner unanticipated by the rightsholder. This is akin to the Court's finding of a new public when content was accessed through a different technical means or platform (see, e.g., TVCatchup, where even freely broadcast TV when retransmitted online was to a new public).

3.4.3 Technological Mediation Does Not Immunise Infringement

One of the central issues in this context is whether the probabilistic or synthetic nature of AI output alters the legal analysis. However, CJEU jurisprudence clearly indicates that the use of new technologies does not exempt users from copyright scrutiny if protected material remains recognisable. In *Pelham*, the Court held that sampling a short musical segment, despite using a new sound technology, still triggered copyright liability where the sample was recognisable in the final work. Likewise, in *Funke Medien*, the Court confirmed that the unauthorised reuse of factual government reports could amount to infringement if the expression of the original was appropriated.[140] These cases collectively affirm that EU copyright law applies independently of the technical method used, and that recognisability of protected material remains the decisive factor.

Thus, whether the reproduction is deterministic or probabilistic, it is the recognisability of the protected expression that remains legally decisive. Generative systems are subject to the same legal thresholds: if their outputs reproduce or paraphrase protected works recognisably, the reproduction and/or communication rights are triggered. What matters is not how the expression is stored or accessed, but whether it can be perceived in a form that reveals the original work's intellectual substance. Generative AI may be technically opaque, but legal responsibility attaches to the effects (i.e., the output), not to the internal logic of the model. This reinforces the principle that legal thresholds should remain robust even as technical architectures evolve.

3.4.4 Market Substitution and the Three-Step Test

Even when the outputs do not reproduce full passages, their capacity to substitute for the original work raises concerns under the three-step test codified in Art. 5(5) of the InfoSoc Directive. As seen already, the test limits exceptions to cases that do not conflict with the normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the rightsholder.[141] Such outputs can replicate the

[140] *See Pelham*, C-476/17, paras. 31–32 (recognisability of the sample renders reproduction unlawful despite technological transformation); and *Funke Medien*, C-469/17, para. 49 (novel technical processes do not exempt use from infringement assessment).

[141] *See* supra.

informational or stylistic value of the original work, thereby short-circuiting its monetisation model, especially when AI services are used as functional replacements for traditional content access. This economic substitution risk extends across sectors and formats, not only in journalism but also in academic, technical, and artistic content. It may be useful to distinguish between sectors in terms of substitution risk. For instance, AI-generated summaries of journalistic articles may immediately displace user traffic, while AI renderings of literary or poetic works might affect derivative or transformative markets. A tentative typology could classify substitution risks as: (i) direct functional substitution (e.g., news content); (ii) partial derivative substitution (e.g., summarised scientific papers); and (iii) stylistic or aesthetic imitation (e.g., artworks, fiction). While all may trigger legal concerns, the degree of market harm and the appropriate remedy may vary accordingly.[142] While all these scenarios raise copyright concerns, the legal analysis, such as whether exceptions apply or how substantial the harm is, may differ in each case. We highlight this typology not as a doctrinal test, but as analytical context to illustrate the breadth of market disruption posed by generative AI. This framing supports a strict interpretation of the three-step test and reinforces the need for a narrow construction of exceptions.

As a result, AI outputs that interfere with commercial pathways, without falling within a narrow permissible exception, fail the second and third prongs of the three-step test and therefore require prior authorisation. This holds true even where the AI-generated version is shorter or less detailed than the original. What matters is functional equivalence: if the output satisfies the user's informational demand in a way that obviates access to the original work, it undermines the work's commercial value and defeats the justification for exceptions.

3.4.5 Conclusion

AI-generated outputs that reproduce protected content in recognisable form, whether through direct quotation, paraphrased restatement, or stylistic imitation, implicate the exclusive rights of reproduction and communication to the public. What matters is the perceptibility and expressive fidelity of the output, its unauthorised accessibility, and its ability to function as a market substitute. As the CJEU's case law confirms, the neutrality of EU copyright law towards technological methods implies that AI services cannot evade liability simply by relying on synthetic or opaque generation processes. Where such outputs displace demand for the original or reroute it through unlicensed channels, authorisation is required.

4 Summary

The *Like Company v. Google Ireland* case presents the CJEU with an unprecedented opportunity to clarify the application of core principles of EU copyright law to the

[142] For a similar distinction according to different categories of works *see* Judge Chhabria in *Richard Kadrey et al. v. Meta Platforms, Inc*., Case 3:23-cv-03417-VC (June 25, 2025), *29–30.

domain of generative AI. The factual and legal issues raised by this dispute reach beyond the specific circumstances of the parties and speak to broader concerns about the compatibility of large-scale AI training and output generation with the *acquis Communautaire*, in particular the InfoSoc and CDSM Directives.

This article can be understood as an amicus curiae brief. It has highlighted four key areas in which legal guidance is essential. First, the unauthorised communication to the public of protected press content through AI-generated responses may fall within the scope of Arts. 3 InfoSoc and Art. 15(1) CDSM Directive, especially where such outputs substitute user access to the original work. Second, the process of training large language models systematically entails acts of reproduction within the meaning of Art. 2 of the InfoSoc Directive, regardless of whether the expressive elements are stored in literal or statistical form. Third, the application of the text and data mining exception under Art. 4 of the CDSM Directive to generative AI training is doctrinally and functionally misplaced. The objectives and technical assumptions underlying the TDM framework diverge fundamentally from the expressive, reconstructive nature of generative AI systems. Moreover, such training fails to satisfy the conditions of the three-step test, particularly with regard to the normal exploitation of the work and the legitimate interests of rightsholders. Finally, where generative outputs contain recognisable fragments or stylistic features derived from protected works, both the reproduction and communication rights can be engaged, necessitating prior authorisation.

Read together, our analysis yields three consequences for application.[143] First, training that stores parameters permitting the reconstruction of recognisable protected expression constitutes reproduction within Art. 2 of the InfoSoc Directive and does not fall within Art. 4 of the CDSM Directive under the three-step test. Second, copies created in retrieval for user-facing generation and delivery will ordinarily not satisfy the conditions of Art. 5(1), although genuinely analytical indexing may be assessed under Art. 4 of the CDSM Directive. Third, where user-facing responses reproduce protected material in a manner that substitutes for access to the source, the reproduction and communication to the public rights under Arts. 2 and 3 of the InfoSoc Directive are engaged and, for press publications, the related right in Art. 15(1) of the CDSM Directive.

These findings suggest that the unchecked deployment of generative AI models trained on copyright-protected content risks undermining the economic foundations of creative industries and eroding the incentive structures that EU copyright law is designed to protect.[144] The Court is therefore invited to adopt a principled and technologically informed interpretation of the relevant provisions, ensuring that

[143] For the articulation of a three-pillar accountability test and its application to generative-AI copyright, *see* Lucchi (2025), p. 112.

[144] Recent policy discussions at the European Parliament reflect a similar concern. A motion for a resolution calls for enhanced transparency, effective opt-out mechanisms, and a reassessment of existing copyright exceptions in light of the disruptive effects of generative AI. *See* European Parliament, Commitee on Legal Affairs, DRAFT REPORT on Copyright and generative artificial intelligence – opportunities and challenges (Motion for a Resolution) (INI/2023/2057), available at https://oeil.secure.europarl.europa.eu/oeil/en/procedure-file?reference=2025/2058(INI).

innovation does not come at the expense of legal certainty, economic fairness, and the continued viability of human authorship.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Baumann M (2023) Generative KI und Urheberrecht—Urheber und Anwender im Spannungsfeld. NJW 76:3673–3678

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? In: FAccT '21: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 610–623. https://doi.org/10.1145/3442188.3445922

Brauneis R (2025) Copyright and the training of human authors and generative machines. Columbia J Law Arts 48:1–59. https://doi.org/10.52214/jla.v48i1.13529

Buick A (2025) Copyright and AI training data—transparency to the rescue? J Intellect Prop Law Pract 20:182–192. https://doi.org/10.1093/jiplp/jpae102

Calderón R (2024) AI training through copyrighted works as infringement: perspectives under the Berne three-step test and the Pane fair use test and plausible solutions. IDEA 65 Special Issue:84–103

Carlini N, Tramèr F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson U, Oprea A, Raffel C (2021) Extracting training data from large language models. In: Proceedings of the 30th USENIX security symposium (USENIX Security 21), pp 2633–2650. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

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. In: Calandrino J, Troncoso C (eds) Proceedings of the 32nd USENIX security symposium (USENIX Security 23), pp 5253–5270. https://doi.org/10.48550/arXiv.2301.13188

Carlini N, Ippolito D, Jagielski M, Lee K, Tramèr F, Zhang C (2023) Quantifying memorization across neural language models. In: Proceedings of the 11th international conference on learning representations (ICLR). https://doi.org/10.48550/arXiv.2202.07646

Carter S, Nielsen M (2017) Using artificial intelligence to augment human intelligence. Distill. https://doi.org/10.23915/distill.00009

Chang K, Cramer M, Soni S, Bamman D (2023) Speak, memory: an archaeology of books known to ChatGPT/GPT-4. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 conference in empirical methods in natural language processing, pp 7312–7327. https://doi.org/10.18653/v1/2023.emnlp-main.453

Colangelo G, Torti V (2019) Copyright, online news publishing and aggregators: a law and economics analysis of the EU reform. Int J Law Inf Technol 27:75–90

Cooper AF, Grimmelmann J (2025) The files are in the computer: on copyright, memorization, and generative AI. Chic-Kent Law Rev 100:141–219

Cooper AF, Gokaslan A, Ahmed A, Cyphert AB, De Sa C, Lemley MA, Ho DE, Liang P (2025) Extracting memorized pieces of (copyrighted) books from open-weight language models. arXiv:2505.12546v1 [cs.CL]. https://doi.org/10.48550/arXiv.2505.12546

Coster H (2023) Global news publisher Axel Springer partners with OpenAI in landmark deal. Reuters. https://www.reuters.com/business/media-telecom/global-news-publisher-axel-springer-partners-with-openai-landmark-deal-2023-12-13/

Davies P (2024) OpenAI partners with European media giants in France and Spain to use content for training. Euronews. https://www.euronews.com/next/2024/03/14/openai-partners-with-european-media-giants-in-france-and-spain-to-use-content-for-training

Dornis TW (2024a) Generative KI, urheberrechtliche Vervielfältigung und öffentliche Zugänglichmachung—Teil 1: Das Modellinnere. Computer & Recht (CR) 40:765–772. https://doi.org/10.9785/cr-2024-401118

Dornis TW (2024b) Generative KI, urheberrechtliche Vervielfältigung und öffentliche Zugänglichmachung—Teil 2: Die öffentliche Zugänglichmachung. Computer & Recht (CR) 40:830–839. https://doi.org/10.9785/cr-2024-401220

Dornis TW (2024c) Generatives KI-Training und Text- und Data-Mining—Eine funktionale Unterscheidung. KIR 1:156–161

Dornis TW (2025) Generative AI, reproductions inside the model, and the making available to the public. IIC 56:909–938. https://doi.org/10.1007/s40319-025-01582-9

Dornis TW (2025b) The training of generative AI is not text and data mining. E.I.P.R. 47:65–78

Dornis TW, Stober S (2024) Urheberrecht und Training generativer KI-Modelle. Nomos, Baden-Baden. https://doi.org/10.5771/9783748949558

Ducato R, Strowel A (2021) Ensuring text and data mining: remaining issues with the EU copyright exceptions and possible ways out. E.I.P.R. 43:322–337

Dusollier S, Kretschmer M, Margoni T, Mezei P, Quintais JP, Rognstad O-A (2025) Copyright and generative AI: opinion. J Intellect Prop Inf Technol E-Com Law 16:121–127

European Commission (2022) Study on copyright and new technologies – copyright data management and artificial intelligence. https://doi.org/10.2759/570559

European Parliament Committee on Legal Affairs (2025) Draft report on copyright and generative artificial intelligence—opportunities and challenges. 2025/2058 (INI). https://oeil.secure.europarl.europa.eu/oeil/en/procedure-file?reference=2025/2058(INI)

European Union Intellectual Property Office (EUIPO) (2025) The development of generative artificial intelligence from a copyright perspective. https://doi.org/10.2814/3893780

Fernández-Molina J, de la Rosa FE (2024) Copyright and text and data mining: is the current legislation sufficient and adequate? Portal Libr Acad 24:653–672. https://doi.org/10.1353/pla.2024.a931775

Geiger C, Gervais D, Senftleben M (2014) The three-step test revisited: how to use the test's flexibility in national copyright law. Am Univ Int Law Rev 29:581–626

Gervais D (2019) Exploring the interfaces between big data and intellectual property law. J Intellect Prop Inf Technol E-Commer Law 10:1–19

Gervais D, Marmanis H, Shemtov N, Rowland CZ (2024) The heart of the matter: copyright, AI training, and LLMs. J Copyright Soc 71:482–517

Ginsburg JC (2016) Berne-forbidden formalities and mass digitalization. Boston Univ Law Rev 96:745–775

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge

Hayes J, Swanberg M, Chaudhari H, Yona I, Shumailov I, Nasr M, Choquette-Choo CA, Lee K, Cooper AF (2025) Measuring memorization through probabilistic discoverable extraction. arXiv:2410.19482 [cs.LG]. https://doi.org/10.48550/arXiv.2410.19482

Heerma JD (2022) § 16 UrhG. In: Wandtke A-A, Bullinger W (eds) Praxiskommentar Urheberrecht, 6th edn. C.H. Beck, München, pp 311–321

Henderson P, Li X, Jurafsky D, Hashimoto T, Lemley MA, Liang P (2023) Foundation models and fair use. J Mach Learn Res 24:400

Hofmann F (2024) Zehn Thesen zu Künstlicher Intelligenz (KI) und Urheberrecht. WRP 11–18

Käde L (2021) Kreative Maschinen und Urheberrecht—Die Machine Learning-Werkschöpfungskette vom Training über Modellschutz bis zu Computational Creativity. Nomos, Baden-Baden. https://doi.org/10.5771/9783748912453

Keller P (2024) Considerations for opt-out compliance policies by AI model developers. Open Future Policy Brief #6. https://openfuture.eu/publication/considerations-for-implementing-rightholder-opt-outs-by-ai-model-developers/

Keller P, Warso Z (2023) Defining best practices for opting out of ML training. Open Future Policy Brief #5. https://openfuture.eu/publication/defining-best-practices-for-opting-out-of-ml-training

Konertz R, Schönhof R (2024) Vervielfältigungen und die Text- und Data-Mining-Schranke beim Training von (generativer) Künstlicher Intelligenz. WRP 289–296
Kur A (2009) Of oceans, islands, and inland water—how much room for exceptions and limitations under the three-step test? Richmond J Glob Law Bus 8:287–350
Lee K, Ippolito D, Nystrom A, Zhang C, Eck D, Callison-Burch C, Carlini N (2022) Deduplicating training data makes language models better. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers), pp 8424–8445. https://doi.org/10.18653/v1/2022.acl-long.577
Lee J, Le T, Chen J, Lee D (2023) Do language models plagiarize? In: Ding Y, Tang J, Sequeda J (eds) WWW '23: proceedings of the ACM WebConference 2023, pp 3637–3647. https://doi.org/10.1145/3543507.3583199
Lemley MA, Casey B (2021) Fair learning. Tex Law Rev 99:743–785
Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-T, Rocktäschel T, Riedel S, Kiela D (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Proceedings of the 34th international conference on neural information processing systems (NIPS '20), pp 9459–9474. https://doi.org/10.48550/arXiv.2005.11401
Lucas A (2010) For a reasonable interpretation of the three-step test. E.I.P.R. 32:277–282
Lucchi N (2025) Generative AI & copyright: training, creation, regulation. Study commissioned by the Parliament's Policy Department for Justice, Civil Liberties and Institutional Affairs at the request of the Committee on Legal Affairs. PE 774.095. https://www.europarl.europa.eu/thinktank/en/document/IUST_STU(2025)774095
Lux H, Noll CJ (2024) Of books and bytes: the copyright dilemma in AI development. Transatl Law J 2:111–115
Margoni T, Kretschmer M (2022) A deeper look into the EU text and data mining exceptions: harmonisation, data ownership, and the future of technology. GRUR Int 71:685–701
McDonagh L (2022) Directive 2019/790/EU (directive on copyright and related rights in the digital single market). In: Lodder A, Murray A (eds) EU regulation of e-commerce: a commentary. Edward Elgar, Cheltenham, pp 308–333. https://doi.org/10.4337/9781800372092.00017
Mezei P (2024) A saviour or a dead end? Reservation of rights in the age of generative AI. E.I.P.R. 46:461–469
Mitchell M, Krakauer DC (2023) The debate over understanding in AI's large language models. Proc Nat'l Acad Sci 120:e2215907120. https://doi.org/10.1073/pnas.2215907120
Ni B, Liu Z, Wang L, Lei Y, Zhao Y, Cheng X, Zeng Q, Dong L, Xia Y, Kenthapadi K, Rossi R, Dernoncourt F, Tanjim MM, Ahmed N, Liu X, Fan W, Blasch E, Wang Y, Jiang M, Derr T (2025) Towards trustworthy retrieval augmented generation for large language models: a survey. arXiv:2502.06872 [cs.CL]. https://doi.org/10.48550/arXiv.2502.06872
Nordemann JB, Pukas J (2022) Copyright exceptions for AI training data—will there be an international level playing field? J Intellect Prop Law Pract 17:973–974. https://doi.org/10.1093/jiplp/jpac106
O'Brien M (2023) ChatGPT-maker OpenAI signs deal with AP to license news stories. The Associated Press. https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a
Oliver J (2002) Copyright in the WTO: the panel decision on the three-step test. Columbia J Law Arts 25:119–170
Peukert A (2005) A bipolar copyright system for the digital network environment. Hastings Commun Entertain Law J 28:1–80
Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning transferable visual models from natural language supervision. In: Meila M, Zhang T (eds) Proceedings of the 38th international conference on machine learning (PMLR), vol 139, pp 8748–8763. https://doi.org/10.48550/arXiv.2103.00020
Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125 [cs.CV]. https://doi.org/10.48550/arXiv.2204.06125
Ricketson S, Ginsburg J (2022) International copyright and neighbouring rights—the Berne Convention and beyond, 3d edn. Oxford University Press, Oxford
Rosati E (2019) Copyright and the court of justice of the European Union. Oxford University Press, Oxford
Rosati E (2020) When does a communication to the public under EU copyright law need to be to a "new public"? Eur Law Rev 45:802–823

Rosati E (2021) Copyright in the digital single market: article-by-article commentary to the provisions of Directive 2019/790. Oxford University Press, Oxford
Rosati E (2024) Infringing AI: liability for AI-generated outputs under international, EU, and UK copyright law. Eur J Risk Regul 16:603–627. https://doi.org/10.1017/err.2024.72
Rosati E (2024) Is text and data mining synonymous with AI training? J Intellect Prop Law Pract 19:851–852. https://doi.org/10.1093/jiplp/jpae092
Rosati E (2025) Copyright exceptions and fair use defences for AI training done for "research" and "learning," or the inescapable licensing horizon. Eur J Risk Regul 1–24
Sag M (2024) Fairness and fair use in generative AI. Fordham Law Rev 92:1887–1921
Scalzini S (2021) The new related right for press publishers: what way forward? In: Rosati E (ed) The Routledge handbook of EU copyright law. Routledge, London, pp 101–119
Schack H (2021) Schutzgegenstand, "Ausnahmen oder Beschränkungen" des Urheberrechts. GRUR 123:904–909
Schack H (2024) Auslesen von Webseiten zu KI-Trainingszwecken als Urheberrechtsverletzung de lege lata et ferenda. NJW 77:113–117
Schulze G (2022) § 16 UrhG. In: Dreier T, Schulze G (eds) Urheberrechtsgesetz, 7th edn. C.H. Beck, München, pp 347–360
Senftleben M (2004) Copyright, limitations and the three-step test. an analysis of the three-step test in international and EC copyright law. Kluwer Law International, Den Haag. https://hdl.handle.net/11245/1.224623
Senftleben M (2014) How to overcome the normal exploitation obstacle: opt-out formalities, embargo periods, and the international three-step test. Berkeley Technol Law J Comment 1:1–19
Senftleben M (2023) Generative AI and author remuneration. IIC 54:1535–1560. https://doi.org/10.1007/s40319-023-01399-4
Sesing-Wagenpfeil A (2024) Trainierte KI-Modelle als Vervielfältigungsstücke im Sinne des Urheberrechts. ZGE 16:212–268. https://doi.org/10.1628/zge-2024-0014
Singh J (2025) New York Times partners with Amazon for first AI licensing deal. Reuters. https://www.reuters.com/business/retail-consumer/new-york-times-amazon-sign-ai-licensing-deal-2025-05-29/
Škiljić A (2021) When art meets technology or vice versa: key challenges at the crossroads of AI-generated artworks and copyright law. IIC 52:1338–1369. https://doi.org/10.1007/s40319-021-01119-w
Sobel B (2020) A taxonomy of training data—disentangling the mismatched rights, remedies, and rationales for restricting machine learning. In: Lee J, Hilty R, Liu K (eds) Artificial intelligence & intellectual property. Oxford University Press, Oxford, pp 221–242. https://doi.org/10.1093/oso/9780198870944.003.0011
Sobel B (2024) Elements of style: copyright, similarity, and generative AI. Harv J Law Technol 38:49–106
Somepalli G, Singla V, Goldblum M, Geiping J, Goldstein T (2022) Diffusion art or digital forgery? Investigating data replication in diffusion models. arXiv:2212.03860 [cs.LG]. https://doi.org/10.48550/arXiv.2212.03860
Somepalli G, Singla V, Goldblum M, Geiping J, Goldstein T (2023) Understanding and mitigating copying in diffusion models. arXiv:2305.20086v1 [cs.LG]. https://doi.org/10.48550/arXiv.2305.20086
Steinrötter B, Schauer LM (2021) Text und Data Mining, Forschung und Lehre. In: Barudi M (ed) Das neue Urheberrecht, Nomos, Baden-Baden, pp 145–164
Synodinou T-E (2012) The principle of technological neutrality in European copyright law. E.I.P.R. 34:618–627
Tyagi K (2024) Copyright, text & data mining and the innovation dimension of generative AI. J Intellect Prop Pract 19:557–570. https://doi.org/10.1093/jiplp/jpae028
United States Copyright Office (2025) Copyright and artificial intelligence—part 3: generative AI training (pre-publication version). https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf
Vesala J (2023) Developing artificial intelligence-based content creation: are EU copyright and antitrust law fit for purpose? IIC 54:351–380. https://doi.org/10.1007/s40319-023-01301-2
Welser M v (2023) Generative KI und Urheberrechtsschranken. GRUR-Prax 15:516–520
World Intellectual Property Organization (1978) Guide to the Berne convention for the protection of literary and artistic works (Paris Act, 1971). WIPO, Geneva

World Trade Organization (2000) United States - Section 110(5) of the US Copyright Act. Report of the Panel, WT/DS160/R. https://www.wto.org/english/tratop_e/dispu_e/cases_e/ds160_e.htm

Wymeersch P (2023) EU copyright exceptions and limitations and the three-step test: one step forward, two steps back. GRUR Int 72:631–642

Zhou Y, Liu Y, Li X, Jin J, Qian H, Liu Z, Li C, Dou Z, Ho T-Y, Yu PS (2024) Trustworthiness in retrieval-augmented generation systems: a survey. arXiv:2409.10102 [cs.IR]. https://doi.org/10.48550/arXiv.2409.10102