

City Research Online

City St George's, University of London

Citation: Bonadio, E., Frosio, G., Geiger, C., Guadamuz, A., Karapapa, S. & Stamatoudi, I. (2026). Preserving balance in the EU digital single market: How Like Company could reframe copyright and innovation in the generative AI era. The Journal of World Intellectual Property, doi: 10.1111/jwip.70032

This is the published version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/38605/

Link to published version: https://doi.org/10.1111/jwip.70032

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Check for updates

Received: 3 March 2026 | Accepted: 9 July 2026

DOI: 10.1111/jwip.70032

ORIGINAL ARTICLE

THE JOURNAL OF WORLD INTELLECTUAL PROPERTY WILEY

Preserving balance in the EU digital single market: How *Like Company* could reframe copyright and innovation in the generative AI era

Enrico Bonadio[1] | **Giancarlo Frosio**[2] | **Christophe Geiger**[3] | **Andres Guadamuz**[4] | **Stavroula Karapapa**[5] | **Irini Stamatoudi**[6]

[1]University of London

[2]Queen's University Belfast

[3]Luiss Guido Carli University

[4]University of Sussex

[5]University of Essex

[6]University of Nicosia

Correspondence
Andres Guadamuz, University of Sussex, United Kingdom.
Email: a.guadamuz@sussex.ac.uk

Correction added on 29 September 2026, after first online publication: For this article, originally published online 27 July 2026, an incorrect version of the revised manuscript had been typeset and published online by mistake. The authors discovered this error and the incorrect version was temporarily removed from online publication on 19 August 2026. The authors provided the correct, final version of the peer-reviewed and extensively revised manuscript for typesetting.

Abstract

This article analyses the Court of Justice of the European Union's first referral on generative AI and copyright, *i.e.*, Case C-250/25 *Like Company v. Google*, which interrogates whether the generation of chatbot summaries of press articles infringes the press publishers' right under Article 15 CDSM Directive and the reproduction and communication rights under Articles 2 and 3 InfoSoc Directive. The analysis shows that large language model training cannot, as such, be characterised as reproduction because it encodes statistical patterns rather than fixing protected expression, while recognising that further empirical evidence would be valuable to clarify the precise technical steps involved in the training pipeline. RAG-generated summaries that paraphrase content within 'very short extracts' fall outside Article 15(1) CDSM and do not communicate publications to a new public. Alternatively, even if training were characterised as engaging temporary reproductions, the text-and-data mining exception under Article 4 CDSM Directive lawfully authorises such uses absent valid opt-outs. Regarding liability, the article concludes that individual users, not platform providers, bear responsibility for outputs

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2026 The Author(s). *The Journal of World Intellectual Property* published by John Wiley & Sons Ltd.

generated through their models' learned parameters and retrieval mechanisms. This comprehensive analysis ensures coherence across EU copyright doctrine and sustains legal certainty for AI innovation within the EU digital single market.

1 | INTRODUCTION

The intersection of generative artificial intelligence ('GenAI') and copyright law has rapidly evolved from a theoretical horizon into a pressing arena of doctrinal conflict, reshaping long-standing assumptions about infringement and lawful uses of protected works in the digital environment. Against this backdrop, this article offers an analysis of *Like Company v Google Ireland Limited*,[1] which represents an important moment for the EU digital single market, marking the first instance where the Court of Justice of the European Union ('CJEU' or 'Court') is directly petitioned to interpret the InfoSoc Directive and the Copyright in the Digital Single Market Directive ('CDSM')[2] in the specific context of Large Language Models ('LLMs'). The case is modest in its facts but potentially large in its consequences: a short press article, a chatbot summary, and four questions capable of shaping the future relationship between EU copyright law and GenAI systems.

The dispute concerns Google's GenAI chatbot Gemini, formerly Bard, made available to users in the European Economic Area. *Like Company*, a Hungarian media publisher operating several news portals, including the regional site Balaton Környéke, published a 579-word article on 21 July 2023 about the singer Kozsó's plan to introduce dolphins into Lake Balaton.[3] According to the referring court, the claimant later tested Gemini by entering a prompt that asked for a Hungarian summary of the specific online press publication appearing on balatonkornyeke.hu concerning that plan. Gemini generated a summary that accurately recapitulated the main information contained in the article.[4] *Like Company* then brought proceedings before the Budapest Környéki Törvényszék, alleging that Google had made electronic copies of its online news articles and had made them available to the public without authorisation.[5] Google denied that the chatbot output involved reproduction or making available to the public and, in the alternative, relied on the text-and-data-mining (TDM) exception under Article 4 CDSM and the temporary reproduction exception under Article 5(1) InfoSoc.[6]

The publicly available record does not disclose the precise technical pathway by which the contested output was produced. It cannot be inferred with certainty whether the response resulted from inference-time retrieval or grounding, from user-supplied text, from indexed or cached material, from third-party reproductions, or from some combination of these possibilities. This matters because the legal consequences differ depending on whether the case is treated as one about training, retrieval, output reproduction, or user attribution. The article therefore proceeds on the basis of publicly available material only and separates the relevant technical possibilities rather than assuming that the facts categorically establish one pathway. That separation is essential: training, retrieval-augmented generation, and user-prompted output are not the same technological act and should not be collapsed into one legal category.

On 3 April 2025, the Hungarian court stayed the proceedings and referred four questions to the CJEU. The first asks whether chatbot responses that partially match press publishers' web pages amount to communication to the public under Article 15 CDSM and Article 3(2) InfoSoc. The second asks whether training an LLM constitutes reproduction within the meaning of Article 2 InfoSoc. The third, contingent on an affirmative answer to the second, asks whether such training falls within the TDM exception in Article 4 CDSM. The fourth asks whether user-prompted outputs that reproduce press publication content should be attributed to the chatbot service provider. These questions do not merely concern one Hungarian article. They require the Court to clarify how EU copyright law distinguishes learning from copying, summaries from reproductions, retrieval from communication, and user agency from provider responsibility.

Although the reference is triggered by the press publishers' right (PPR) in Article 15 CDSM, the deeper issue is the relationship between that specific regime and the general rights in Articles 2 and 3 InfoSoc. The referring court asks, in substance, whether the general communication to the public right may exclude the application of the more

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

specific PPR. That formulation is important. Article 15 CDSM was introduced as a calibrated regime for the online use of press publications by information society service providers. It contains deliberate internal limits, including the exclusion of hyperlinks, individual words and very short extracts, as well as rules on the relationship between publishers and authors. If the same conduct could simply be recharacterised under the general InfoSoc rights whenever Article 15's limits are inconvenient, the legislative compromise would be hollowed out. This article therefore treats Article 15 not as a redundant add-on, but as the specific interface through which online uses of press publications should primarily be assessed.

The article proceeds functionally rather than by mechanically following the sequence of the four referred questions. Section 2 sets out the technical foundations, explaining how GenAI systems such as Gemini operate, including training, memorisation and retrieval-augmented generation. Section 3 examines the PPR under Article 15 CDSM, its scope, limits and relationship with the InfoSoc regime. Section 4 analyses the reproduction right and explains why training, properly characterised, should not be treated as reproduction of protected expression. Section 5 considers the TDM exception and its relevance to AI training and inference-stage processing. Section 6 turns to communication to the public and the treatment of summaries, snippets and very short extracts. Section 7 addresses platform and user liability, focusing on volition, attribution, intermediary responsibility and workable remedies. Section 8 concludes.

Although the authors' analysis supports the view that GenAI models such as Gemini can operate within the existing copyright framework and do not inherently violate established legal principles, the article's doctrinal position should not be mistaken for indifference to creator remuneration. The fact that GenAI systems can, in principle, operate within the existing copyright framework does not settle the broader political economy of value allocation. Authors, journalists, publishers and other creators supply much of the cultural and informational material from which GenAI systems derive utility. Whether through improved licensing infrastructures, collective management, statutory remuneration, or other mechanisms, EU law may need to develop fairer ways of ensuring that value generated by AI systems is not captured only at the platform layer.[7] The point of the present article is narrower but foundational: before designing new remuneration models, the CJEU should preserve the conceptual architecture of copyright law by keeping training, retrieval, outputs and liability analytically distinct.

2 | TECHNICAL ANALYSIS

There are various working definitions of the term 'artificial intelligence'.[8] In the context of this article, the focus is on large-scale machine-learning systems, especially so-called foundation models such as LLMs and large multimodal models, that are trained on vast corpora of digital material to produce general-purpose capabilities.[9] This section explains how such models, and specifically Google's Gemini system deployed in *Like Company*, function, in order to ground the subsequent legal discussion on reproduction, communication to the public, TDM and platform liability in that concrete technical setting.

2.1 | The Gemini model

Google's Gemini system is a GenAI chatbot built on an LLM architecture. In the *Like Company* scenario, the relevant system is Gemini v1.0. Rather than storing static, human-readable copies of training data, Gemini processes input texts by breaking them down into minimum units known as 'tokens'.[10] During training the model learns statistical patterns that govern how tokens co-occur. The resulting system does not retrieve pre-existing documents upon request; it generates new text, code, or summaries by predicting the next token in a sequence on the basis of probability distributions learned from the training corpus.[11]

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Operationally, Gemini thus behaves less like a database and more like a probabilistic language generator.[12] Because it relies on predictions, it is capable of producing both accurate summaries and 'hallucinations', where it generates inaccurate or fabricated content. In *Like Company*, the Hungarian-language summary of the Balaton Környéke article may therefore be analysed as a new linguistic realisation guided by patterns in data and (under the primary hypothesis explored here) by retrieval, rather than as the system 'playing back' a stored copy of the press publication. From a technical perspective, Gemini's core model is not an accessible archive of works but a set of learned parameters.[13]

2.2 | Training stages

At a high level, an LLM's life-cycle can be divided into (i) data gathering and pre-processing, (ii) training, and (iii) post-training and deployment. These stages are directly relevant to the *Like Company* dispute because the claimant alleges reproduction of its article during the input phase.

In the input phase, large datasets are assembled from multiple sources.[14] Raw data are then pre-processed: duplicates, corrupted files and irrelevant material are filtered out; data are cleaned and normalised; [15] and texts are tokenised into words or sub-word units, while other modalities (audio, video, images) are segmented into workable chunks.[16] This stage may also involve annotation or labelling.[17] Pre-processing serves to improve data quality, reduce bias and noise, and make the corpus computationally manageable so that the model learns meaningful patterns rather than artefacts of messy input.[18]

Training begins once this processed dataset is assembled into a training corpus.[19] At this point, temporary copies of the data necessarily exist within the technical pipeline, during ingestion, caching and tokenisation, to permit learning. These copies are functional and short-lived but, strictly speaking, the data consists of a transient copy. During training, the system converts input into numerical representations and adjusts its internal parameters (weights) to minimise prediction error. Gemini v1.0, like other LLMs, is therefore an architecture plus a very large array of numbers that encode statistical regularities such as how often words or visual elements co-occur.[20]

Two observations are here crucial for the subsequent legal analysis of the questions referred to the CJEU in *Like Company*. First, the model does not contain accessible copies of the works in the training data.[21] Models can memorise some material, but they are not static repositories from which users can browse or retrieve particular works at will. In normal operation, a trained model such as Gemini v1.0 cannot be regarded as making specific training works available to the public. Second, each individual work contributes only a negligible fraction to overall model capability: what matters is the aggregate information extracted from billions or trillions of tokens.[22] The marginal impact of a single news article, or even an entire regional news site such as Balaton Környéke, on the model's capabilities is vanishingly small.

In addition, there is a temporal constraint that is decisive in *Like Company*. Gemini v1.0 has a knowledge cut-off of February 2023, while the disputed article was published in July 2023.[23] The article therefore could not have been present in the training corpus of the specific model version that generated the impugned output. As a matter of technical fact, no reproduction of that press publication occurred during Gemini v1.0's training.

2.3 | Memorisation

Nonetheless, LLMs like Gemini can under some circumstances memorise and regurgitate specific items from their training data. Memorisation describes the situation where a model retains concrete snippets rather than only abstract patterns, so that it can reproduce them substantially when suitably prompted.[24] This phenomenon is

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

associated with overfitting, where the model has been trained too closely on particular examples, often in highly duplicated or repetitive data.[25]

Empirical studies show that memorisation is relatively rare and tends to affect highly repeated or salient sequences (famous poems, lyrics, well-known images), rather than ordinary texts such as a single regional news story.[26] It can be substantially reduced by deduplication of training data,[27] regularisation to discourage over-reliance on specific features,[28] and other techniques such as differential privacy, early stopping, and gradient clipping.[29] Memorisation therefore exists but is statistically bounded and technically mitigable; it does not convert Gemini or similar models into a database of their inputs.

2.4 | Post-training and retrieval-augmented generation

After pre-training, models like Gemini are often further optimised through post-training techniques. This phase does not fundamentally change the model's nature as a parametric function, but it shapes how it responds to prompts. One widely used method is Reinforcement Learning from Human Feedback, in which humans rate or rank outputs and their preferences are used to fine-tune the model so that it produces more helpful, safe and 'aligned' answers.[30] Post-training techniques include also processes based on so-called 'inference',[31] where a trained model applies its learned parameters to new data to produce an output, as in the case of retrieval-augmented generation (RAG).[32]

A RAG-based operation offers a primary hypothesis for understanding the *Like Company* facts. When users interact with Gemini, their prompts and histories of the conversations are packaged into a 'context window', *i.e.*, the temporary working memory for that interaction.[33] This is not part of the model, but an instance of interactions between the model and the user. Within the 'context window', RAG supplements what is in the model's weights, at the knowledge cut-off point, with information retrieved in real time from external sources (search indices, document repositories), or the web, at inference time.[34] The RAG component selects the most relevant passages, which are then injected into the 'context window' and supplied to the LLM. The model is instructed to generate an answer conditioned both on its learned parameters and on this fresh retrieved material. The effect is to bridge traditional information retrieval and GenAI: the system behaves partly like a search engine (finding relevant documents) and partly like a generator (synthesising an answer in natural language). This design allows models to stay up to date despite a fixed training cut-off and reduces hallucinations by grounding responses in external sources. It also enables citation or attribution of those sources and modular updates to the retrieval layer without retraining the entire model.

In Google's publicly described 2025 Gemini architectures, this retrieval-based workflow is complemented by an explicit 'grounding' phase.[35] The model issues one or more search queries to Google Search, processes the returned results into candidate answers, and selects a response that is explicitly anchored in those search results.[36] The search phase itself is a classic information-retrieval operation on an index; no additional 'making available' of content occurs inside the model beyond what is already possible via Google Search.[37] The generative component then reformulates and condenses that content for the specific user, within the confines of the context window.[38]

2.5 | Technical characteristics of the *Like Company* case

Against this background, the *Like Company* dispute turns on two technical assertions. As mentioned, *Like Company* alleges, first, that Gemini was trained on its content during the input phase, thus allegedly infringing its right of reproduction; and second, that the chatbot's detailed summaries amount to making it available to the public.[39] The claimant's core intuition is that such summaries would only be possible if Gemini had first reproduced the article in its entirety during training.

But as already noted, the publicly reported timing of Gemini v1.0's training and the publication date of the *Balaton Környéke* article point in the opposite direction: the article appears to post-date the relevant model's knowledge

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

cut-off.[40] On that basis, the page could not have been included in the model's training corpus as a matter of timing. Accordingly, on the best reading of the public record, there was no reproduction of that article at any stage of training for the model used in these proceedings. Nor could Gemini v1.0, as a set of parameters, make that specific article available to the public, because it neither contained the article nor allowed users to retrieve it as such.

Moreover, even apart from timing, it is highly unlikely that Gemini would reproduce an ordinary regional news article verbatim or in substantial part. Google reports that verbatim memorisation rates for the Gemini family are extremely low, and independent studies of similar large models find that regurgitation tends to be limited to highly duplicated or famous material.[41] Combined with the fact that the *Like Company's* article was not in the training data at all, this makes substantial memorisation in this case technically implausible.

Instead, one plausible interpretation of the available technical evidence suggests that the contested summary was generated through RAG and grounding. By explicitly supplying the URL and asking for a summary of 'the online press publication that appeared on balatonkornyeke.hu',[42] the claimant's prompt triggered Gemini's retrieval mechanisms. A retrieval-based system would typically issue one or more search queries to Google Search, retrieve the relevant page, process it into embeddings, and use those retrieved passages within the context window to generate a condensed narrative in Hungarian.[43] The underlying model provided linguistic capability and summarisation skills, but the factual content was drawn in real time from the indexed web page, not from training-time copies.

Technically, then, under this framework, Gemini v1.0 operates in *Like Company* as an LLM equipped with RAG and grounding. It does not contain or reproduce the news article in its weights, nor does it, by virtue of its architecture, make that article available to members of the public beyond those who already have access via the Balaton Környéke website and Google Search. The output at issue is best characterised as a probabilistic summary generated on-the-fly from lawfully accessible online content, rather than as a communication of stored copies from within the model itself.

3 | THE PRESS PUBLISHERS' RIGHT

To answer Question 1, *i.e.*, whether chatbot responses that display text partially matching press publisher's web pages constitute communication to the public, particularly when the reproduced text is substantial enough to fall within the scope of the PPR, it is necessary to start with Article 15 CDSM's internal threshold. The notion of a 'very short extract' matters here because it operates as a built-in limitation on the PPR's scope. Indeed, the referring court explicitly frames the dispute around chatbot answers that 'partially' match press web pages and exceed what can be regarded a 'very short extract' within the meaning of Article 15 CDSM.

In this context, the Hungarian referring court signals that liability for communication to the public arises only once the output falls outside this statutory carve-out. We will return to the general reproduction and communication to the public questions in Sections 4 and 6, respectively. For present purposes, the preliminary task is to clarify the scope of Article 15 CDSM, *i.e.*, including whether it extends to AI-generated content, and its structural relationship with Articles 2 and 3 InfoSoc.

3.1 | The relationship between Art. 15 CDSM and the InfoSoc regime

Crucially, the relationship between this specific regime and the general copyright framework is central to the *Like Company* referral. The referring court explicitly asks whether the application of the general communication to the public right under Article 3(2) InfoSoc 'may exclude' the application of Article 15 CDSM.[44] The referral's 'may exclude' formulation brings an architectural question to the surface: if conduct *i.e.*, in substance, the online use of a press publication could be re-labelled and assessed solely under the general making available right in Article 3(2) InfoSoc (i.e., shorn of the internal carve-outs that Article 15 layers onto that same right), Article 15 CDSM's internal limits, above all 'individual words or very short extracts', risk becoming easy to plead around. This suggests a

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

structural hierarchy: Article 15 was designed as a calibrated interface for press content online, containing specific balances, most notably the exclusion of 'very short extracts'. If claimants could bypass these internal limits simply by invoking the general InfoSoc rights for the same conduct, the legislative compromise would be effectively neutralised. Consequently, the analysis of press content usage by information society service providers must arguably be governed primarily by the specific provisions of Article 15, preventing the general law from swallowing the specific regulation.[45]

One clarification keeps this hierarchy analytically clean without disturbing its practical operation. The subject matter of the PPR is not identical to that of the Article 2 and 3(2) InfoSoc rights on which it is modelled. The PPR protects the press publication as an entrepreneurial and investment object, irrespective of, and independently of the degree to which, the publication or any item within it satisfies the originality threshold; in this it resembles the related rights of performers and of phonogram and film producers, which attach to the performance, recording or fixation as such rather than to authorial expression. Copyright in the individual works incorporated in a press publication (the journalist's article as an 'author's own intellectual creation') and any copyright subsisting in the press publication as a whole by virtue of an original selection or arrangement are distinct objects of protection that Article 15(2) CDSM expressly leaves 'intact'.[46]

3.2 | Does the PPR extend to AI-generated content and to content resembling the one covered by the PPR?

Despite the fact that the PPR is not expressly linked to the existence of copyright in the protected subject matter, the philosophy behind it is not to extend protection where protection does not apply under copyright, or the content has fallen into the public domain. On the contrary, one of the publishers' arguments for its introduction was that this additional related right is needed because it is not always clear from the chain of transfers or the contracts/assignments with creators whether a publisher owns or possesses the rights in a particular piece of content, to block the use or press charges for unauthorised use based on a copyright licence.[47]

If the PPR were interpreted to cover content that is not protected by copyright, it could also be invoked in relation to AI-generated content that inherently lacks copyright protection, because, under EU copyright's human-authorship requirement, output generated autonomously by an AI system will generally fail the 'author's own intellectual creation' standard and therefore fall outside copyright in the first place.[48] In other words, even where AI-generated content falls outside copyright and would normally be free for anyone to use, it could be pulled back under exclusive control through the PPR. That would be conceptually incoherent: it would treat the same content as 'unprotectable' for copyright purposes, yet still protecting it via a related right defeating its original purpose for non-protection.[49] For that reason, AI-generated content that is not protected by copyright should not be brought within the scope of the PPR.

In light of the above, the PPR does not cover 'very short extracts' and should not cover AI-generated content or content that is not or is no longer protected by copyright. In addition (and this is the decisive point) what triggers the related right is not strictly word-for-word copying but the recognisable taking of the publication's protected expression. The distinction matters because related rights are not confined to literally verbatim borrowings. In *Pelham*, the Court held that a protected sound sample reused in modified form still infringes the phonogram producer's right, and escapes it only where the modification renders the sample 'unrecognisable to the ear'; modification in the course of creative reuse is thus no bar to infringement, and recognisability, not verbatim identity, is the operative threshold.[50] The Court has since extended that criterion beyond phonograms to the reproduction of protected expression generally. What separates infringement from lawful reuse is accordingly not whether the borrowing was altered, but what was taken: a related right attaches to the specific protected subject matter (i.e., the publication, performance or recording as fixed), so it reaches a recognisable taking from that subject matter, whether verbatim or recognisably paraphrased, but not the independent re-expression of the unprotected

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

facts or information it conveys. A text that retells the same events in genuinely independent wording no more reproduces the publisher's protected expression than a song re-recorded from scratch reproduces the producer's phonogram. Any other interpretation would unjustifiably block other uses or creations that could express the same facts in a different manner. If the wording produced by chatbots like Gemini is based on the facts of the publication but is technically the outcome of predicted language and hallucination by the chatbot (e.g., as a person in real life would narrate the same things), then no infringement could take place as this narration would be unlikely to infringe the content of the publication.

In conclusion, chatbots like Gemini, which generate language that is the outcome of predicted language and hallucination, without reproducing the protected expression in a recognisable form, cannot give rise to any infringement. Even more so where the extract appearing in the output qualifies as a 'very short extract'.

4 | INPUT AND REPRODUCTION

This section addresses Question 2: *i.e.*, whether the training of an LLM-based chatbot such as Gemini, built on the observation and matching of linguistic patterns, constitutes an act of reproduction within Article 2 InfoSoc particularly regarding press publications, whether such technical acts are pre-empted or governed by the specific regime of Article 15(1) CDSM.

This work suggests that, on the better reading of current EU law, model training, properly characterised, cannot, as such, be treated as reproduction of protected expression. That conclusion preserves the idea/expression dichotomy,[51] and avoids creating a *de facto* control right over 'learning'. Learning statistical regularities is better understood as distinct from the making of copies of a work, and stretching reproduction to cover statistical parameterisation risks being both doctrinally unsound and economically counter-productive in the EU acquis.[52] That position is compatible with an important qualification. As noted in Section 2 above and developed further below, models can, in some rare circumstances, regurgitate passages from training material that they have effectively memorised,[53] often under conditions of duplication, rarity, overfitting, and, in practice, adversarial or extraction-style prompting. However, it is crucial to emphasise that the incidental memorisation of a single work does not imply that the model systematically retains copies of its entire training corpus.[54] Nor is there a reliable, *ex ante* way to infer from the trained weights whether any particular work has been memorised, or to extract that information directly from the trained model. The proof lies solely in the output.[55]

4.1 | A preliminary note: Keeping training and RAG legally distinct

As explained earlier, the factual record indicates that no reproduction occurred at the training stage in this case, because the website at issue was published after the knowledge cut-off date for Gemini v1.0, that is,the first generation of the Gemini model family, announced in December 2023 as Google's new flagship multimodal foundation model. Any allegedly infringing conduct must therefore be ascribed, if at all, to RAG at inference time (on which see the analysis in Section 6). Put bluntly, the factual matrix behind the reference is about what Gemini displayed to users when answering a prompt, *i.e.*, a RAG-based response that allegedly incorporated fragments from a press article, rather than about the offline training phase in which the model's parameters were learned.

Keeping these two stages distinct helps avoid conflating 'learning from' a text' with 'showing' that text to users. RAG and model-training are technologically and legally distinct: the latter raises communication to the public and output-layer reproduction/attribution issues[56]; the former raises the question addressed here, *i.e.*, whether the *learning process itself* fixes protected expression. Given the wording of Question 2, it is appropriate to analyse training *as such*; however, conflating inference-time retrieval with training risks inviting the CJEU to opine on

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

matters not necessary to resolve the national dispute. The Court presumes the referring court's questions are relevant, yet it may reformulate or decline aspects that are hypothetical or bear no relation to the facts essential to judgment.[57] Our discussion of 'training ≠ reproduction' therefore proceeds on a clean technical footing, without relying on inference-time behaviours, and is intended to offer a useful interpretation should the CJEU consider the training issue necessary to the disposition of the case.

4.2 | The EU concept of reproduction and its limits

Article 2 InfoSoc confers on rightholders the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part, of their works. The CJEU has given this provision a technology-neutral but expression-oriented reading. In *Infopaq*, it held that the fixation of 11-word extracts may be a reproduction 'in part' if the extract contains elements which are the author's own intellectual creation; crucially, the finding turned on the existence of copies embodying protected expression, even if brief and transient in the scanning and storage stages of a news-clipping workflow.[58] In later cases the CJEU insisted that protection attaches to expression, not abstract ideas, styles or methods (the 'idea/expression' dichotomy remains a structural limit), and that recognisability of the original protects the boundary between lawful influence and infringement.[59] In *Pelham*, for example, the Court stated that a sound sample may escape infringement where it is unrecognisable to the ear in the new context.[60]

Two immediate consequences follow for the *Like Company* reference. First, analysis as such should not be treated as reproduction. An analytical process, even one that ingests protected works, does not itself constitute reproduction unless it fixes expression. *Infopaq* condemns the making of copies, not the mere act of processing or 'reading' inputs. The decisive question is whether training fixes protected expression. Secondly, EU law has consistently declined to extend rights to functional or abstract features, *i.e.*, program functionality, programming languages and formats are outside copyright.[61] That logic, rooted in the idea/expression dichotomy, matters when assessing what is stored inside a modern model. Models internalise patterns rather than store recognisable and retrievable expressive sequences.

4.3 | What training does (and does not) store: From text to statistics

Modern LLMs ingest large text corpora and use gradient-descent optimisation to adjust parameters (weights) in a neural network so that, given a context, the model minimises prediction error for the next token.[62] The learned artefact consists of a high-dimensional parameter matrix and associated embeddings.[63] It is not a database of strings but a statistical representation of co-occurrence patterns captured via latent-space interpolation.[64] In information-theoretic terms, the process is lossy: it discards vast amounts of surface form, *i.e.*, verbatim character strings, and retains compressed correlations that improve prediction.[65] Retrieval of specific protected passages from the parameter matrix is neither the training objective nor a generally available capability absent special circumstances, *e.g.*, overfitting to rare, duplicated sequences combined with attack-style prompts - a point to which we return later in this Section.[66]

Doctrinally, this matters because parameters are not expression. They are numeric encodings of statistical relationships[67]; they do not, in ordinary operation, fix or embody the protected form of any given work.[68] On the CJEU's own terms, recognisability is pivotal[69]: what the model retains is not recognisable expression, but abstraction.[70] Treating the parameter matrix as a copy of every input would collapse the idea/expression boundary[71] and by parity of reasoning imply that human learning, which also internalises regularities,[72] infringes whenever one studies protected works: *i.e.*, an outcome EU law has never endorsed. Here the model 'memorises' patterns through lossy compression and latent-space interpolation: it discards verbatim detail and retains only

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

statistical patterns, not the original.[73] That looks much closer to human reading and learning, *i.e.*, internalising ideas and styles, than to storing expressive copies.[74] In this regard, the point could be made that an AI model's internalisation of data, like human learning, does not produce expressive copies but instead encodes abstract patterns or statistical relationships. A recent US decision provides a useful external confirmation of the same basic intuition: in *Bartz and others v Anthropic PBC*, the US court characterised LLM training on lawfully acquired books as 'quintessentially transformative' and expressly likened it to a reader learning in order to write, training 'not to […] replicate or supplant [the works]', but to 'create something different'.[75] While that fair use holding is not transferable to EU law, the description of what training does is: it frames training as learning from works rather than warehousing them, which is exactly why equating parameterisation with "reproduction" is doctrinally overbroad.

4.4 | Why comparing training to reproduction is doctrinally overbroad

An automatic equivalence between training and reproduction may sit uneasily with three well-settled strands of CJEU jurisprudence. First, the CJEU has repeatedly anchored protection in authorial expression, not in ideas, methods, styles or functional elements. Read together, cases such as *Infopaq*, *Painer*, *Football Dataco* and *SAS Institute* confirm that functionality, programming languages and data formats fall outside copyright, while protection attaches only to the author's original expressive choices.[76] Against that backdrop, LLM training does not store expressive sequences but abstracts style and statistical regularity from a corpus, precisely the kind of non-expressive regularities that EU law has refused to monopolise.

Secondly, the Court has now made recognisability of creative elements the practical limiting principle for reproduction. As mentioned, in *Pelham*, the phonogram right was not infringed where a sample had been altered so as to be 'unrecognisable to the ear': copyright protects expression, not every trace of influence.[77] The recent CJEU's decision in *Mio/Konektra* generalises that logic to works of applied art and, by implication, to works under Articles 2–4 InfoSoc: infringement requires that the original creative elements of the protected work are reproduced in a recognisable manner in the later subject matter, whereas the 'overall impression' test belongs to design law.[78] Stylistic similarity, *i.e.*, 'this looks like Eames/sounds like Dylan/reads like Woolf', may matter for design, but it is not the copyright test.[79] Nor do fragment size or degree of originality modulate scope: once a work qualifies, any recognisable reproduction of its creative elements is covered.[80] Transposed to LLMs, this points away from treating a trained parameter matrix as a 'copy': the learned weights are not recognisable, nor retrievable, as the text of any particular article; they encode correlations, not the protected form. The infringement risk instead lies at the output layer, where memorisation and regurgitation may produce recognisable passages from training works, *i.e.*, an issue to be addressed with targeted safeguards, not by reclassifying the entire learning process as reproduction.

Thirdly, the *Infopaq* line draws a clear distinction between analysis and fixation. The finding of reproduction there depended on the creation of actual string copies (snippets scanned, stored and printed) however brief their duration.[81] Where a process does not generate or retain recognisable expressions that can be retrieved or perceived by technical means, but yields only statistical traces in a parameter space, it is better characterised as analysis than as reproduction.[82] Reproduction under Article 2 InfoSoc subsists 'in any form', and copies that no person can read directly (an encrypted file, a compressed audio or image file, object code) are reproductions all the same, because the protected expression remains embodied in them and can be retrieved or perceived by technical means.[83] The criterion is therefore the recoverability of recognisable expression: a process produces a reproduction where its output embodies a particular work in a form from which that work can be reconstituted, whether directly or with the aid of a machine. A trained parameter matrix is different in kind. It does not embody the expression of any individual input and, in ordinary operation, cannot be used to reconstitute it[84]; what it retains are statistical correlations distilled across the whole corpus, from which no particular work can be recovered.[85] To extend Article

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2 to such non-expressive traces would be to regulate learning itself, rather than copying, and to collapse the idea/expression dichotomy that structures EU copyright.[86]

A contrary rule would carry far-reaching and technology-distorting consequences.[87] It would equate any statistical model trained on protected materials, whether in language, music or vision, with a copy of all its sources, thereby creating a *de facto* control right over learning.[88] That result is alien to the EU tradition and difficult to administer in practice.

4.5 | Transient copies as technical ephemera in the pipeline, not training itself

Thus, on the best reading of the acquis and the CJEU's expression-centred jurisprudence, the process of training an LLM, understood as the analytical adjustment of parameters to capture statistical regularities in lawfully accessible text, may not constitute an act of reproduction under Article 2 InfoSoc. The trained artefact is a parametric model, not a store of retrievable and recognisable copies of any particular work; it does not fix protected expression and so falls on the 'analysis' side of the idea/expression dichotomy that the Court has consistently maintained.

This conclusion does not ignore engineering realities. Training pipelines may produce fleeting, functional artefacts, *i.e.*, ingestion buffers, caches, tokenisation shards,[89] that, taken strictly, can amount to acts of reproduction.[90] But their existence may not make '*training* itself a 'reproduction' for the purposes of Question 2.[91] These artefacts are ancillary, transient and incidental to the technological process; they are not the learning event.[92] EU law resolves them with the instruments built for such technical ephemera: where inputs are lawfully accessible and no reservation is breached, Article 5(1) InfoSoc (which, it should be noted, is mentioned in the *Like Company* referral but not in the questions posed to the CJEU) covers the temporary, integral copies; where the analysis runs beyond that narrow shelter, Article 4 CDSM (as we will see more in depth in the next section) supplies the dedicated, technology-neutral TDM mechanism, including a machine-readable opt-out. This two-step design provides legal certainty and avoids the high transaction costs of litigating every cache or batch, or worse, chilling research through inaction.[93] At the same time, Article 4's opt-out gives rightholders a practical route to internalise value where licensing is economically sound, without converting learning into a property-rule control, as mentioned already above. The default is permission to mine; a rightholder may switch a given use onto a licensing track by making a machine-readable reservation, enabling remuneration in markets that can bear it. That approach aligns with copyright fundamentals, namely the idea/expression dichotomy: mining/training extracts patterns, trends and correlations rather than expressive form, and so it is analysis, not copying.

Framed this way, the answer to Question 2 is negative in the sense that matters for authorisation: *training as such* cannot be a rights-relevant reproduction. At the same time, legal certainty is preserved because any ancillary, short-lived pipeline copies arising in a compliant workflow are either exempt under Article 5(1) InfoSoc; or, where applicable, authorised under Article 4 CDSM. That preserves the settlement under the latter Directive, respects rightholders' choices through reservation and licensing where they wish to exercise control, and avoids creating a *de facto* monopoly over the act of learning. On that footing, Section 5 of this article will turn to the exceptions-and-limitations framework: if a court nevertheless treats some pipeline elements as reproductions, the correct analysis lies in applying the legislature's chosen tools, *i.e.*, temporary copying (Article 5(1) InfoSoc) and TDM exception (Article 4 CDSM), rather than expanding Article 2 InfoSoc.

4.6 | Memorisation, regurgitation and output-layer responsibility

As an additional clarification to avoid talking past the technology, it is useful to distinguish rare 'regurgitation' phenomena from the nature of training itself.[94] As anticipated in Section 2, computer-science research shows that

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

memorisation, *i.e.*, the verbatim emission of training snippets, does occur, but it is statistically bounded and materially mitigable through dataset deduplication, regularisation, sampling controls and similar engineering choices[95]; leading providers also deploy safety filters and anti-extraction measures that further reduce the likelihood that training results in expressive regurgitation.[96]

Even where string-level memorisation can be induced by adversarial prompting,[97] that behaviour concerns the output layer and service design rather than the training process as 'reproduction'.[98] Put differently: the occasional appearance of long passages is not proof that the learned parameters are copies; it is an operational risk to be addressed by provider diligence (*e.g.*, filters, guardrails, notice-and-action) and, in exceptional cases, targeted injunctions against specific extraction behaviours. Those questions of diligence and attribution are properly analysed under Question 4 (see Section 7 below), not as a reason to collapse the doctrinal boundary between analytical training and rights-relevant reproduction.

Crucially, it is vital to distinguish targeted regurgitation driven by adversarial prompting from continuous, systematic market substitution for ordinary readers. A recent study illustrates this precise distinction.[99] While the researchers demonstrated that text extraction from a book is technically possible, their results were achieved only through a highly specific, non-standard 'sliding window' attack that involved thousands of successive prompts designed to bypass security filters,[100] a process far more resource-intensive than simply purchasing the book.[101] Furthermore, the attack's success depended heavily on the specific target, *Harry Potter and the Sorcerer's Stone*, a work repeated thousands of times across the public web, making it a statistical outlier particularly susceptible to memorisation.[102] Consequently, while media coverage has understandably amplified the headline risk,[103] the paper itself treats the result primarily as a safety/security vulnerability addressable by technical means,[104] not as a freestanding copyright conclusion that LLMs are 'free libraries'. Rather, these findings confirm that while determined actors using sophisticated tools can elicit memorised sequences from highly duplicated data, standard guardrails such as data deduplication and output filtering remain effective in preventing the model from serving as a market substitute in ordinary operation. Two caveats temper how far this example travels. First, it concerns verbatim regurgitation of a literary work (a memorisation phenomenon) and not the distinct question of informational substitution, where a model conveys the gist of a source in fresh language. Secondly, books and press publications are not similarly situated in this respect: a novel is consumed for its expression, so its market is threatened mainly by reproduction of that expression, whereas a news article is consulted largely for the information it carries, so a chatbot that accurately summarises the facts may divert readership from the source even though it reproduces no protected expression at all. Press publications are, on that account, more exposed than books to substitution by generative summarisation.[105] That exposure is real and is taken seriously below; but it is, by its nature, a harm to the informational and readership economy of journalism rather than an appropriation of the publication's protected expression, and EU copyright has deliberately placed facts and information outside the scope of protection. The implication, developed in the Conclusion of this article, is not that the reproduction and making-available rights should be stretched to capture facts-based substitution (which Recital 57 and the same-scope principle confirmed in *Meta* forbid),[106] but that, if this harm is to be addressed, the appropriate instrument is a remuneration mechanism keyed to such use, not an expansion of the exclusive rights.[107]

Recent national litigation, however, illustrates how easily the debate can slide from 'rare regurgitation' to a general theory that treats learning itself as copying. The recent decision by the Munich Regional Court in *GEMA* is widely being presented as a landmark for 'training-as-reproduction', but much of the public takeaway rests on conflating proof of memorisation (via prompted lyric outputs) with a general claim that works are stored 'inside' a model as such.[108] That framing is technically contestable and doctrinally risky, because it substitutes an output-driven inference (recognisable passages can be elicited) for the Article 2 InfoSoc question whether the trained artefact *as such* fixes protected expression.[109]

In sum, legally and economically, the implications of such memorisation (regurgitated via adversarial prompting) are narrow. Verbatim emissions are rare, prompt-dependent and concentrate on duplicated training items; they fall

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

sharply with deduplication and alignment. Occasional output that matches training text does not prove a latent internal copy of each work; it shows memorisation at the margin, which can be mitigated. Treating these tail-events as proof that the entire learning process equals reproduction would turn Article 2 InfoSoc into a property rule over learning capacity. The fact that a model *can* be forced, through sustained adversarial interaction, to emit recognisable passages is evidence that output-layer guardrails matter and that liability analysis should focus on the concrete instance of recognisable reproduction in the output, rather than treating training or ordinary conversational summarisation as market-substituting 'publication' by default.

4.7 | A cautionary note: Harmonisation with the PPR's scope

The second referred question explicitly links Article 15(1) CDSM and Article 2 InfoSoc, asking whether training constitutes reproduction under either provision. It is crucial to insist that the definition of 'reproduction' remain consistent across both.[110] While related rights typically protect the specific 'fixation' of a publication rather than its 'originality', interpreting 'reproduction' under Article 15 to cover non-expressive technical acts (such as statistical parameterisation) would lead to a paradoxical result: the related right would become functionally broader than the copyright it is meant to complement.[111]

Specifically, if Article 15 CDSM were read to capture any technical processing of the fixation 'as such', even where no original expression is preserved, it would grant publishers a monopoly over the informational content of news, a protection explicitly denied to authors under copyright.[112] This would contravene the legislative intent that the rights granted to press publishers should have the same 'scope' as the reproduction right provided for in InfoSoc.[113] Consequently, to avoid inverting the hierarchy between the author's primary right and the publisher's related right, 'reproduction' under Article 15 must be interpreted as requiring the same focus on recognisable expression as Article 2 InfoSoc. Therefore, Article 15 must be interpreted restrictively to exclude acts of machine learning that do not preserve the protectable substance of the publication.

5 | TDM EXCEPTION

The related and crucially important Question 3 concerns the applicability of the TDM exception to the *Like Company* factual scenario. This section examines not only the scope of the CDSM's TDM exception in that dispute, but also the explicit legislative intent of EU institutions, particularly as codified in the EU AI Act,[114] to subject AI training to the TDM exception framework.

As mentioned, the CJEU is asked whether, in case an LLM-based chatbot like Gemini constitutes an instance of reproduction under Article 15(1) CDSM and Article 2 InfoSoc, this reproduction of lawfully accessible works falls within the TDM exception under Article 4 CDSM. As we have exposed before,[115] we consider that in scenarios such as in *Like Company*, reproductions in the sense of Article 2 InfoSoc may not materialise; however, should the transient copies and any related step in the training process qualify as 'reproductions' by the Court, we are of the opinion that the TDM exception applies in any case in this context, legitimising such use in the absence of a right reservation by *Like Company*.

Before analysing the TDM exception, we preliminarily note that a hierarchy of defenses applies here: the RAG of the article in question likely constitutes a transient, incidental act under Article 5(1) InfoSoc. The issue is of interest, even though as mentioned such exception falls outside the express scope of the questions referred to the CJEU in *Like Company*, which suggests that the referring court may have treated this defense as settled or secondary to the copyright infringement analysis. That the RAG in question is such a copy under Article 5(1) InfoSoc is confirmed by the technical analysis within Section 2 of this article, which shows that the grounding mechanism through which Gemini has retrieved the Balaton article did operate in real time, creating temporary buffers and

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

cache copies that persisted only for the duration of the inference process. These ephemeral copies were integral to the technical functioning of the search and summarisation process itself, much like the transient reproductions occurring in RAM or cache memory during any ordinary web retrieval.

Said that, how the training of an LLM like Gemini can be evaluated from a copyright perspective, in particular when the algorithm is 'fed' with copyright protected works,[116] is here the crucial question. Different from other jurisdictions such as the US, where the legality of machine learning is discussed in the framework of the fair use exception[117] or Japan, where GenAI training is legalised under a specific 'non enjoyment' exception,[118] the EU legislator took the decision in the AI Act to subject the training of GenAI models to the CDSM's TDM exception. Indeed, Article 53, 1-c) AI Act imposes obligations on providers of general-purpose AI models like Gemini to 'put in place a policy to comply with Union copyright law, and in particular *to identify and comply with*, including through state-of-the-art technologies, *a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790* [CDSM]' (emphasis added).[119] This provision needs to be read together with Recital 105 CDSM, which clarifies the EU legislator's will by making explicit reference to GenAI models and their training on copyright protected works.

This has always been the position of the European Commission also prior to the AI Act: when asked to clarify its position, Commissioner Thierry Breton on behalf of the Commission considered that the existing EU legislative framework fit for the purpose of GenAI as the 'creation of art works by AI does not deserve a specific legislative intervention'; and this was the case because the TDM exceptions give content creators opt-out chances, 'providing balance between the protection of rightholders including artists and the facilitation of TDM, including by AI developers'.[120] It has been equally the position of a majority of scholars that the training of GenAI models like Gemini is subject to the CDSM's TDM rules.[121] It can nevertheless be regretted that this important legislative decision in 2024 did not follow an informed debate in the European Parliament. The relevant provision of the AI Act was introduced last minute in the very last round of discussions over the TDM exceptions. Those exceptions, even if fairly recent, also predate the GenAI boom.[122] Therefore, as determined by the EU legislator, the technical steps and reproductions necessary for training GenAI models fall within the scope of the CDSM TDM exception. This is because GenAI systems learn from datasets through TDM.[123]

Crucially, the framework described above seems to encompass the specific technical process at issue in the *Like Company* referral: *i.e.*, the 'grounding' or RAG process. When the model retrieves the Balaton article to generate a summary, it is performing an 'automated analytical technique' under Article 2(2) CDSM to extract semantic information, thereby triggering the exception. This real-time retrieval and analysis of lawfully accessible content, *i.e.*, the inference-stage processing, constitutes TDM regardless of whether the training phase itself involves reproductions, ensuring that derivative works and summaries generated from publicly available sources remain lawfully exempt. More specifically, Article 2(2) CDSM defines TDM as 'any automated analytical technique applied in order to analyse text and data in digital form in order to generate information'. The grounding process precisely matches this definition: it is automated, analytical, applied to digital text, and generates new information, *i.e.*, the summary. This functional alignment grounds the TDM exception not in theoretical disputes about parameterisation, but in the concrete observable act of content analysis.

As is known, Article 3 CDSM introduces an exception for TDM for scientific research, which solely benefits research organisations and cultural heritage institutions. By contrast, Article 4, which was added later in the legislative process that led to the CDSM, is not limited to specific types of institutions and therefore applies more broadly.[124] The latter is the provision effectively relied upon by the AI Act, as GenAI systems like Gemini are usually developed and operated by private commercial companies that do not fall within the scope of Article 3. According to Article 4(1) CDSM, 'Member States shall provide for an exception or limitation [...] for reproductions and extractions of *lawfully accessible works* and other subject matter for the purposes of text and data mining' (emphasis added). Interestingly, under Article 4(2) CDSM, 'reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining', which is relevant to solve any possible question of storage or memorisation of protected works by GenAI models in the learning process.[125]

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Recent case law has started to clarify how Article 4 CDSM operates in practice for online services that crawl and index protected content and for AI-training data sets. Back in 2024, the Hungarian Municipal Court of Appeals held that a leading search engine's web-crawling, indexing and display of short previews of an online press publisher's articles fell within the national implementation of the general-purpose TDM exception, given that the service had lawful access and the publisher had not validly reserved its rights in machine-readable form (robots.txt).[126] Similarly, the Amsterdam District Court in *DPG Media B.V., Mediahuis Netherlands B.V., and NRC Media B.V. v. Knowledge Exchange B.V.* accepted that a commercial news-alert service scraping publishers' RSS feeds and displaying very short snippets with links could rely on Article 4 CDSM, in the absence of an effective machine-readable TDM opt-out on the publishers' websites and feeds.[127] Another strand of reasoning appears in the German *Kneschke v LAION* litigation, where the court addressed AI-training uses and the German implementation of Articles 3 and 4 CDSM. While recognising that TDM exceptions can in principle cover large-scale dataset construction, the court engaged directly with whether natural-language reservations in website terms (rather than robots.txt or similar protocols) can satisfy the requirement of a 'machine-readable' opt-out. The higher-instance guidance emphasised that the opt-out must be perceivable and interpretable by a machine and, in the 2025 appeal, ultimately rejected the specific natural-language opt-out at issue as not machine-readable.[128]

Taken together, these emerging national decisions on Article 4 CDSM delineate the practical contours of the EU's TDM framework against which the *Like Company* reference must now be read. In this case, the Balaton article was freely available on the open web without paywalls, authentication walls, or technical restrictions, *i.e.*, a critical factual foundation for establishing lawful access under Article 4(1) CDSM. The publisher made the content publicly indexable and retrievable by standard web protocols and search mechanisms, thereby consenting to its accessibility in the information ecosystem. This contrasts with scenarios involving hacked databases, circumvented paywalls, or deliberately obscured repositories. The absence of technical or legal barriers to access means that the retrieval of the article for purposes of analysis and summary generation, whether during training or inference, occurs in a context where the content's availability is lawfully assured. This satisfaction of the Article 4(1) prerequisite is fundamental to triggering the TDM exception's protections.

A radically different approach has been taken recently by some scholars, arguing that the TDM exception does not cover training activities[129] or at least not all copyright-relevant uses of GenAI training activities[130]; or that from a policy perspective, it should not cover them because of structural imbalances this would introduce.[131] In this sense, the decision of the Munich court in *Gema* caught a lot of attention, as it decided that only parts of the training activities are covered by the TDM exception, while the memorisation/storage of song texts via training involves reproductions that are not covered and therefore infringing.[132] This ruling, however, is very debatable on the TDM side (input), as we have seen that the AI Act clearly intends to submit GenAI training to the TDM exception under Article 4 CDSM and the opt-out mechanism. Splitting the training process into different technical parts seems to be a *contra legem* interpretation which will be detrimental to a unified approach in the EU member states which the regulation of the AI Act meant to achieve.[133] For this reason, we believe the Munich court should have referred the matter to the CJEU for clarification as it clearly involves the interpretation of EU law.

The outcome reached by the Munich Court in *Gema*, which as mentioned divided the training process into separate technical phases and denied TDM protection to the storage/memorisation component, underscores the critical importance of the *Like Company* referral to the CJEU. Unlike the German court's *contra legem* approach, which seems to conflict with the AI Act's explicit subjection of GenAI training to Article 4 CDSM, the Hungarian referring court has framed its questions to encompass both the training process and the inference-stage (RAG/grounding) operations. This broader scope affords the CJEU an opportunity to establish authoritative, unified guidance that TDM protection covers the entire training pipeline, including memorisation and parameter storage, as an integrated analytical technique under Article 2(2) CDSM. In other words, by anchoring its analysis in the concrete facts of *Like Company*, where the alleged infringement pivots on real-time content retrieval and lawful access, the Court can establish binding EU-wide clarity: *i.e.*, that Article 4 CDSM applies holistically to GenAI development

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

absent valid opt-outs, and that no artificial subdivision of technical stages can circumvent the exception's protections. Such a ruling would pre-empt fragmentation and preserve the digital single market's coherence.

In sum, also considering what the EU legislator meant by passing the AI Act, the reproductions aimed at training GenAI chatbots like Gemini (assuming reproductions do materialise) are subject to the TDM exception under Article 4 CDSM. As we have underlined, this solution is surely not ideal and might call for a revision de *lege ferenda* in order to implement a model more favourable to GenAI innovation in the EU, while safeguarding the core European values such as a fair remuneration for creators. But according to the current legislative framework, there is little doubt that the use for training should be exempted unless the rightholders have opted out. The *Like Company* referral presents the CJEU with an ideal vehicle to crystallise this legislative intent into binding judicial doctrine, thereby foreclosing national-court deviation and anchoring the TDM exception as the governing framework for lawful GenAI development across the EU.

Finally, while the *Like Company* referral does not explicitly turn on whether the publisher validly exercised an opt-out, the legal uncertainties surrounding Article 4(3) CDSM's opt-out mechanism remain central to the CJEU's broader interpretative task. The opt-out provision has been extensively critiqued for generating legal insecurity[134]: when precisely is content 'lawfully available' online for TDM purposes? What constitutes 'an appropriate reservation by machine-readable means', particularly in the absence of harmonised technical standards? May derivative rightsholders (such as press publishers) exercise the opt-out independently of authors, or must consent align vertically within the rights chain? The absence of both a uniform technical standard and a centralised EU-level registry where opt-outs can be consulted compounds these uncertainties. The European Commission is actively addressing these gaps: the November 2025 launch of the Copyright Knowledge Centre at the EUIPO, tasked to 'explore solutions for copyright to keep powering creativity and innovation in Europe', signals institutional recognition of the problem.[135] Given the Centre's existing management of the Orphan Works Database and the Out-of-Commerce Works Portal, and the Commission's commissioning in January 2025 of a feasibility study on a centralised TDM opt-out registry, a unified, interoperable opt-out infrastructure may eventually emerge as a practical complement to the CJEU's doctrinal clarification in *Like Company*.[136]

6 | OUTPUT AND COMMUNICATION TO THE PUBLIC

Question 1 concerns whether the display, by a GenAI chatbot such as Gemini, of text that partially reproduces press publisher web pages constitutes a communication to the public within the meaning of Article 15(1) CDSM and Article 3(2) InfoSoc. In essence, the Hungarian court asks whether AI-generated summaries or responses that incorporate verbatim segments or close paraphrases of online press articles amount to a new act of making those publications available to an indeterminate public, thereby falling within the PPR's scope. This section examines that question in light of the CJEU's existing communication to the public case law, the specific architecture of RAG systems such as Gemini, and the internal balancing mechanism built into Article 15(1) CDSM through the exclusion of 'individual words or very short extracts'.

6.1 | Summaries, paraphrase, and when an output becomes 'reproduction in part'

Having established that the *training* phase, *i.e.*, the offline internalisation of statistical patterns into model parameters, cannot be characterised as reproduction under Article 2 InfoSoc, it is crucial to distinguish this from the separate issue of *inference-time* generation. The legal questions raised by chatbot summaries, such as that in *Like Company*, concern an output-layer (often RAG-mediated) phenomenon: summaries and paraphrases arise at inference time, when the system responds to a user prompt and may retrieve and display (or restate) material from external sources. While this section addresses whether such displays constitute a 'communication to the public', we

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

must first resolve the antecedent question of reproduction at the output stage: does a summary or paraphrase generated by GenAI at the moment of inference constitute a 'reproduction in part' of the original article?

The answer generally leans towards the negative. A summary expressed in fresh language should not be treated as a *reproduction* of the press article. As mentioned, EU copyright (and the PPR under Article 15 CDSM) protect *expression*, *i.e.*, the author's original choices in wording, structure, and presentation, not the underlying facts, themes, or 'information' conveyed by a report. A chatbot output that restates an article's informational content in different words will therefore ordinarily fall on the 'ideas/facts' side of the boundary, even if it is accurate and even if it is prompted by a user who has access to the original publication.[137] The difficult cases arise where the output crosses from *paraphrase* into *recognisable copying*, *e.g.* by reproducing distinctive turns of phrase, sentences, or other concrete expressive selections that embody the author's own intellectual creation.[138]

That 'hinge' is increasingly well-articulated in the CJEU's jurisprudence. As mentioned above, in *Infopaq* the Court made clear that reproduction 'in part' turns on whether the material taken contains elements that are the expression of the author's intellectual creation[139]; and most recently, in *Mio/konektra*, the CJEU sharpened the infringement inquiry for works of applied art: infringement requires that the original *creative elements* of the protected work are reproduced 'in a recognisable manner' in the later subject matter, and the design law criterion of 'overall impression' must not be imported into copyright analysis.[140] Although those cases arise outside journalism, the limiting principle travels well: what matters is not that an output 'resembles' the article in topic or tone, but whether protected expressive choices survive *recognisably* in what is shown. As also mentioned above, the same logic underpins *Pelham*: where a fragment is altered so that it is unrecognisable (there, 'to the ear'), it falls outside reproduction of the protected subject matter.[141] In the summary setting, this helps explain why a probabilistic paraphrase will ordinarily not count as reproduction: it yields a new, distinct work and, in ordinary cases, does not undermine the publisher's ability to obtain a satisfactory return on its investment in the publication's protected expression.[142] That qualification matters: a summary may still divert readers, and so revenue, from the source by conveying the unprotected facts the article reports; but such informational substitution is not a harm the reproduction right addresses, precisely because facts and information lie outside protection. Whether that residual, informational harm warrants a remuneration response is a distinct question, taken up in the Conclusion of this article.[143] The doctrinal point is modest but important: transformation is not a free-standing defence, yet recognisability functions as a practical boundary-marker between protected expression and non-protected abstraction.[144]

Several features sometimes emphasised in the present factual matrix may be relevant. The referring court in *Like Company* describes scenarios where different responses may display different parts of a press publication (including summaries), and notes the risk of inaccurate or 'hallucinated' content.[145] It is also said that (i) the chatbot output points to, or is assessable only by consulting, the original site; (ii) the full article remains available at the source; and that (iii) the output is generated only after a user prompt that already signals awareness of the publication.[146] Those considerations may bear on *market substitution* and on the practical function of the output (*e.g.*, whether it operates as a navigational aid rather than a substitute): also, they help characterise the output as a probabilistic *answer* that often pushes users back to sources: hence why disputes of this kind typically turn on Article 15 CDSM's calibrated interface for press content online (and, where applicable, quotation), rather than on treating summarisation as if it were the showing of the work itself.[147] As will be detailed shortly, Article 15 CDSM applies to extracts that are reproduced verbatim and do not qualify as 'very short'.

This distinction between abstract paraphrase and recognisable reproduction is where the 'may exclude,' question from the referral becomes decisive. It is analytically cleaner to separate the inquiry into two stages. The first is the threshold question of reproduction under general copyright principles: has the output reproduced protected expression in a recognisable way? If the answer is negative, as with most generative summaries that maintain sufficient distance from the source text, no rights are infringed. However, if the answer is positive, *i.e.*, where the system outputs verbatim or near-verbatim blocks of text, or strings together distinctive phrases so that

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

protected expression is reproduced recognisably, the second step must be to apply the specific filter of Article 15 CDSM.

To treat such displays solely under Article 3(2) InfoSoc would ignore the legislature's intent to create a dedicated regime for press aggregators. Therefore, if the displayed content qualifies as a 'very short extract' under Article 15(1), this specific limitation should effectively displace or 'exclude' liability under the general communication right, ensuring the fair balance between rights holders and technology providers intended by the CDSM is preserved. The precise boundary of this liability, specifically regarding the 'very short extracts' exception, is the subject of the next subsection, while the broader question of whether such displays constitute an act of dissemination is addressed later.

6.2 | Exception for short extracts and fair balance

The specific balancing mechanism built into Article 15(1) CDSM, *i.e.*, the exclusion of 'individual words or very short extracts', sets the general thresholds for copyright protection within the concept of 'reproduction in part' under Article 2 InfoSoc. More specifically, the referring court seeks clarification on whether AI-generated summaries containing multiple verbatim phrases constitute 'very short extracts'; and how to assess extract length when content appears across multiple chatbot interactions. As mentioned earlier, in the specific case before the Hungarian court, Gemini's summary incorporated elements from *Like Company's* article about Kozsó's dolphin project, but there is no evidence of direct quotations; there were also references to related articles on the singer's criminal history and health issues.[148] The national court noted that this might exceed typical snippet lengths allowed under national implementations of Article 15(1) CDSM.

It goes without saying that the PPR does not extend to mere facts or data reported in press publications since this is a basic premise of copyright law. Also, it does not apply to individual words or very short extracts of a press publication, as this has also been set out in the CJEU's case law with regard to copyright works.[149] In particular, in *Infopaq*[150] it was famously held that words as such are not protected since they do not constitute elements covered by protection, nor are they the intellectual creation of the author who employs them. It is only through the choice, sequence, and combination of these words that the author can express their creativity in an original manner and achieve a result—*i.e.*, an intellectual creation. Isolated sentences, or even certain parts of sentences in a text, are protected if they are suitable for conveying to the reader the originality of a publication, such as a newspaper article, by communicating to that reader an element which is, in itself, the expression of the intellectual creation of the author of that article. It is only such sentences or parts of sentences that may therefore be protected.[151] It would be inconsistent for the free use of those extracts to be allowed under copyright law and barred via the PPR. That would render the related right broader than copyright and would unjustifiably block the permitted use of unprotected works.[152] This train of thought can also be derived from the fact that there is an express provision in the CDSM clarifying that journalists are to receive an appropriate share of the revenues for the use of their works: this is Article 15(5) CDSM. It would make no sense for journalists to be remunerated for uses for which they are not entitled, *i.e.*, if the extract used was so minimal that it could not actually qualify for copyright protection.[153] The *Like Company* referral implicitly raises important policy questions about how to reconcile the PPR with the EU's objectives for GenAI development within the digital single market. The CJEU will need to consider whether existing exceptions sufficiently accommodate machine learning workflows, and how to apply the CDSM's 'fair balance' principle between rights holders and technology providers. Clarifying the threshold for 'short extracts' will directly affect the permissible level of detail in output tokens generated by transformer-based systems.

Based on our analysis, Gemini-style summaries are structurally aligned with Article 15(1) CDSM's 'very short extracts' threshold because, even when they incorporate brief quotations, they primarily re-express factual content in independent language and compressed form rather than re-presenting the press publication's protected expression. They function as dynamic, query-specific snippets, comparable in length and purpose to search result

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

previews that Recital 58 CDSM and national implementations treat as permissible, so long as they do not substitute for reading the original article.[154] By providing high-level overviews, such extracts preserve incentives for users to click through to the publisher's site while facilitating information discovery and AI-assisted research, thus maintaining the CDSM's intended fair balance between press publishers' monetisation interests and broad public access to news.

It might be objected that this reading is insufficiently purposive, *i.e.*, that 'reproduction' and 'communication to the public' are autonomous concepts to be construed in light of the aims of the instruments in which they appear, and that Article 15 was enacted precisely to counter the erosion of publishers' revenues by services that mediate access to news. The objection has force, but it does not carry as far as it first appears. Purposive interpretation operates within, not against, the structural choices the legislature has made; and the EU legislature made two such choices that bound the aim here. It tied the related right to the same scope as the reproduction and making available rights in InfoSoc (a co-extensiveness the Grand Chamber has since confirmed)[155] and it expressly excluded 'mere facts', 'individual words' and 'very short extracts' from the right's reach. The aim of recouping investment is therefore an aim to protect investment in expression, pursued through rights no broader than copyright and subject to express carve-outs; it is not a freestanding mandate to capture any service that competes with the press for readers' attention. A reading that brought facts-based summarisation within the right in order to serve the provision's economic purpose would not interpret those limits but override them, converting the PPR into a monopoly over news as information, *i.e.*, a result Recital 57 forecloses, and that would sit uneasily with the access and innovation interests the same directive is meant to accommodate. The purposive case, properly bounded, thus confirms rather than displaces the analysis offered here; the gap it leaves is one for the legislature to close by other means.

6.3 | Act of communication in the context of GenAI systems and the notion of 'public'

The CJEU has consistently understood 'communication' to cover any transmission or making available that enables the public to access a work from a place and at a time individually chosen by them, provided there is a deliberate intervention by the person who organises that access. The focus falls on an intentional act which makes protected subject matter perceptible to others and which goes beyond mere provision of physical facilities. This is equally true for Article 3(2) InfoSoc and the PPR in Article 15 CDSM, which builds on that provision.[156]

Two aspects are central: Whether protected content is actually communicated, *i.e.*, whether the service makes the original press publication, or a substantial part of it, perceptible to users; and whether the communication is directed at a 'public', an indeterminate but fairly large number of potential recipients, rather than an isolated individual or closed, predetermined circle. In the *Like Company* scenario, the chatbot's core contribution is to generate a probabilistic summary in response to a user's prompt. Technically and legally, that summary is a new text. It draws on information that may be retrieved at inference time (via RAG or grounding mechanisms) but is not a reproduction of the original article as such. As explained above in this article, the model does not store press publications in a retrievable and recognisable format; it operates on learned statistical regularities and, in the RAG pipeline, on transiently retrieved passages that are processed into the context window. In *GEMA*,[157] by contrast, the Munich court proceeded on the factual finding that ChatGPT had generated output containing substantial segments of specific song lyrics; this made it significantly easier for the court to characterise the output as a direct communication of the phonogram or musical work, as the user was confronted with text that effectively was the original song, rather than a mere synopsis of its content. That factual matrix diverges sharply from the summarisation of short news pieces that is at issue in *Like Company*.

Where the chatbot produces a condensed narrative in its own words, we are dealing with an act of communication of a newly generated work, not a communication of the original press publication. The mere fact that the summary concerns the same events as the article does not transform that summary into a communication of the

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

original protected expression. Facts and data remain unprotected.[158] What matters is whether the chatbot reproduces the author's original selection and arrangement of words in a manner that remains recognisable.

Even if one were to identify isolated verbatim passages in the output, these would typically be limited to single phrases or sentences, embedded in an otherwise reformulated narrative. Based on the CJEU's own logic in *Infopaq* and subsequent case law, only those parts of the output that convey the author's intellectual creation, through a recognisable choice and arrangement of words, could be treated as reproducing the work in part. In the typical chatbot summarisation context, that threshold is rarely met and in any event it affects only fragments, not the summary as a whole. In this context, *GEMA* serves as a cautionary example without displacing the general rule: it confirms that where a GenAI system does cross the line into substantial verbatim reproduction of protected text, courts will have little difficulty applying the communication and reproduction rights. But it also implicitly acknowledges that such cases are at the extreme end of the spectrum and do not define the norm for AI-assisted summarisation.

As regards the notion of the 'public', chatbot interactions such as those triggered by Gemini are ordinarily individual and on demand: a single user submits a prompt and receives a response generated specifically for that session. However, under the jurisprudence of the CJEU, the assessment of 'communication to the public' is not confined to the one-to-one technical structure of the transmission. The Court has consistently defined the 'public' as an indeterminate number of potential recipients and implies a fairly large number of persons, assessed cumulatively rather than individually.[159] It has therefore accepted that successive, individual acts of making available may be considered together in determining whether there is a communication to the public. On that approach, the fact that access occurs through discrete, user-initiated requests does not preclude a finding that the same protected subject matter is communicated to an indeterminate public where the service enables access by a potentially unlimited number of users.

Yet, the scenario highlighted in this article, *i.e.*, a GenAI platform which provides an interactive tool rather than broadcasting pre-selected content to a mass audience, is factually different. In other words, such scenario is conceptually distinct from streaming, broadcasting or the permanent provision of content to an undifferentiated public, exactly because the GenAI platform generates responses dynamically and transiently for each user, without making predefined works available or enabling simultaneous access to identical content by multiple users, thereby maintaining the inherently individual nature of each interaction. Accordingly, at the level of primary qualification, there are good reasons to conclude that the GenAI summarisation in question does not amount to an act of communication to the public of the underlying press publication. Rather, it is the communication of a newly generated text, which only incidentally overlaps with the original expression.

6.4 | The notion of 'new public'

Even if one assumed *arguendo* that some chatbot interactions could amount to an 'act of communication' within the meaning of Article 3 InfoSoc, the separate requirement that such communication be directed at a 'new public' would not be satisfied in the *Like Company* scenario. The 'new public' criterion, developed in *Svensson* and refined in *GS Media* and *VG Bild-Kunst*, requires that the communication reach a public that was not taken into account by the rightholder when the initial communication to the public was authorised.[160] Where a work is freely accessible on the internet without technical or contractual restrictions (e.g., paywalls or login barriers), the rightholder is presumed to have taken into account all internet users as the relevant public. Hyperlinks, including embedded links, that merely facilitate the finding of such content do not generally address a new public; they act as a "pointer" to an existing communication.[161]

In the Hungarian reference, the *Like Company* article about Kozsó was made available free of charge on a public website, without technological restrictions. The relevant public thus consists of all internet users who might access the site. Gemini's RAG-based operation, when prompted with the URL, essentially performs a targeted search and

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

summarisation of content already accessible to that same public. It does not circumvent technological protection measures, nor does it unlock content that was previously limited to a restricted audience. Moreover, the behaviour of the user in this case is telling: by explicitly including the URL of the article in the prompt ('the online press publication that appeared on balatonkornyeke.hu'), the user demonstrates awareness of the original publication's existence and accessibility. The chatbot thus presupposes, instead of creating, the user's connection to the underlying content. This could be seen as the opposite of reaching a new public; the service facilitates interaction with content that the user has *already* identified.

To treat such an interaction as addressing a new public would be to conflate any mediation (*e.g.*, search, indexing, recommendation and summarisation) with communication to a new public whenever the intermediary provides added value. That would extend the notion of 'new public' far beyond its justificatory basis and would risk capturing commonplace acts of online navigation and information retrieval, which the CJEU has so far carefully excluded from the scope of Article 3 InfoSoc, especially in cases where no circumvention of access restrictions is involved.[162] In *Getty Images v Stability AI* ('*Getty*'), the UK High Court did not accept that the mere provision of a general-purpose generative model constituted infringement per se, instead directing attention to whether infringing reproductions could be shown to subsist within the model or its deployment in the UK.[163] The focus remained on whether, in concrete instances, the output reached users as recognisable reproductions of the claimant's works. By contrast, the Munich Court's decision in *GEMA*[164] illustrates a situation where the 'public' element was easily satisfied because the allegedly infringing outputs, in particular song lyrics, were made available to ordinary users in a form that was indistinguishable from the original; yet even there, the court's reasoning turned on the specific expressive character and commercial importance of lyrics, not on any general hypothesis of AI-mediated access as such. That logic militates against any blanket characterisation of chatbot summarisation as communication to a 'new public'.

6.5 | Hyperlinking case law revisited for GenAI outputs

The hyperlinking jurisprudence of the CJEU offers an indispensable reference point even if chatbots do more than simply reproduce URLs. In *Svensson*, *BestWater*, *GS Media* and *VG Bild-Kunst*, the Court sought to balance authors' rights with the architecture of the web by drawing a line between (i) links that enable users to access lawfully available content without reaching a new public and (ii) links that circumvent restrictions or target clearly unlawful sources.[165]

Chatbots that like Gemini operate via RAG can be understood as combining two functions: Information retrieval, which is functionally equivalent to a search engine querying an index and returning snippets and links; and generative summarisation, which synthesises retrieved information into a new text. From the perspective of communication to the public, the first component aligns closely with the hyperlink cases. Where a chatbot surfaces a link (or other in-text reference) to the original press publication, the situation is indistinguishable from that of a search engine results page displaying the article title, a short snippet and a clickable URL. In both instances, the user must still visit the original site to access the full content; the public reached remains the same. The second component, *i.e.*, summarisation, adds a layer of transformation. Yet, this added functionality does not, by itself, create a new public or a new communication of the underlying work. On the contrary, the summary is a distinct work generated by the system, typically of a transformative nature, which does not fully or substantially reproduce the original expression. As explained earlier in Section 3 of this article, the PPR cannot be interpreted in a manner that would render it broader in substance than the underlying copyright, particularly where the extracts used fall beneath the threshold of originality, are 'very short', or amount to unprotected facts and ideas.

Seen in this light, the hyperlink cases remain conceptually fit for purpose. The mere fact that the 'pointer' is embedded in a richer, AI-generated narrative does not change the basic structure: the chatbot, like a search engine, facilitates discovery and understanding of content already lawfully accessible on the web, and does so without

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

expanding the circle of persons who can access the original article. The CJEU can therefore build on its existing case law, adapting its language to account for generative features but preserving the core insights about lawful access and lack of a new public. The approach taken by the UK High Court in *Getty* supports the analogy: rather than treating GenAI services as infringing *per se*, the court focused on whether the specific acts of reproduction and use could be established on the evidence in the relevant jurisdiction. And the *GEMA* judgment by the Munich Court shows what happens when the factual pattern changes: once the GenAI output ceases to be a pointer or paraphrase and instead delivers large quantities of recognisable expressive material (song lyrics) on demand, the mediation begins to look much closer to a classic act of making available. Crucially, however, nothing in *GEMA* suggests that summarising non-creative factual reporting, without extensive verbatim reproduction, should be treated in the same way.

6.6 | Parallels between chatbots and search engine results pages

Like Company presents, in fact, a paradigm instance of how chatbots and search engines converge. By explicitly asking Gemini to 'provide a summary in Hungarian of the online press publication that appeared on balatonkornyeke.hu', the user is functionally performing a highly specific search query. In a retrieval-based scenario, the system's RAG or grounding module would then retrieve the article (or relevant cached/indexed versions), analyse it, and provide a condensed account, frequently accompanied by an in-text reference or link to the source. From the user's perspective, this behaviour is close to what a modern search engine already does when answering "rich" queries: it may show a snippet, a 'knowledge panel', or a short extract together with the URL. What is new is not the basic function, *i.e.*, mediating access to web content, but the sophistication of the natural-language interface and the degree of paraphrasing.

From the standpoint of EU copyright law, chatbots employing RAG therefore operate much like enhanced search engines. Search engines routinely display verbatim or paraphrased snippets as part of their indexing functions, and the CJEU has never held that such displays amount to acts of communication to the public, provided no substantial protected expression is reproduced. If one were to classify such summarisation as an independent communication to the public of the press publication, it would be difficult to avoid the conclusion that search engine snippets and many forms of enriched search results likewise constitute communication to the public. That reading would stand in tension with the careful calibration reflected in Article 15 CDSM, which as mentioned addresses search engines and aggregators by granting press publishers a related right but expressly carves out individual words and very short extracts, hyperlinks, and private/non-commercial uses. It would also risk undermining the CJEU's insistence that technological intermediation which does not expand the circle of recipients or circumvent access restrictions should not be treated as an autonomous communication.[166]

Gemini's output in *Like Company*, insofar as it arises from real-time retrieval, is thus conceptually adjacent to a search engine snippet. Treating it as a communication to the public would produce an unprecedented expansion of Article 3 InfoSoc and would effectively override the CJEU's balancing in the search-engine context. The *GEMA* decision[167] confirms the limits of this analogy: the Munich Court treated long verbatim reproductions of song lyrics as communication to the public, but only because the outputs contained substantial expressive copying. The reasoning turned not on a rejection of the search engine analogy, but on the presence of direct expressive appropriation. As the output in *Like Company* is merely a summary of a 579-word article generated via probabilistic prediction, the *GEMA* logic does not apply. Read together, *GEMA* and *Getty* support a calibrated standard: while *GEMA* shows that GenAI services may incur liability where outputs closely replicate protected works, *Getty* illustrates that such services are not treated as infringing *per se* in the absence of evidence that the model or its outputs embody reproductions of the claimant's works. The *Like Company* facts fall well short of that threshold.

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Accordingly, a principled handling of *Like Company* requires the Court to recognise the functional parallel between chatbot summarisation and conventional search engine results. Both operate as discovery and navigation tools; both rely on content that is freely accessible; both, at most, reproduce brief excerpts to assist users in deciding whether to consult the original source. The presence of an additional generative layer does not alter this basic communicative structure.

7 | PLATFORM'S LIABILITY FOR AI-GENERATED OUTPUTS PROMPTED BY USERS

This section addresses Question 4, *i.e.*, whether, when a user instructs an LLM-based chatbot like Gemini to reproduce or refer to the text of a press publication and the system responds by displaying part or all of that text, the service provider thereby commits an act of reproduction under Article 2 InfoSoc and Article 15(1) CDSM. On the best interpretation of the acquis, the answer should be no. Any rights-relevant reproduction in this configuration is attributable to the user who issues the prompt and causes the specific output; any potentially infringing passage appears only because the user deliberately requests it. The provider operates a general-purpose, content-agnostic, analytical tool whose functioning is automatic and neutral with respect to any given work. That allocation accords with (i) the CJEU's focus on who performs the protected act and on the difference between neutral technical intermediation and deliberate, content-specific intervention; (ii) the design of the PPR, which is aimed at uses by information society service providers while excluding private or non-commercial uses by individual users, and (iii) the intermediary liability framework (*i.e.*, e-Commerce Directive and the DSA),[168] which predicates provider responsibility on knowledge or control over specific items and rejects general monitoring obligations.

7.1 | Who performs the "reproduction"? attribution, volition and the nature of the tool

The scenario is user-initiated: a prompt instructs the system to reproduce, or closely paraphrase, a specific press passage. The model then generates a response probabilistically from that input, and in some deployments from documents retrieved at inference time. The attribution question is therefore who makes the copy. At the output layer, the proximate actor is the user: she selects the work, specifies the extent, and triggers the generation event, *i.e.*, the prompt. The provider, by contrast, supplies a general-purpose tool that tokenises input and predicts continuations; it does not exercise content-specific choice over the article or passage at issue. In short, volition and selection lie with the user; the provider's role is instrumental.

Multiple arguments support this conclusion. First, EU copyright, and the CJEU case law, do look at who performs the act in functional terms and whether the alleged infringer plays a deliberate role in bringing about the protected act. That logic tracks the CJEU's insistence, across rights, on a deliberate intervention by the person said to perform the protected act, and its distinction between neutral, automatic intermediation and content-specific control. In communication to the public cases, liability turns on intentional intervention directed at a public and, in some lines, on knowledge of illegality.[169] In hosting and platform cases, the Court distinguishes a passive, neutral technical service, benefiting from safe harbour, from an operator that plays an active role giving it knowledge or control over specific unlawful content. In *Google France* which focused on AdWords and trademarks, the Court emphasised the provider's 'mere technical, automatic and passive' role; and in *L'Oréal/eBay* and in *YouTube/Cyando* it framed liability around knowledge and control over specific content and the presence, or absence, of an active role.[170]

Second, in this context, the role of user agency and system unpredictability matters. LLM outputs are user-initiated and stochastic. The probabilistic nature of LLM outputs means providers cannot reasonably foresee every verbatim reproduction *ex ante*, especially with complex prompts. Even diligent providers cannot foresee, *ex ante*,

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

every possible chain of tokens that a creative or adversarial prompt may elicit, especially where the prompt itself embeds protected text, a common pattern in extraction experiments. The computer-science evidence dovetails with this allocation. Empirical work demonstrates that verbatim training-data emissions, or regurgitation, are uncommon, tend to arise under attack-style prompting and high-duplication training examples, and are substantially reduced through dataset deduplication, privacy-oriented training, output filters, and retrieval policies that favour linking/summarisation over long verbatim spans. Carlini et al. show verbatim extraction is possible but conditional[171]; Lee et al. document an order-of-magnitude reduction in verbatim emissions after deduplication[172]; later studies quantify how model scale, duplicate frequency and prompt length interact with memorisation.[173] These findings support a targeted, notice-and-mitigate governance model over strict provider-side primary liability for user-prompted reproductions. These are precisely the guardrails a negligence-based standard should reward.

Third, that the provider's servers instantiate the bits does not decide attribution: the legally relevant act is user-directed and user-specific. In this regard, two familiar objections fall away: that the provider is the reproducer merely because its infrastructure instantiates the output; and that its design and tuning of the model make it the reproducer of any protected text the system may emit. First, a 'server-location' theory, *i.e.*, whoever's hardware hosts the output is the reproducer, has no footing in the case law; attribution turns on deliberate, content-specific intervention, not on where the bits reside.[174] Second, model design and tuning do not make the designer the reproducer of any particular text: in *Infopaq*, the reproducer was the party that organised and ran the copying process for its own ends; in *PRCA v NLA*, transient technical copies integral to a process were held exempt to preserve technological development.[175] Where the user specifies the subject and extent of any reproduction and the provider offers a general-purpose tool with reasonable guardrails, recasting the provider as the primary 'reproducer' asks Article 2 InfoSoc to do work it was not designed to do. The reproduction right protects fixation of expression; it is not a general control over analytical capacity. Stretching it would collapse the acquis' neutral-tool/content-actor distinction and, in practice, invite *ex ante* screening of prompts and outputs, *i.e.*, general monitoring that EU law rejects on proportionality grounds.[176] The result would be to convert copyright into near-strict liability for learning and generation system design, a function the EU legal order assigns instead to horizontal due-diligence regimes such as the DSA and the AI Act and to targeted, output-layer remedies.

Fourth, assigning the act to the user preserves the functioning of exceptions and limitations where appropriate. Article 5(3)(d) InfoSoc enables quotation for purposes such as criticism or review, subject to fair practice and acknowledgment of source; whether a particular reproduction qualifies depends on context and purpose, which are facts known to the user who issued the prompt.[177] By contrast, a regime that imposed strict primary liability on providers would encourage blunt *ex ante* filtering that cannot reliably distinguish quotation from verbatim copying for other purposes, threatening the fair balance the CJEU polices across exceptions.[178]

Finally, from a law-and-economics perspective, placing primary liability on the prompting user minimises both error costs (over-blocking lawful uses such as quotation) and administrative costs (monitoring and assessing billions of prompts). In that environment, the least-cost avoider[179] for one-off, prompt-engineered reproductions is typically the user who chooses the target sequence and drafts the instruction, not the provider, absent evidence of design choices that encourage regurgitation. The user knows whether her prompt targets a protected text and whether an exception, *e.g.*, quotation, criticism/review or parody, is implicated. Providers, as repeat players, bear a duty of care to reduce foreseeable, systematic regurgitation but are not taxed for idiosyncratic, one-off prompts that only a general monitoring obligation could catch. Assigning strict primary liability to providers would effectively require generalised anticipatory filtering, which the acquis disfavours.[180] It would also amplify anti-commons effects.[181] While both users and providers face anti-commons risks, a provider-primary liability rule creates a 'Schrödinger's-cat' uncertainty: before any prompt, at the model's design and tuning stage, each prospective output sits, *from the provider's vantage point*, in a superposition of lawful and unlawful, collapsing only when the user's purpose materialises at the point of use.[182] This is because the provider is not itself using the works at that stage; it builds a neutral, general-purpose instrument that merely enables innumerable potential uses of the corpus of human knowledge (some lawful, some unlawful) by users, so the legal character of any given output turns on

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

context the provider cannot know *ex ante*. Yet the provider, a neutral toolmaker rather than a user of the works, would be forced to behave as if collapse had already occurred, pre-clearing or blocking *ex ante*. Lacking the user's context, it cannot rely on purpose-dependent exceptions (*e.g.*, quotation), so any arguable overlap must be treated as potentially infringing, multiplying veto points and inflating transaction costs.

7.2 | Intermediary liability under the E-Commerce directive and DSA: Knowledge, neutrality and no general monitoring

Under the EU intermediary liability regime, chatbot service providers' position is analogous to that of a hosting service protected by the E-Commerce Directive and the DSA. As noted earlier, the E-Commerce Directive establishes hosting safe harbours for information provided by recipients and prohibits general monitoring.[183] The DSA preserves the no-general-monitoring principle and adds structured *ex post* accountability, *i.e.* notice-and-action, statements of reasons, and for *Very Large Online Platforms* (VLOPS) and *Very Large Online Search Engines* (VLOSEs) systemic risk assessments, and audits.[184]

The CJEU has consistently rejected general and permanent filtering obligations as incompatible with fundamental-rights balancing[185]; however, it has accepted specific and targeted injunctions (including "equivalent content" in limited conditions),[186] and has framed platform responsibility around knowledge, control and active role, not the mere availability of a general-purpose service.[187] The Court's platform jurisprudence across rights, from *Google France* through *L'Oréal/eBay* to *YouTube/Cyando*, sets out a fault-sensitive allocation of responsibility. Hosting providers are not directly liable merely because illegal content passes through their systems; they risk primary liability only when they actively select or promote unlawful content, or when, having actual or constructive knowledge of specific illegality, they fail to act expeditiously.[188] Nothing in that case law suggests that a provider becomes the maker of the copy simply because its infrastructure executes the user's instruction.

For outputs generated by chatbots like Gemini, that translates to: (i) no strict *ex ante* policing of all possible verbatim emissions, *i.e.*, general monitoring is barred; (ii) expeditious action on notice concerning concrete passages; and (iii) proportionate, system-level safeguards under the DSA without collapsing into strict primary liability for user-prompted reproductions. This reading aligns with the CJEU's reasoning in *YouTube/Cyando*, which shields diligent platforms whose business model is not predicated on copyright infringement.[189] The Court draws a sharp line between a neutral host and an operator that, through knowledge/control or a model that incentivises infringement, plays an active role.[190] Where a provider implements proportionate guardrails and preventive tools and operates a reliable notice-and-action mechanism, the hosting exemption remains available and the provider is not treated as performing a protected act on users' behalf. Transposed to LLM chatbots, a service that deploys reasonable output safeguards and acts expeditiously on specific notices meets the E-Commerce Directive and DSA standard; any rights-relevant reproduction triggered by a prompt is attributable to the user, not the toolmaker. However, while the general-purpose model remains a neutral tool, liability could re-emerge if a provider specifically fine-tunes or curates the model to 'regurgitate' protected works, thereby playing an essential role in the communication of that content.[191]

7.3 | Interplay with the PPR and exceptions

The PPR under Article 15 CDSM is specifically designed to target online uses by information society service providers, while expressly exempting 'acts by individual users' and the use of 'individual words or very short extracts'.[192] Reading these statutory definitions alongside the CJEU's platform jurisprudence just mentioned implies a clear division of labour: service-provider liability attaches only where the provider *itself* performs the act of communication, *e.g.*, by actively curating or pushing protected content to the public, and not where an individual user elicits a specific extract *ad hoc* from a neutral, general-purpose tool.[193]

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

To hold otherwise would be to nullify the 'individual user' exemption by re-characterising user-initiated acts as provider acts. This connects directly to the referring court's query regarding whether Article 3(2) InfoSoc 'may exclude' the application of Article 15 CDSM. This framing is decisive: it highlights the risk that a broad, unanchored application of the general making-available right could bypass the specific legislative compromises of Article 15. If the dispute is not channeled through the *lex specialis* of Article 15, with its tailored thresholds for 'very short extracts' and user exemptions, claimants could effectively ignore the limits the EU legislature imposed on the related right.

A contrary reading would invert the logical hierarchy of the acquis. It would risk treating the related right (Article 15 CDSM) as a strict-liability regime that is functionally broader and more burdensome than the author's core rights (Articles 2 and 3 InfoSoc) upon which it is built. By design, Article 15 is framed as a complementary related right, not a mechanism to exceed the protection afforded to the underlying author's right.[194] If Article 15 CDSM were interpreted to impose strict primary liability on a provider for every user-elicited output, it would create a regime functionally stricter than the author's core rights under Articles 2 and 3 InfoSoc. As detailed earlier, under established case law, a neutral platform is not liable for user-posted content absent knowledge or an active role. If Article 15 were to bypass this neutrality defense and hold providers automatically liable for user prompts, it would grant the derivative rightholder (the publisher) a 'super-right' more powerful than that of the original creator (the author), thereby inverting the system's intended balance.

Therefore, the question of whether a chatbot provider 'reproduces' press content must be answered in a way that preserves the integrity of Article 15 CDSM. A fault-conditioned, neutrality-preserving approach to provider responsibility helps maintain that architecture. It preserves meaningful remedies for press publishers where a provider plays an active role or fails to act after concrete notice, without letting the mere provision of a general-purpose conversational tool turn Article 15 into a route to *de facto* strict primary liability.

7.4 | Liability allocation, remedies and workable standards

Some may worry that placing primary liability on the user under-protects rightholders. But the horizontal framework already supplies effective remedies without general monitoring. This is not a plea for laissez-faire but for calibration: provider duties should track the DSA/E-Commerce model, *i.e.* notice-and-action, narrowly proportionate *ex-ante* checks for long verbatim passages, rate-limits and guardrails on prompt patterns associated with regurgitation, transparent redress, and where necessary targeted injunctions against specific extraction behaviours.[195]

Nothing in this allocation deprives right-holders of remedies, but it is worth being candid about where those remedies operate. A claim against the individual user is of limited practical value where the user has merely elicited a passage in a private, ephemeral exchange: there is no persistent, publicly accessible item to notice or remove, the interaction is frequently a private or non-commercial use falling outside the PPR in any event, and litigation against dispersed end-users is neither feasible nor, given those exemptions, generally warranted. Such a claim becomes meaningful only where the user disseminates the excerpt (by republishing it on a website or social platform) at which point it is a classic act of making available, actionable against the user and removable from the host through ordinary notice-and-action.[196] For the recurring, system-level phenomenon that actually concerns rightholders, the operative remedies run not against the user but against the provider, on the fault-based footing the intermediary liability framework supplies: (i) DSA notice-and-action, which obliges the provider to act expeditiously on a substantiated notice, here not by deleting a vanished output but by adjusting the system prospectively so the offending class of output ceases to recur[197]; (ii) targeted, proportionate injunctive relief against the provider, crafted with *UPC Telekabel* proportionality to address the specific extraction behaviour without imposing general monitoring[198]; and (iii) where a provider's design choices systematically produce regurgitation and it fails to adopt obvious mitigations, loss of safe harbour and damages aligned with *YouTube/Cyando*.[199]

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Properly understood, then, these levers operate prospectively and at the system level: they do not so much remove a unique, already-vanished output as require the provider to prevent its recurrence (through output filtering, retrieval design, or pattern-level blocking) without turning providers into strict-liability guarantors for every private prompt. A concrete illustration shows how this operates. Suppose a publisher finds that a particular prompt (say, an instruction to 'continue' its article from a supplied opening line) reliably causes the model to emit a recognisable passage of the original. The publisher cannot take down that output, which is generated afresh on each request and retained nowhere; what it can do, under Articles 16 and 17 DSA, is notify the provider, supplying the prompt and the reproduced passage. An expeditious and effective response then operates entirely on the provider's side and prospectively: adding the work, or a fingerprint of it, to an output classifier of the kind already deployed to detect near-verbatim reproduction of protected text[200]; constraining the retrieval or grounding layer so that the source is summarised rather than reproduced; or rate-limiting and blocking the specific adversarial prompt pattern, so that the class of offending output ceases to recur, rather than a single instance being removed after the fact. Where voluntary action proves inadequate, a *UPC Telekabel*-style injunction can require the provider to achieve that result by means of its own choosing, provided the measure targets the specific behaviour and stops short of general monitoring[201]; and where the provider's own design or tuning is the source of the regurgitation, the neutrality defence falls away and ordinary remedies in damages follow. In none of these cases does effective enforcement depend on pursuing the end-user or on capturing an ephemeral output after the fact; it depends on the provider's prospective, system-level diligence, which is precisely what the DSA and the Court's injunction case law are designed to elicit. For the residual informational substitution that the exclusive rights do not reach, no such per-output mechanism is available or needed: there, the operative response is a remuneration scheme, administrable precisely because it bypasses output-by-output enforcement and operates at the licensing layer instead.[202] That is precisely the compliance pattern endorsed in *YouTube/Cyando* (diligent platforms are shielded; actors whose business models hinge on infringement are not) and it preserves space for innovation and for lawful uses such as quotation and very short extracts.[203]

8 | CONCLUSION

This article has examined the *Like Company* reference to the CJEU to elucidate how Articles 2 and 3 InfoSoc, read alongside Article 15 CDSM, apply to GenAI systems that summarise press content. Beyond the specific facts of the Hungarian dispute, the analysis tries to clarify fundamental doctrinal questions with EU-wide implications and draw conclusions: training cannot, as such, constitute reproduction under Article 2 InfoSoc as it fixes, in ordinary operation, parameters (weights) not protected expressions; anyway, if it is accepted that technical copies arise in the training pipeline, they are covered by the Article 4 CDSM TDM exception (absent a valid opt-out reservation); RAG-generated summaries that remain within 'very short extracts' fall outside Article 15(1) CDSM; and outputs that paraphrase rather than reproduce do not communicate the original publication to a new public. On platform liability, the article concludes that rights-relevant reproduction should be attributed to the user formulating the prompt, rather than the neutral service provider, with liability conditioned on knowledge of specific infringements. Applied to *Like Company*, this means Gemini's training process, operating under the CDSM's TDM framework, does not require publishers' authorisation, and its summarisation outputs do not infringe the PPR.

Critically, the interpretation proposed here preserves the structural relationship between Articles 2 and 3 InfoSoc and Article 15 CDSM: the PPR remains a targeted, complementary related right that does not extend protection beyond the underlying authors rights, but applies those rights in a defined online context, including *vis-à-vis* GenAI services. As the referring court captures in its language of whether Article 3(2) InfoSoc "may exclude" the application of Article 15 CDSM,[204] the two regimes must be understood hierarchically in this context. Article 15 operates as the specific, governing interface for the use of press publications by information society service providers. While it borrows the doctrinal definitions of Articles 2 and 3 InfoSoc, its specific limitations, particularly

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

regarding 'very short extracts', must take functional precedence. This ensures that the two regimes do not simply cumulate to trap lawful uses, but rather articulate a coherent system where the specific choices of the EU legislature regarding press content are respected.

However, this framework also reveals a critical gap: while the TDM exception enables lawful training and the InfoSoc-CDSM interaction prevents overreach, neither clearly nor fully ensures equitable remuneration for individual creators whose works fuel GenAI systems. This brings into focus an apparent paradox, and the stance that resolves it. Article 15 was enacted to arrest the erosion of publishers' revenues by services that mediate access to news; yet, on the analysis offered here, the technology now most capable of diverting readers from the source (i.e., generative summarisation grounded in retrieval) may fall outside the related right altogether, even where no opt-out has been exercised.

The paradox is genuine, but it is a property of the structure of copyright rather than a flaw in the argument. Copyright and the PPR protect expression and the investment in producing it, not the information a publication conveys; a service substitutes more, in the sense that matters to the exclusive rights, only when it appropriates more expression, not when it more efficiently conveys unprotected facts. A summary that reports the news in fresh language therefore neither reproduces the publication nor communicates it to a new public, however much it competes with the source for attention. To read the rights otherwise (to chase the economic harm by capturing facts-based substitution) would override the legislature's express exclusion of facts and very short extracts, breach the same-scope ceiling the CJEU's Grand Chamber has now affirmed, and convert the PPR into a monopoly over news as information, at a heavy cost to access and downstream innovation.

It does not follow, however, that the harm is illusory or that this article is indifferent to it, and the point is better stated plainly than left to emerge from the technical analysis. The diversion of informational demand away from journalism is real, and may bear more heavily on the press than on other content industries; what the analysis shows is not that this harm is negligible, but that the exclusive rights are the wrong instrument to meet it. Nor is Article 15 CDSM indifferent to authors: under Article 15(5), publishers must share an appropriate part of any PPR revenues with the journalists whose work they carry, so the related right is already a conduit of author remuneration, but a conduit that runs dry precisely where, as here, no protected expression is taken, and the right is not engaged. The gap is therefore twofold and structural: where only facts are conveyed, neither publishers (under the PPR) nor authors (through Article 15(5)) are compensated, and no purposive reading of the existing rights can supply that compensation without breaching the limits just described. That is why the response belongs with the legislature rather than the court, *i.e.*, a remuneration mechanism, whether a statutory or compulsory licence, a collective scheme, or a levy keyed to the use of journalistic content by GenAI systems, designed to channel value to publishers and, through them, to authors, for the informational use that the exclusive rights leave untouched. Recent scholarship and economic analysis converge on such 'permitted-but-paid' models as the most defensible way to reconcile broad access with sustainable creation,[205] and it is to that reform, not to a distortion of 'reproduction' and 'communication to the public', that the case for protecting journalism should be directed.

The doctrinal conclusion of this article and its distributive concern thus point the same way: keep the conceptual architecture of the exclusive rights intact, and address the political economy of AI and news where it belongs, in a purpose-built remuneration regime. This article, therefore, as mentioned in the Introduction, acknowledges that sustainable copyright governance must couple rigorous doctrinal analysis with a binding remuneration obligation that channels value to individual authors and creators whose intellectual contributions power GenAI systems, ensuring the creative workforce remains adequately compensated and incentivised across the EU.

ACKNOWLEDGMENTS

Enrico Bonadio is Professor of Intellectual Property Law at City St. George's, University of London, UK; Giancarlo Frosio is a Professor of Intellectual Property and Technology Law at the School of Law of Queen's University Belfast, UK; Christophe Geiger is Professor of Law at the Luiss Guido Carli University in Rome, Italy; Andres Guadamuz is Reader in Intellectual Property Law at University of Sussex, UK; Stavroula Karapapa is Professor of

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Intellectual Property and Information Law at the University of Essex, UK; Irini Stamatoudi is Professor of Intellectual Property Law at University of Nicosia, Cyprus. The authors are grateful to Alexander Clevewood NG for editing assistance. This article results from a research project on Generative AI and copyright law supported by Google LLC. The project was conducted in full academic independence, and all views and recommendations herein reflect the authors' personal academic judgment, for which they remain solely responsible. This article was prepared exclusively on the basis of publicly available materials. The authors had no access to non-public evidence (including internal logs, prompt transcripts, model weights, training datasets, indexing/caching configurations, or confidential technical documentation). Any description of the system's internal technical pathway is therefore necessarily inferential and is framed as such.

ORCID

Andres Guadamuz https://orcid.org/0000-0001-6092-9143

ENDNOTES

[1] See *Like Company v Google Ireland Limited*, C-250/25, request for a preliminary ruling from Budapest Környéki Törvényszék (Hungary) <https://curia.europa.eu/juris/liste.jsf?num=C-250/25&language=en> ("*Like Company* referral") accessed 3 April 2025. For another analysis of the case, see Mezei, P., and others, Comment of the European Copyright Society on the Request for Preliminary Ruling in Case C-250/25 (*Like Company*), *IIC*, 2026.

[2] 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, OJ 2001 L 167; 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, OJ 2019 L 130

[3] See *Like Company* referral, para 7.

[4] Ibid, para 8.

[5] See *Like Company* referral, para 9.

[6] Ibid, paras 19–21 and 23.

[7] The *Like Company* referral arrives at a critical juncture where proposals for a "permitted-but-paid" future legislative reform have been put forward by scholars in the debate, proposals which deserve close attention when reflecting on the possible revision of the AI Act and the CDSM. See Geiger, C., 'Elaborating a Human Rights friendly Copyright Framework for Generative AI' (2024) 55(7) IRIPCL 1129; Geiger, C. and Iaia, V., 'The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI' (2024) 52(105925) CLSR 1; Frosio, G., 'Should We Ban Generative AI, Incentivise It or Make it a Medium for Inclusive Creativity', in Bonadio, E. and Sganga, C. (eds.), *A Research Agenda for EU Copyright Law* (Elgar 2025), 61; Lucchi, N., 'Generative AI and Copyright', *Study for the JURI Committee*, July 2025, p. 122, admitting "that the long-term viability of individualized licensing across billions of works is limited" and that "a statutory licensing scheme or collective remuneration mechanism may ultimately provide a more scalable and equitable model"; Carlini, R., Menéndez, N. and Schiffrin, A., 'How to Update EU and US Copyright Regimes in the Age of AI', IPD Working Paper 2026, p. 17, according to whom "shifting to a default system of prior authorization coupled with a fixed payment scale seems to us to be the most likely to address the challenge of preserving journalism in the era of AI". For a similar proposal but focusing on the output, see Senftleben, M., 'Generative AI and Author Remuneration' (2023) 54 IIC 1535. In a recent economic study for the European Parliament, one of the key findings was that this regulatory solution was the optimal to balance all interests involved: see Peukert, C., 'The Economics of Copyright and AI, Empirical Evidence and Optimal Policy', In-Depth Analysis Requested by the JURI Committee, Dec. 2025, p. 7 (noting that "[s]tatutory licensing appears optimal. Among realistic options, *statutory licensing emerges as a robustly welfare-superior framework under most calibrations*. Making the license compulsory ensures wide access to works, a royalty rate set by a regulatory authority balances interests of rightsholders, AI developers and end users, enabling royalty payments that create incentives for enough new creation to keep AI systems valuable over time") (emphasis added). In general, on the necessity to reach a balance between the rights of content creators and the need not to jeopardise GenAI innovation, see Bonadio, E. and Pollicino, O., 'From Wall to Compass: Modernising EU Copyright in the Generative AI Era' (2026) 2 EIPR.

[8] See, *for example*, Russell, S.J. and Norvig, P., *Artificial Intelligence: A Modern Approach* (4th edn, Pearson 2020), 45.

[9] See, *for example*, for definitions used in a policy perspective, Bas, G. and others, 'The EU AI Act: A Pioneering Effort to Regulate Frontier AI?' (2024) 27 IA 55.

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[10] See Comanici, G., and others, 'Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities' (*arXiv*, 7 July 2025) <https://doi.org/10.48550/arXiv.2507.06261>; The paper discusses that while memorisation of training data is possible, models do not act as static databases where any input has been stored.

[11] See Georgiev, P., and others, 'Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context' (*arXiv*, 8 March 2024) <https://doi.org/10.48550/arXiv.2403.05530>

[12] See *Like Company* referral, para 5. See also Rohan, J. and others, 'Gemini: A Family of Highly Capable Multimodal Models' (*arXiv*, 13 December 2023) <https://doi.org/10.48550/arXiv.2312.11805>

[13] *Like Company* referral, para 5.

[14] See, *for example*, Frankle, J., Schwab D.J., and Morcos A.S., 'The Early Phase of Neural Network Training' (*arXiv*, 24 February 2020) <http://arxiv.org/abs/2002.10365>

[15] Ibid.

[16] See, *for example*, Choo, S. and Kim, W., 'A Study on the Evaluation of Tokenizer Performance in Natural Language Processing' (2023) 37 AAI.

[17] See Frankle, Schwab and Morcos (n 14).

[18] See, *for example*, Hendrycks, D., Lee, K., and Mazeika, M., 'Using Pre-Training Can Improve Model Robustness and Uncertainty' (*arXiv*, 20 October 2019) <http://arxiv.org/abs/1901.09960>

[19] See, on the process of training, Goodfellow, I., Bengio, Y., and Courville, A., *Deep Learning* (MIT Press 2016), Chapter 6.

[20] See, on the process of extracting meaning, Vaswani, A. and others, 'Attention is All You Need', in Guyon, I. and others (eds), *Advances in Neural Information Processing Systems 30* (2017), 5998–6008.

[21] See Comanici and others (n 10). See also Kudo, T., and Richardson, T., 'SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing' (*arXiv*, 19 August 2018) <https://doi.org/10.48550/arXiv.1808.06226> (noting that LLMs do not store recognisable words; and that they are broken up into tokens, not documents from the training data).

[22] See *for example*, Kaplan, J. and others, 'Scaling Laws for Neural Language Models' (*arXiv*, 23 January 2020), ss 1.1–1.3 <https://arxiv.org/abs/2001.08361>

[23] For a list of different model knowledge cut-off dates, see: <https://github.com/HaoooWang/llm-knowledge-cutoff-dates?tab=readme-ov-file>

[24] See, *for example*, Carlini, N. and others, 'Extracting Training Data from Large Language Models', in *Proceedings of the 30th USENIX Security Symposium*, USENIX Association (2021), 2633–2650.

[25] See, *for example*, Ying, X., 'An Overview of Overfitting and Its Solutions' (2019) 1168 *Journal of Physics: Conference Series* 022022.

[26] See, *fopr example*, Carlini, N. and others, 'Quantifying Memorization across Neural Language Models', in *Proceedings of the 11th International Conference on Learning Representations (ICLR)* (2023), 7–9 (on effects of model scale, duplication and prompt/context length; and deduplication as mitigation); Morris, J.X. and others, 'How Much Do Language Models Memorize?' (*arXiv*, 18 June 2025) <http://arxiv.org/abs/2505.24832> See also Sag, M., 'Copyright Safety for Generative AI' (2023) 61 Houston LR 295, 310, 326–336, 338 (discussing de-duplication from a copyright safety perspective and noting that de-duplication makes "memorisation of the training data more likely"); Cooper, A.F. and others, 'Extracting Memorized Pieces of (Copyrighted) Books from Open-Weight Language Models' (*arXiv*, 30 November 2025) <http://arxiv.org/abs/2505.12546>

[27] See, *for example*, Xia, W. and others, 'A Comprehensive Study of the Past, Present, and Future of Data Deduplication' (2016) 104 *Proceedings of the IEEE* 1681; Lee, K. and others, 'Deduplicating Training Data Makes Language Models Better', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics (2022), 8429–8431 (noting that dataset-level deduplication reduces verbatim regurgitation by up to an order of magnitude while providing quality gains).

[28] See, *for example*, Baptista, R. and others, 'Memorization and Regularization in Generative Diffusion Models' (*arXiv*, 18 March 2025) <http://arxiv.org/abs/2501.15785>

[29] See, *for example*, Zhang, T. and others, 'Balancing Learning Model Privacy, Fairness, and Accuracy with Early Stopping Criteria' (2023) 34 IEEE TNNLS 5557–5569.

[30] See, *for example*, Chaudhari, S. and others, 'RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs' (2025) 58 ACM CS. See also Church, K.W., Chen, Z. and Ma, Y., 'Emerging Trends: A Gentle

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Introduction to Fine-Tuning' (2021) 27 NLE 763–778 (describing fine-tuning as a way of optimising a model with specific information after training).

[31] See Cui, P. and Athey, S., 'Stable Learning Establishes Some Common Ground between Causal Inference and Machine Learning' (2022) 4 NMI 110–115.

[32] See, *for example*, Gupta, S., Ranjan, R., and Singh, S.N., 'A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions' (*arXiv*, 3 October 2024) <http://arxiv.org/abs/2410.12837>

[33] See, *for example*, Paulsen, N., 'Context Is What You Need: The Maximum Effective Context Window for Real World Limits of LLMs' (*arXiv*, 21 September 2025) <http://arxiv.org/abs/2509.21361>

[34] See, *for example*, Lewis, P. and others, 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks', in Hugo Larochelle and others (eds), *Advances in Neural Information Processing Systems 33* (Curran Associates 2020), 9459–9474 (explaining that in RAG mode the model looks up relevant documents when answering and then writes using both its learned knowledge and those retrieved texts, so the response stays grounded in external sources).

[35] See Google, 'Grounding with Google Search' (2025) <https://ai.google.dev/gemini-api/docs/google-search> It should be noted that public documentation describes these mechanisms at a general level and across model iterations; the publicly available record in *Like Company* does not specify the exact model/version and retrieval configuration deployed at the relevant time.

[36] Ibid.

[37] Ibid.

[38] Ibid.

[39] See *Like Company* referral, paras 17–18.

[40] See (n 14).

[41] See Mesnard, T. and others, 'Gemma: Open Models Based on Gemini Research and Technology' (*arXiv*, 13 March 2024) <https://arxiv.org/abs/2403.08295>; Cooper and others, see (n 26).

[42] See *Like Company* referral, para 8.

[43] See Google (n 35).

[44] See *Like Company* referral, para 26.

[45] The Grand Chamber's judgment in Case C-797/23, *Meta Platforms Ireland Ltd v Autorità per le Garanzie nelle Comunicazioni (AGCOM)*, ECLI:EU:C:2026:395, confirms that the first subparagraph of Article 15(1) CDSM, read with Articles 2 and 3(2) InfoSoc, "unequivocally defines the exclusive rights of reproduction and making available to the public enjoyed by publishers of press publications" and "constitutes a measure of full harmonisation of the corresponding substantive law" (paras 48–50), so that those rights "have the same scope as the rights of reproduction and making available to the public provided for by [InfoSoc], in so far as online uses of those publications by information society service providers are concerned" (para 59; see also Recital 57 CDSM). The asymmetry invoked here therefore lies not in the scope of the protected act, which is co-extensive across the two instruments, but in the additional internal carve-outs that Article 15 superimposes (individual words or very short extracts, hyperlinks, private or non-commercial uses by individual users, and a 2-year term) and in the PPR's distinct subject matter. The "same scope" holding reinforces our conclusion: because the related right is co-extensive with, and may not exceed, the underlying InfoSoc rights, the general right cannot be deployed to circumvent Article 15's calibrated limits (see further Section 4, "A cautionary note: harmonisation with the PPR's scope").

[46] Article 15(2) CDSM provides that the PPR "shall leave intact and shall in no way affect" the rights of authors and other rightholders in the works and other subject matter incorporated in a press publication, and "shall not be invoked against" them; see also Recital 57 CDSM.

[47] See indicatively the Study for the JURI Committee, 'Strengthening the Position of Press Publishers and Authors and Performers in the Copyright Directive', Legal and Parliamentary Affairs, Policy Department for Citizens' Rights and Constitutional Affairs, Directorate General for Internal Policies of the Union, PE 596.810-September 2017 <https://www.europarl.europa.eu/RegData/etudes/STUD/2017/596810/IPOL_STU%282017%29596810_EN.pdf?utm_source=chatgpt.com>

[48] See, *for example*, Gervais, D., 'The Machine as Author' (2020) 105 Iowa LR 2053–2106, esp. pp. 2053–2054 and 2065; Hugenholtz, P. B., Quintais, J. P., 'Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?' (2021) 52(9) IIC—International Review of Intellectual Property and Competition Law 1190–1216, esp.

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1195–1196, 1211–1212; Frosio G., 'Four theories in search of an A(I)uthor', in Abbott, R., and Geffen, D., (eds), *Research Handbook on Intellectual Property and Artificial Intelligence* (Elgar 2022), 155–177.

[49] Gaidartzi, A. and Stamatoudi, I., 'Authorship and Ownership Issues Raised by AI-Generated Works: A Comparative Analysis' (2025) 14(4) Laws 57 <https://www.mdpi.com/2075-471X/14/4/57>; Stamatoudi, I., 'Copyright Protection for AI-generated Works: Is this Something We Want?' (2025) CR.

[50] See Case C-476/17, *Pelham GmbH v Hütter*, EU:C:2019:624, paras 31 and 37–39 (finding that a protected sample reused in modified form infringes the phonogram producer's related right unless rendered "unrecognisable to the ear"); Joined Cases C-580/23 and C-795/23, *Mio AB and Others v Galleri Mikael & Thomas Asplund AB and USM U. Schärer Söhne AG v konektra GmbH and LN*, ECLI:EU:C:2025:941 ('*Mio/konektra*'), paras 86–87 (on recognisability of the protected elements, rather than overall impression or strict identity, as the test for reproduction). It follows that modification during reuse does not in itself defeat infringement of a related right; the question is whether the protected subject matter remains recognisable in the later material. For the PPR, whose subject matter is the press publication's expression as fixed rather than the facts it reports (Recital 57 CDSM), this means the right is engaged by a recognisable taking of that expression (verbatim or recognisably paraphrased) but not by the independent conveyance of the underlying information. That reading keeps the PPR aligned with, and no broader than, the reproduction right it borrows from InfoSoc, consistent with the "same scope" principle confirmed in *Meta* (n 45, para 59).

[51] On the dichotomy idea/expression, see Article 2(1) of the Berne Convention for the Protection of Literary and Artistic Works (Paris Act of 24 July 1971, as amended on 28 September 1979); Article 9(2) of the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS), annex 1C to the Marrakesh Agreement establishing the World Trade Organization, 15 April 1994; Article 1(2) of Directive (EU) No 2009/24 of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs; Case C-406/10, *SAS Institute Inc. v World Programming Ltd*, EU:C:2012:259, paras 39–41 and 45; Masouyé, C., *Guide to the Berne Convention for the Protection of Literary and Artistic Works* (Paris Act, 1971), WIPO, Geneva (1978), 11–12.

[52] Such an interpretation, recently adopted by the Regional Court of Munich in *GEMA v OpenAI* (see LG München I, Urteil vom 11.11.2025—42O 14139/24 <https://aifray.com/wp-content/uploads/2025/11/42-O-14139-24-Endurteil.pdf)> seems to rely on a misinterpretation of the technical functioning of GenAI models as generally described in scientific literature, even though some ambiguities seem to persist. For an in-depth critique of *GEMA*, see Frosio, G., 'Copyright in Formaldehyde: How *GEMA v OpenAI* Freezes Doctrine and Chills AI—Part 1 and 2' (*Kluwer Copyright Blog*, 10–11 December 2025) <https://legalblogs.wolterskluwer.com/copyright-blog/copyright-in-formaldehyde-how-gema-v-openai-freezes-doctrine-and-chills-ai-part-1> Guadamuz, A., 'Victory for GEMA in Germany against OpenAI' (*TechnoLlama*, 22 November 2025) <https://www.technollama.co.uk/victory-for-gema-in-germany-against-openai> On this decision, see also Dornis, T.W., Ginsburg, J., Lucchi, N., 'GEMA v. OpenAI (Munich I Regional Court, 2025): Doctrinal, Comparative, and International Perspectives on Copyright and Generative AI', 81(6) JZ 235 (2026) <https://scholarship.law.columbia.edu/faculty_scholarship/4774>

[53] See Cooper and others (n 26).

[54] A growing technical and doctrinal literature cautions against treating "memorisation" in model training as a proxy for copyright "reproduction" or infringement absent an infringing, substantially similar output: see, *for example*, Haviv, A. and others, 'We Should Separate Memorization from Copyright' (*arXiv*:2602.08632v1 [cs.CY], 9 February 2026) <https://doi.org/10.48550/arXiv.2602.08632>

[55] See, *for example*, Guadamuz, A., 'The Persistence of Memorisation' (*Technollama Blog*, 28 November 2025) <https://www.technollama.co.uk/the-persistence-of-memorization>

[56] The legal questions raised by chatbot summaries concern inference-time retrieval and display (RAG/output-layer use) and are addressed below in Section 6 (especially when it comes to focusing on Article 15(1) CDSM and the concept of "very short extracts").

[57] See, classically, Case C-379/98, *PreussenElektra AG v Schleswag AG*, EU:C:2001:160, para 39 (clarifying that the Court may refuse to answer where the question bears no relation to the facts of the main action or is hypothetical); Case C-621/18, *Wightman and Others v Secretary of State for Exiting the European Union*, EU:C:2018:999, para 26 (holding that the Court may, in order to provide a useful answer, reformulate questions referred).

[58] See Case C-5/08, *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465, paras 37–51, 64–66.

[59] See, for example, *SAS Institute* (n 51), paras 32–46 (confirming that functionality, programming language and data formats are not protected by copyright); Case C-604/10 *Football Dataco Ltd v Yahoo! UK Ltd*, EU:C:2012:115, paras 37–39 (on originality standard).

[60] See *Pelham* (n 50), paras 31–40 (clarifying that use "unrecognisable to the ear" falls outside reproduction of a phonogram).

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[61] *SAS Institute* (*n* 51), paras 39–46.

[62] See, *for example*, Goodfellow, Bengio, and Courville (n 19), Ch 8–10, pp. 294–296, 388–390 (regarding standard reference on gradient descent and parameter optimisation in neural networks).

[63] See,*for example*, Bengio, Y. and others, 'A Neural Probabilistic Language Model' (2003) 3 JMLR 1137, 1141; Vaswani and others (n 20), paras 3.2.2 and 3.4.

[64] See, *for example*, Bengio and others (n 63), 1141; Kingma, D.P. and Welling, M., 'An Introduction to Variational Autoencoders' (2019) 12(4) FTML 58–60 (noting that variational autoencoders provide a continuous latent space that enables smooth interpolation between inputs, capturing patterns rather than fixed copies of works). See also, confirming this key point, *Like Company* referral, para 5.

[65] See, *for example*, Hinton, G.E. and Salakhutdinov, R.R., 'Reducing the Dimensionality of Data with Neural Networks' (2006) 313 Science 504–506.

[66] See, *for example*, Carlini and others, (n 24), 2633–2635; Lee, K. and others, (n 27), 8424–8437, at 8424–8426 (on unprompted memorisation and role of duplicates), 8429–8430 & 8430–8431 (on 10× reduction in *verbatim* emissions after deduplication).

[67] See Bengioand others, (n 63), 1141; Vaswani and others, (n 20), paras 3.2.2 and 3.4.

[68] See *SAS Institute*, (n 51), paras 39–41 and 45 (clarifying that programming language, functionality and formats are not protected "expression"; and that copyright attaches to the author's expressive form, not functional/abstract elements).

[69] See *Pelham* (n 50), paras 31–40.

[70] See, *for example*, LeCun, Y. and others, 'Deep Learning' (2015) 521 Nature 436–437 (noting that deep models "learn representations of data with multiple levels of abstraction"; and that higher layers capture increasingly abstract features); *Infopaq* (n 58), paras 37, 48 and 51 (stating that reproduction "in part" requires fixation of expressive elements, not mere ideas or abstractions).

[71] This error lies at the heart of the *GEMA* decision by the Munich Court (see n 52), which conflated the probabilistic regeneration of text with the retention of fixed copies, a view that ignores the "lossy" nature of training and effectively freezes the doctrine of reproduction. See, *for example*, Guadamuz (n 52).

[72] See Saffran, J.R. and others, 'Statistical Learning by 8-Month-old Infants' (1996) 274 Science 1926–1927 (noting that infants track transitional probabilities in syllable streams, learning distributional regularities from exposure); Frost, R. and others, 'Domain Generality of Statistical Learning' (2015) 19(3) TCS 124–134, at pp. 124–126 (noting that humans acquire statistical structure across modalities; and that learning reflects sensitivity to co-occurrence and sequence regularities); Saffran, J.R. and Kirkham, N.Z., 'Infant Statistical Learning' (2018) 69 ARP 182–185 (on survey of evidence that human learners extract patterns and regularities from continuous input, forming abstract representations rather than memorising surface forms).

[73] See (n 51 and 52).

[74] See, *for example*, Tenenbaum, J.B. and others, 'How to Grow a Mind: Statistics, Structure, and Abstraction' (2011) 331 Science 1279–1280 (noting that human learning builds abstract structural representations from statistical evidence rather than storing surface details); Clark, A., 'Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science' (2013) 36(3) BBS 181–204, at 181–183 (noting that, on a predictive-processing picture, cognition encodes generative, abstract models of causes, not verbatim inputs).

[75] See *Bartz and others v Anthropic PBC*, N.D. Cal., Case No. 3:24-cv-05417-WHA, Order on Fair Use, 23 June 2025, p. 12 (holding that copying and storing pirated books to build a central, permanent library was not fair use, but describing training as "quintessentially transformative" and comparing it to "any reader aspiring to be a writer").

[76] See *Infopaq* (n 58), paras 37, 48 and 51; Case C-145/10, *Painer v Standard VerlagsGmbH and Others*, EU:C:2011:798, paras 89–92 (finding that a work is protected where it is the result of authors' free and creative choices and reflects their personality; and that protection attaches to those expressive choices); *SAS Institute* (n 51), paras 39–41 and 45; *Football Dataco* (n 59) paras 37–39 (confirming that protection concerns expressive choices, not mere data or effort).

[77] *Pelham* (n 50), paras 31–40.

[78] See *Mio/konektra* (n 50), paras 86–87.

[79] *Id*., para 87. See also Opinion of Advocate General Emiliou in Case C-590/23 *CG*, *YN v Pelham GmbH*, *SD*, *UP*, ECLI:EU:C:2025:452, para 64 (noting that style "is not, as such, protected by copyright" and belongs to the realm of ideas, not expressions; and that overall stylistic features and common motifs are not individually protected).

[80] Ibid.

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[81] *Infopaq* (n 58), paras 50–66.

[82] See Recital 18 and Article 4(1) CDSM (clarifying that TDM consists of automated analytical techniques that derive "patterns, trends and correlations" from digital text/data, *i.e.*, a recognition that such operations are analysis, not the copying of expressive strings).

[83] Lossiness alone is thus not decisive: a lossy MP3 or JPEG is a reproduction even though it discards perceptually redundant data, because what it retains still embodies, and permits perception of, the protected expression. What distinguishes a trained model is not only that the process is lossy but that it is abstractive: it discards expression in favour of non-expressive statistical correlations from which no particular work can, in ordinary operation, be reconstituted.

[84] See Bommasani, R., and others, 'On the Opportunities and Risks of Foundation (*arXiv*:2108.07258 [cs.LG], 2021) <https://arxiv.org/abs/2108.07258> 30–35 (noting that original expression is not recoverable from the model's weights alone); Wright, C., 'Training on the Tightrope: AI Copyright and Data Privacy as Colliding Regulatory Regimes' (2026) 61 CLSR 106343, paras 2.1 and 2.3 (arguing that expression is not recoverable from model weights alone).

[85] This is precisely why the contrary view (i.e. that works are reproduced "inside" the model) conflates the bounded, output-layer phenomenon of memorisation with the routine character of the trained artefact; where recoverable expression can be elicited from highly duplicated items, the appropriate response operates at the output layer, not by recharacterising the entire parameter space as a stored copy (see infra).

[86] See (n 42).

[87] See Hugenholtz, P.B., 'Against "Data Property"', in Ullrich, H. and others (eds), *Kritika: Essays on Intellectual Property: Vol. 3* (Edward Elgar 2018), 48–71 (warning that new exclusion rights over information would distort markets and innovation).

[88] See Lemley, M. and Casey, B., 'Fair Learning' (2021) 99 Texas LR 743, 748–49 (arguing that copyright should not create a *de facto* control right over the ability of machines, or people, to learn); Margoni, T. and Kretschmer, M., 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2022) 71 GRUR Intr. 688–690, 696–702 (warning that allocating control upstream over analytic uses risks regulating AI via property rights in inputs).

[89] For the engineering reality that training pipelines generate *ephemeral artefacts* (buffers, caches, tokenisation shards and sharded files) as part of efficient preprocessing and input streaming, see, *for example*, Kudo, T. and Richardson, J., 'SentencePiece: a Simple and Language-independent Subword Tokenizer and Detokenizer for Neural Text Processing', in *Proceedings of EMNLP 2018: System Demonstrations*, Association for Computational Linguistics (2018), 66–71 (focusing on tokenisation as a distinct preprocessing stage producing model files and segmented corpora used downstream); Murray, D.G. and others, 'tf.data: A Machine Learning Data Processing Framework', 14 *Proceedings of the VLDB Endowment* 12 (2021) 2945–2949, 2951, 2954–2955 (noting that standard pipeline operators include caching, prefetching and buffering to overlap I/O and compute, *i.e.*, transient, functional copies integral to training throughput).

[90] See, *for example*, Geiger, C. Frosio, G. and Bulayenko, O., 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market—Legal Aspects' (European Parliament, In-Depth Analysis, 2018) (explaining that TDM processes inherently trigger restricted acts—*i.e.*, reproduction/extraction—and recommending an exception framework, precisely to avoid treating machine *learning* as *copying*, while still addressing ancillary copies); Margoni and Kretschmer (n 77) 685–699 (emphasising that TDM entails copying steps but that EU law channels them through specific exceptions/opt-outs rather than recasting *analysis* as *reproduction*).

[91] See contra, *for example*, Dornis, T.W. and Stober, S., *Urheberrecht und Training generativer KI-Modelle—technologische und juristische Grundlagen* (Nomos, Baden-Baden 2024), 1–2 (stating expressly that during pre-training and fine-tuning "copyright-relevant reproductions [...] materialize 'inside' the AI model"; and that this constitutes copying in the legal sense); Dornis, T.W., 'Generative AI, Reproductions Inside the Model, and the Making Available to the Public' (2025) 56 IIC– IRIPCL 916–922 (sections 2.2.1 & 2.2.1.2: arguing that works used for training are reproduced "inside" the model in a vector/latent space and that regurgitating outputs imply an internal copy); Lucchi, N., *Generative AI and Copyright: Training, Creation, Regulation*, Study for the European Parliament (Policy Department for Justice, Civil Liberties and Institutional Affairs), PE 774.095, July 2025, p. 30 (noting that "AI training entails systematic reproduction of protected works"); pp. 34–45 and 78 (maintaining that ingestion by generative systems "typically involves acts of reproduction beyond analytical use"), and p. 125 (pointing out that reproduction right is engaged even if use serves a new purpose).

[92] See (n 78).

[93] See, *for example*, European Commission, Commission Staff Working Document, *Impact Assessment on the modernisation of EU copyright rules*, SWD (2016) 301 final, Part 1/3, Brussels, 14 September 2016, s. 4.3 "Text and data mining", pp.

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

105–106 (on legal uncertainty and diversity of licensing practices generating transaction costs, including for compliance).

[94] See Emanuilov, I. and Margoni, T., 'Forget Me Not: Memorisation in Generative Sequence Models Trained on Open Source Licensed Code', SSRN research paper n. 4720990, 2024 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4720990> (showing that problematic verbatim "memorisation" is the exception, underscoring why models should not be equated with their training corpora).

[95] See (n 16–21).

[96] See, *for example*, Microsoft, 'Protected material detection filter' (*Microsoft Learn*, 16 September 2025) <https://learn.microsoft.com/en-us/azure/ai-foundry/openai/concepts/content-filter-protected-material?tabs=text> (describing output-side filters that flag and block content closely matching copyrighted text or code).

[97] See, *for example*, Kumar, P., 'Adversarial Attacks and Defenses for Large Language Models (LLMs): Methods, Frameworks & Challenges' (2024) 13 IJMIR, art. 26.

[98] See Haviv and others (n 54), 1–14 (arguing that technical "memorisation" findings in foundation models are often not evidence of copyright "reproduction" or unlawful copying, because they can reflect lawful generalisation or high-frequency material, and calls instead for an output-level, risk-based assessment aligned with copyright standards to identify *actual* infringement risk). See also, for the same conclusion in the US fair use context, E. Lee, 'Fair Use and the Origin of AI Training' (2025) 63(1) Houston LR 104, esp. p. 212–215 (noting that 'Internal "Memorization" Should Not Be Infringing Absent a Corresponding Infringing Output').

[99] Ahmed, A. and others, 'Extracting Books From Production Language Models' (arXiv preprint *arXiv*:2601.02671, 2026) <https://arxiv.org/abs/2601.02671>

[100] Ibid.

[101] Ibid, p. 19 ("for the cases in which we retrieved whole copies of near-verbatim books, it was often quite costly (> $100) to do so").

[102] Ibid.

[103] See, *for example*, Robison, K., 'AI Models Can Memorize Books, and That's a Problem' *The Atlantic* (7 January 2026) <https://www.theatlantic.com/technology/2026/01/ai-memorization-research/685552>

[104] See Ahmed and others (n 99). Some further joint interdisciplinary research involving tech experts and legal scholars would be useful to firmly clarify all issues, though (see in this spirit Guadamuz, A., 'Why We Need Tech Lawyers to Shine Again' (2025) 56 IIC 1797), as the current technical results give rise to apparently very different interpretations among legal scholars.

[105] See generally, on market and subject matter dependant substitution risk, Senftleben (n 7), 1535, 1535–37, 1549–50 (linking remuneration to the substitution risk posed by AI outputs and proposing repartition by work repertoires, genres and prompt data); Frosio (n 7) 61, 81–83 (proposing an input-side levy and calibrating statutory remuneration to the creative markets most disrupted by AI-generated creativity); Christian Peukert, *The Economics of Copyright and AI: Empirical Evidence and Optimal Policy*, European Parliament, Policy Department for Justice, Civil Liberties and Institutional Affairs, PE 778.859, December 2025, 14, 17–18, 37, 49 (finding heterogeneous labour and substitution effects, with demand falling for translation, writing and graphic design services, while aggregate substitution effects may be limited but distributional impacts on specific sectors or rightsholder groups may be significant); Lucchi (n 7), 125, 147–48 (stating that the claim that AI training does not compete with original works is context-dependent and that sectors such as illustration, journalism and stock photography face direct market disruption, while proposing output-side measures linked to substitution thresholds and content categories).

[106] See *Meta* (n 46), para 59.

[107] See (n 8).

[108] See *GEMA* (n 52).

[109] See, *for example*, Guadamuz (n 45).

[110] Recital 57 CDSM.

[111] Ibid; see also Reto M. Hilty and others, 'Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules' (2017) para 28 (warning against related rights creating super-protection for non-original content).

[112] See again Recital 57 CDSM (clarifying that the "rights granted to publishers of press publications should not extend […] to mere facts reported in press publications").

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[113] Ibid, (mentioning that the "rights granted to the publishers of press publications under this Directive [CDSM] should have the same scope as the rights of reproduction and communication to the public provided for in Directive 2001/29/EC [InfoSoc]…").

[114] See Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), OJ L, 2024/1689, 12.7.2024, Recital 105.

[115] See above the technical analysis (Section 2) and the legal assessment of the right of reproduction issues (Section 4).

[116] See Guadamuz, A., 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs' (2022) 73 GRUR Intr. 11; Lucchi, N., 'ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems' (2023) European JRR 1.

[117] Around 70 copyright-related lawsuits have been filed thus far against GenAI system operators in the US. Several scholars argue it is likely fair use. See Sag (n 27) 295; Lemley, M., 'How Generative AI Turns Copyright Law Upside Down' (2024) 25 CSTLR 21; Samuelson, P., 'Fair Use Defenses in Disruptive Technology Cases' (2024) 71 UCLA LR 1484; Lemley and Casey (n 89) 743. This position however is not undisputed: see Brauneis, R., 'Copyright and the Training of Human Authors and Generative Machines' (2024) 48 Columbia JLA 1; Ginsburg, J.C., 'AI Input Data and Fair Use: A View from the U.S.' (2024) AM 165.

[118] Japan has established one of the most permissive legal environments globally for TDM, and has therefore been sometimes called by Japanese scholars a "paradise for GenAI training". Indeed, Article 30-4 of the Japanese Copyright Act allows the use of copyrighted materials for "non-enjoyment purposes", which includes TDM for commercial purposes, without the need for rights holder's permission or compensation: see Ueno, T., 'The Flexible Copyright Exception for 'Non-Enjoyment' Purposes—Recent Amendment in Japan and Its Implication' (2021) GRUR Intr. 145.

[119] For a comment see de Champris, B., 'Artificial Intelligence and Copyright: The EU Should Preserve the Copyright Directive's Delicate Balance to Safeguard and Promote Innovation' (2024) 1 *AIRe* 113; Peukert, A., 'Copyright in the Artificial Intelligence Act—A Primer' (2024) 73(6) GRUR Intr. 497.

[120] Answer given on 31 March 2023 by Commissioner Breton on Behalf of the European Commission (Parliamentary written question for written answer, E-000479/2023, posed on 15 February 2023: "How does the Commission plan to regulate this use of AI, which harms artists and rights holders?") <https://www.europarl.europa.eu/doceo/document/E-9-2023-000479-ASW_EN.html>

[121] See, *for example*, Senftleben, M., 'GenAI and the Copyright Three-Step Test—Do TDM Exceptions for AI Training Conflict with a Work's Normal Exploitation?' (2025) 74 GRUR Intr. <https://doi.org/10.1093/grurint/ikaf113> (also denying a possible conflict of the TDM exception used for AI training with the three-step test); Senftleben, M., 'Text and Data Mining, Generative AI, and the Copyright Three-Step Test' 57 (2025) IIC 67–107; 'Copyright and Generative AI: Opinion of the European Copyright Society' (2025) 16(1) JIPITEC 121.

[122] The argument that the TDM exception has not been tailor-made for GenAI purposes (which is correct) and therefore might not cover all reproductions relevant to the process seems therefore not relevant to determine the legality of training, as the EU legislator decided to submit the training to the exception.

[123] See on the issue more broadly Manteghi, M., 'Readjusting the EU's Regulation of Text and Data Mining in the Age of Artificial Intelligence', PhD defended at the University of Turku, 28 Nov. 2025, Annales Universitatis Turkuensis, 2025.

[124] On these exceptions see Geiger, C., Frosio, G., and Bulayenko, O., 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU', in Saiz Garcia, C. and Evangelio Llorca, R. (eds.), '*Propiedad* Intelectual Y Mercado Único Digital Europeo' (Tirant lo Blanch 2019), 27. For an analysis prior to the CDSM see also Stamatoudi, I., 'Text and Data Mining', in Stamatoudi (ed.), 'New Developments in EU and International Copyright Law' (Leiden, Netherlands: Kluwer Law International 2016) 264–265.

[125] Question which was addressed by the Munich Regional Court in *Gema* (on this decision see n 52 and again *infra*).

[126] See Municipal Court of Appeals, Case 9.Pf.20.353/2024/6-II, 3 December 2024.

[127] See ECLI:NL:RBAMS:2024:6563, 30 October 2024.

[128] See Regional Court (LG) of Hamburg (LAION vs Kneschke), 27 Sept. 2024, Case No. 310 O 227/23 confirmed by Fifth Civil Senate of the Hanseatic Higher Regional Court (OLG Hamburg), 10 December 2025, Case 5 U 104/24.

[129] See Dornis, T.W. and Stober, S., 'Generative AI Training and Copyright Law', Transactions of the International Society for Music Information Retrieval, V (2025) <https://arxiv.org/pdf/2502.15858> and from the same authors, (n 80); Dornis, T.W., 'The Training of Generative AI is not Text and Data Mining' (2025) 47 European IPR 65–78. See also, for

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

a rejection of a broad commercial TDM exception in the UK, HM Government, *Report on Copyright and Artificial Intelligence*, CP 2602959, 18 March 2026, p. 11, paras 25–27 (stating that the broad copyright exception with opt-out is no longer the Government's preferred way forward); House of Lords Communications and Digital Committee, *AI, copyright and the creative industries*, 4th Report of Session 2024–26, HL Paper 267, 6 March 2026, pp. 4–5 and 74 (stating that a broad commercial TDM exception would harm rightsholders and recommending that the Government rule out a new commercial TDM exception with an opt-out model).

[130] See Rosati, E., 'Infringing AI: Liability for AI-Generated Outputs under International, EU and UK Copyright Law' (2025) 16 European JRR 610–611 (stressing that AI training entails some copyright-relevant acts, *i.e.*, reproduction, communication/making available to the public, are not covered by the TDM exception under Articles 3 and 4 CDSM); Dornis, T.W., and Lucchi, N., 'Generative AI and the Scope of EU Copyright Law: A Doctrinal Analysis in Light of the Referral in *Like Company* v. Google' (2025) IIC. See also European Parliament, JURI, 'Report on copyright and generative artificial intelligence—opportunities and challenges' (Rapporteur: Axel Voss), A10-0019/2026, 25 February 2026, adopted on 10 March 2026. Even if in the opinion of the Rapporteur, Article 4 is not applicable to GenAI, the report makes (welcomed) proposals on how to make the opt-out more operational, thus endorsing implicitly the application of the TDM. Ultimately, on the TDM exception and on other issues, the Report has many contradictory statements and incoherences (see for an analysis in this sense, Geiger, C., Bossi, L. and Di Lazarro, F., 'The Voss Parliamentary Report on Copyright and Generative AI: A Wrong Step in the Right Direction?', ILEO Research Paper No 26-01.

[131] Alonso, E., and Lucchi, N., 'AI And Copyright 'Hallucinations': Does the Text and Data Mining Exception Really Support Generative AI Training?' (2025) 47(9) European IPR 515–526.

[132] For an early comment on this case, see Peukert, A., 'Vervielfältigung im KI Model? Zugleich Anmerkung zu LG München I, Urteil vom 11.11.2025', ZUM 2026, p. 17. On memorization see also Cooper, A.F. and Grimmelmann, J., 'The Files are in the Computer: on Copyright, Memorization, and Generative AI' (2025) 100 *Chicago-Kent LR* I141–219. Likewise, in *GEMA* (n 52) the Munich court decided that the output of large portion of song texts based on prompts qualify as communication to the public, which however has not been much disputed so far, as verbatim take-over of identical parts of the protected work in the output can amount to a copyright infringement (see our analysis below in Section 5).

[133] See in this sense Geiger, C. and Iaia, V., 'Is There a Captain in the Ship? The EU Copyright Regulator's Quest in the Generative AI Era', *ILEO Research Paper Series* No. 25-09 (2025).

[134] More detailed in this sense is *for example*, Mezei, P., 'A Saviour or a Dead End? Reservation of Rights in the Age of Generative AI' (2024) 46(7) European IPR 461.

[135] The Copyright Knowledge Center at EUIPO was launched on the 21 November 2025 <https://www.euipo.europa.eu/en/copyright-knowledge-centre> the Center will in particular "explore new services that enhance copyright transparency and accessibility across the Single Market".

[136] On the potential future role of the EUIPO, and on more generally on the need for governance on this issue at EU level, see (n 124); Christophe Geiger and Vincenzo Iaia, 'Towards an Independent EU Regulator for Copyright Issues of Generative AI: What Role for the AI Office (But More Importantly: What's Next)?' (2024) (2) AM 185.

[137] See *Football Dataco* (n 59), paras 38–39 (confirming that copyright does not protect "information" as such, but the author's own intellectual creation in selection/arrangement); *SAS Institute* (n 51), paras 39–41 (noting that functionality, programming language and data formats are not protected expression).

[138] See *Infopaq* (n 58), paras 37, 48 and 51 (confirming that reproduction "in part" depends on fixation of elements expressing the author's own intellectual creation).

[139] Ibid.

[140] See *Mio/konektra* (n 50), paras 86–87 and 92–93.

[141] See *Pelham* (n 50), para 38–39.

[142] For the same reason, reliance on the quotation exception is generally misplaced. Quotation presupposes the incorporation of an identifiable extract from the protected work (in *Pelham*, a sample recognisable to the ear; for literary works, an extract that is, in substance, taken verbatim). Where the output is a paraphrase rather than an identifiable excerpt, it is more naturally analysed as summarisation than as "quotation", much as, in ordinary writing practice, quotation marks signal a word-for-word extract, not a restatement.

[143] The distinction tracks Recital 57 CDSM and the idea/expression dichotomy: the reproduction right is concerned with the market for the work's expression, not with competition for the informational demand that unprotected facts satisfy. See further (n 43) and the Conclusion.

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[144] This distinction is vital for Article 15 CDSM. Because the PPR is a neighbouring right, it applies to publications regardless of their 'originality'. This creates a specific legal risk: if the concept of 'reproduction' were interpreted loosely to cover the mere overlap of facts, themes, or information, the PPR would effectively mutate into a monopoly over information itself, something EU law has consistently refused to grant. See *Infopaq* (n 58), paras 45-46; *Football Dataco* (n 59), para 42; *SAS Institute* (n 51), paras 39-41; see also again Recital 57 CDSM (stating as mentioned that the "rights granted to publishers of press publications should not extend […] to mere facts reported in press publications"). To prevent this, the test for 'reproduction' under Article 15 CDSM must remain anchored to the taking of recognisable wording or expression, not just the underlying facts. As discussed in Section 4, this alignment is essential to preserve the hierarchy of the *acquis*: if the PPR were allowed to capture informational overlap that copyright leaves free, the related right would become functionally broader and more restrictive than the author's primary right it was designed to complement.

[145] See *Like Company* referral, paras 11 and 22.

[146] Ibid, paras 4–5.

[147] See, *for example*, Recital 58 CDSM (on the limitations of Article 15).

[148] *Like Company* referral, para 7.

[149] Stamatoudi I., 'Originality under EU Copyright Law', in Torremans, P. (ed), *Copyright Law: A Handbook of Contemporary Research* (Elgar 2017), 57; Stamatoudi I., "More' Originality for Cypriot Copyright Law according to the CJEU's Case Law' (2018) 30 CR(1) 207.

[150] See *Infopaq* (n 58).

[151] Ibid, para 47.

[152] Even the *sui generis* right for databases requires that substantial investment (either in terms of quality or in terms of quantity) has taken place in either the obtaining, verification or presentation of the contents, as provided in Article 7 of the 1996 Directive 96/9/EC on the Legal Protection of Databases ("Database Directive"). Also, only substantial takings are prevented and not any taking unless it is done repeatedly and systematically. In the case of the PPR there is no need for investment despite the fact that the notion of investment is referred to in the reasoning of this provision. See Recital 58 CDSM: "[…] [T]he use of individual words or very short extracts of press publications by information society service providers may not undermine the investments made by publishers of press publications in the production of content".

[153] Stamatoudi, I., and Torremans, P., 'The Digital Single Market Directive', in Stamatoudi, I. and Torremans, P. (eds), *EU Copyright Law. A Commentary* (Elgar 2021) 651, and 727–728.

[154] Recital 58 CDSM.

[155] See *Meta* (n 46), para 59; Recital 57 CDSM.

[156] However, the structural question raised by the national court, *i.e.*, whether Article 3(2) "may exclude" Article 15, remains vital here. While Article 15 builds on the concepts of Articles 2 and 3 InfoSoc, it functions as a *lex specialis* that tailors these rights for the online press sector. This implies that where the specific conditions of Article 15 are not met (*e.g.*, because the use involves only very short extracts), a fallback claim under the general Article 3(2) InfoSoc right should not be available to re-capture the same conduct. To hold otherwise would render the Article 15 carve-outs nugatory.

[157] See (n 35).

[158] *Infopaq* (n 58). As acknowledged in the Evaluation Report on the Database Directive, "[t]here is a long-standing principle that copyright should not be extended to cover basic information or "raw" data". See Commission of the European Communities, DG Internal Market and Services Working Paper. First Evaluation of Directive 96/9/EC on the Legal Protection of Databases, 12 December 2005, p. 23. The principle has been applied by the CJEU with reference to the *sui generis* right in *The British Horseracing Board Ltd v William Hill Organization Ltd*, Case C-203/02, EU:C:2004:695; in general, see Karapapa, S., Defences to Copyright Infringement: Creativity, Innovation and Freedom on the Internet (Oxford University Press 2020), Chapter 2.

[159] See in this regard C-135/10, *Società Consortile Fonografici (SCF) v Marco Del Corso*, EU:C:2012:140, para 84; C-89/04 *Mediakabel BV v Commissariaat voor de Media*, EU:C:2005:348, para 30; C-192/04 *Lagardère Active Broadcast, v Société pour la perception de la rémunération équitable (SPRE)* et al., EU:C:2005:475, para 31; and C-306/05 *Sociedad General de Auditores y Editores de España (SGAE) v Rafael Hoteles SL* EU:C:2006:764, paras 37 and 38; see also in general Karapapa, S., 'The Requirement for A 'New Public' in EU Copyright Law' (2017) 1 *European* LR 63–81.

[160] C-466/12, *Nils Svensson and Others v Retriever Sverige AB*, EU:C:2014:76; C-279/13, *C More Entertainment AB v Linus Sandberg*, EU:C:2015:199; C-160/15, *GS Media BV v Sanoma Media Netherlands BV, Playboy Enterprises International*

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Inv., *Britt Geertruida Dekker*, EU:C:2016:644; C-527/15, *Stichting Brein v Jack Frederik Wullems (Stichting I)*, EU:C:2017:300. Relevant cases also include: Case C-360/13, *Public Relations Consultants Association Ltd v The Newspaper Licensing Agency Ltd*, EU:C:2014:1195 (on 'browsing'); and C-610/15, *Brein v Ziggo BV and XS4ALL Internet BV*, EU:C:2017:456; Stamatoudi, I., "Linking' and 'Browsing' in the Light of the EU Court of Justice's Recent Case Law', in Pistorius, T. (ed.), *IP Perspectives on the Regulation of New Technologies* (Elgar 2018), 179.

[161] Quintais, J.P., 'Untangling the Hyperlinking Web: In Search of the Online Right of Communication to the Public' (2018) 21 JWIP 385–420, at 398.

[162] See *Svensson*, *C More Entertainment*; *GS Media BV*; *Wullems (Stichting I)*; *Ziggo (Stichting II)* (n 160).

[163] *Getty Images vs. Stability AI* [2025] EWHC 2863 (Ch).

[164] See *GEMA* (n 52).

[165] See *Svensson*, *BestWater*, *GS Media* and *VG Bild-Kunst* (n 160).

[166] Ibid.

[167] See (n 45).

[168] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market, OJ 2000 L 178, Artt 12-15; Regulation (EU) No. 2022/2065 of 19 October 2022 on a Single Market For Digital Services (Digital Services Act) [2022] OJ L277/1, Articles 4-9.

[169] See *Ziggo* (n 160), paras 36–45 (on deliberate intervention; knowledge of infringement; substantial infringing uses and lack of neutrality; and inducement of infringement); Case C-392/19, *VG Bild-Kunst v Stiftung Preußischer Kulturbesitz*, EU:C:2021:181, paras 33–34 (on intentional intervention and control); *GS Media* (n 160), paras 34–51 (on the knowledge element).

[170] Joined Cases C-682/18 and C-683/18, *YouTube LLC v Cyando AG*, EU:C:2021:503, paras 102–115 (on criteria indicating active role/knowledge); Case C-324/09, *L'Oréal SA v eBay International AG*, EU:C:2011:474, paras 113–124 (clarifying that active role removes safe harbour); Joined Cases C-236/08 to C-238/08, *Google France and Google v Louis Vuitton Malletier and others*, EU:C:2010:159, paras 113–120 (on hosting when neutral).

[171] See, *for example*, Carlini and others (n 25) 2642–2646 (on rarity/duplication and attack prompts).

[172] See, *for example*, Lee, K. and others (n 27), pp. 8429–8431 (on order-of-magnitude reduction in verbatim emissions after deduplication).

[173] See, *for example*, Abadi and others, 'Deep Learning With Differential Privacy' in *Proceedings of CCS 2016* (ACM 2016), 309–312 (on DP-SGD limits per-example memorisation); Carlini and others (n 26), 7–9 (on scale, duplication and prompt length).

[174] See *Google France* (n 170), paras 113–114 (confirming that neutral, mere technical/automatic/passive role implies no knowledge/control); *Svensson* (n 160), paras 24–25, 29–31 (clarifying that the assessment does not turn on the site where the work resides; and that even where framing makes a work appear on the linker's site, the legal analysis is unchanged); *YouTube LLC* and *Cyando AG* (n 170) paras 68 and 90, 100–102 (noting that attribution hinges on an operator's deliberate intervention beyond merely making a service available); Opinion of AG Hogan in Case C-433/20, *Austro-Mechana v Strato*, EU:C:2022:217, paras 59–60, 69 (emphasising that who makes the copy is key; and that cloud providers can remain outside the "reproducer" role absent active involvement). See also Stamatoudi, I., Torremans, P.L.C. (eds), *EU Copyright Law: A Commentary* (Cheltenham, 2nd edn., Edward Elgar 2021), Art. 2 comment (noting that attribution in reproduction turns on the party that performs the act); Peukert, C., 'Copyright Levies and Cloud Storage: Ex-ante Policy Evaluation With A Field Experiment' (2024) 53 RP article 104114 (pointing out that *Austro-Mechana* treats cloud-server copies as private copies by users; and distinguishing hosting capacity from making the reproduction).

[175] See *Infopaq* (n 58), paras 50–51 and 66-70; Case C-360/13, *PRCA v NLA*, EU:C:2014:1195, paras 21–27 and 52–63.

[176] See, *for example*, Case C-70/10, *Scarlet Extended SA v SABAM*, EU:C:2011:771, paras 40–53; Case C-360/10, *SABAM v Netlog NV*, EU:C:2012:85, paras 34–51 (focusing on "no general monitoring", proportionality and fundamental rights).

[177] See, *for example*, Case C-516/17, *Spiegel Online GmbH* v *Volker Beck*, EU:C:2019:625, paras 76 and 94 (on conditions for quotation; contextual assessment, fair practice and extent required by the specific purpose); *Pelham* (n 50), paras 67, 69 and 71–74 (on definition and purpose of quotation; intention to enter into "dialogue" with the quoted work; and proportionality).

[178] See, *for example*, European Commission, Communication—Guidance on Article 17 of Directive (EU) 2019/790 on Copyright in the Digital Single Market, COM(2021) 288 final, Brussels, 4.6.2021, para 171 (noting that at the present

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

state of the art, automated tools cannot reliably recognise context-dependent lawful uses such as quotation, parody or pastiche, so ambiguous cases should not be blocked *ex ante*); Case C-401/19, *Republic of Poland* v *European Parliament and Council*, EU:C:2022:297, paras 80, 85, and 90–94 (noting that filtering measures must include robust safeguards to prevent blocking of lawful content and maintain a fair balance of rights; and that only manifestly unlawful content may be prevented automatically); *Scarlet Extended* (n 176), paras 36–40, 52; and *Netlog* (n 176), paras 34–38, 50 (both clarifying that general, preventive filtering obligations are incompatible with fundamental rights balancing and risk over-blocking lawful uses); see also Article 8 DSA (titled "No general monitoring or active fact-finding obligations").

[179] See Calabresi, G. and Melamed, A.D., 'Property Rules, Liability Rules, and Inalienability: One View of the Cathedral' (1972) 85 Harvard LR 1116–1123 (on choosing instruments under high transaction costs; and least-cost avoider logic).

[180] See (n 161).

[181] See, *for example*, Michael Heller, 'The Tragedy of the Anticommons: Property in the Transition from Marx to Markets' (1998) 111 Harvard LR 621. See also empirical discussion in the European Commission, Communication—*A European strategy for data*, COM(2020) 66 final, Brussels, 19.2.2020, pp. 2–11 (on documents concentration of data with a few Big Tech players; fragmentation and unequal conditions harming SMEs; and B2B frictions such as imbalanced bargaining power and legal uncertainty: all classic sources of overlapping claims/transaction costs that skew markets toward well-resourced incumbents).

[182] See Frosio, G., 'Schrödinger's Chatbot: Platform Liability, Volition, and Generative AI', SSRN research paper n. 7039658, July 2026, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7039658.

[183] Article 14 E-Commerce Directive.

[184] Article 8 DSA.

[185] See *Scarlet Extended* and *SABAM/Netlog* (n 176); *Republic of Poland* (n 178).

[186] See *L'Oréal* (n 170), paras 131–134, 139 and 143-144; Case C-18/18, *Glawischnig-Piesczek* v *Facebook Ireland*, EU:C:2019:821, paras 37, 41, 45-47 and 53 (on targeted injunctions; and "equivalent content" within limits consistent with Article 15 E-Commerce Directive).

[187] See Article 14 E-Commerce Directive; and DSA, especially Articles 14–17, 20–24, 34–35 (on due diligence; and systemic risks).

[188] See, *for example*, *L'Oréal* (n 170), paras 116–123; *YouTube/Cyando* (n 170), paras 107–115.

[189] See *YouTube/Cyando* (n 170), paras 93–96 (confirming that YouTube had taken "credible and effective measures"; and that its business model is not based on illegality).

[190] *Idem*, paras 84–87 (highlighting that relevant factors include whether the operator deploys appropriate measures and whether its financial model encourages unlawful sharing).

[191] See, *for example*, *Ziggo* (n 160) (deciding that providers lose the protection afforded to neutral intermediaries if they actively intervene to facilitate infringement, for instance, by knowingly indexing and managing illicit content).

[192] Articles 15(1)-(3) CDSM.

[193] See, *for example*, *YouTube/Cyando* (n 170), paras 71, 102, 106 and 117–118.

[194] See Article 15(2) and Recital 57 CDSM. Article 15 is expressly "without prejudice" to authors' and other rightholders' rights and may not be invoked against them, signalling that the PPR is meant to add protection rather than re-allocate value. A provider-primary, strict-liability reading can still economically harm authors in two concrete ways: (i) rent-shifting: platforms will pay to clear risk through publisher licences, so money that could otherwise flow to authors via author-right licensing (or remain available for lawful dissemination that supports authors' market value) is diverted to publishers unless it is shared; and (ii) over-blocking: because providers cannot know the user's purpose in advance (*e.g.*, quotation, criticism, reporting), they will block or heavily filter borderline uses, reducing lawful visibility, readership and downstream revenue opportunities for authors. The "prejudice" is therefore practical and monetary: authors may capture less value and lose audience-building uses, even though their formal rights remain intact).

[195] See DSA's Articles 16 and 17 (on notice-and-action; and statements of reasons); Article 20 (on repeat-infringer policy); Articles 34–35 (on risk assessment/mitigation for VLOPs/VLOSEs); see also Case C-314/12, *UPC Telekabel Wien GmbH v Constantin Film*, EU:C:2014:192, paras 62–63 (on proportionate, targeted injunctions).

[196] Where the excerpt is republished, the resulting item is hosted and persistent, so the ordinary Articles 16–17 DSA notice-and-action mechanism applies against the host in the usual way; the difficulty addressed here is specific to the unpublished, elicited output, which has no host and no persistence.

[197] See Articles 16–17, 34–35 DSA.

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[198] See Article 8(3) InfoSoc; *UPC Telekabel* (n 195), paras 56–63 (confirming that injunctions must be proportionate and avoid imposing general monitoring).

[199] See *YouTube/Cyando* (n 170), paras 102–115 (on diligent platforms not being primarily liable; and criteria for knowledge/active role).

[200] See *Microsoft* (n 96), describing output-side filters that flag and block content closely matching protected text; and Articles 16–17 DSA.

[201] See *UPC Telekabel* (n 195), paras 56–63.

[202] See (n 7) and the Conclusion.

[203] See *YouTube/Cyando* (n 170), paras 102–115

[204] See *Like Company* referral, para 26.

[205] See the remuneration scholarship cited (n 7).

How to cite this article: Bonadio, E., Frosio, G., Geiger, C., Guadamuz, A., Karapapa, S., & Stamatoudi, I. (2026). Preserving Balance in the EU Digital Single Market: How *Like Company* Could Reframe Copyright and Innovation in the Generative AI Era. *The Journal of World Intellectual Property*, 1–41. https://doi.org/10.1111/jwip.70032

17471796, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jwip.70032 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License