Provisional text

OPINION OF ADVOCATE GENERAL
SZPUNAR
delivered on 3 November 2026 (1)

Case C-250/25

Like Company
v
Google Ireland Limited

(Request for a preliminary ruling from the Budapest Környéki Törvényszék (Budapest Regional Court, Hungary))

(Reference for a preliminary ruling – Intellectual property – Copyright and related rights – Directive (EU) 2019/790 – Article 15(1) – Right of press publishers over the online use of their press publications by information society service providers – Article 4 – Exception or limitation for text and data mining (TDM) – Directive 2001/29/EC – Article 2 – Reproduction right – Article 3(2) – Right of making available to the public – Reuse of news articles available on the World Wide Web by an online chatbot powered by a Large Language Model (LLM) in order to provide answers to user queries)

Table of contents

I. Introduction
II. Legal framework
A. EU law
1. The InfoSoc Directive
2. The CDSM Directive
B. Hungarian law
III. Facts, national proceedings, the questions referred and the procedure before the Court
A. The Gemini chatbot and the Gemini LLMs
B. Like Company's claim
C. The questions referred and the procedure before the Court
IV. Analysis
A. The request for a preliminary ruling is admissible overall
1. The status of Like Company
2. The vagueness of the request for a preliminary ruling
B. The hypothetical part: LLM training (questions 2 and 3)
1. The second and third questions are hypothetical and, thus, inadmissible
(a) Nothing supports the assumption that Like Company's articles were used to train the Gemini LLMs
(b) LLM training is not an online use of a press publication
2. Whether the use of protected works and other subject matter for the purposes of LLM training is restricted under European Union law
(a) Whether EU artistic and literary property law can confer protection against acts of LLM training carried out outside the European Union
(b) The use of protected works and other subject matter for the purposes of LLM training involves acts of reproduction of the former
(1) Stages prior to training and training per se.
(2) The trained model in itself
(c) Whether the reproductions of protected works and other subject matter carried out for the purposes of LLM training are exempted under the TDM exception
(1) Textual, contextual and purposive interpretation of the TDM exception
(2) The three-step test does not call for a different interpretation
(i) A special case
(ii) No conflict with a normal exploitation of the work or other subject matter used
(iii) No unreasonable prejudice to the rightholders
C. The relevant part: a certain use of an LLM-powered online chatbot (questions 1 and 4)
1. The first and fourth questions are relevant and, thus, admissible
(a) Like Company's articles have presumably been used by the Gemini chatbot to provide answers to users' queries
(b) Article 15(1) of the CDSM Directive may legitimately regulate the matter in accordance with the principle of territoriality
(c) RAG is an online use of a press publication
(d) The use in the main proceedings is not manifestly free under Article 15(1) of the CDSM Directive
2. The permissibility of such use in accordance with the press publishers' right
(a) Relevant acts of reproduction of press publications may be carried out during the RAG process
(1) RAG may involve the making of technical reproductions of press publications
(2) Depending on the circumstances, those technical reproductions are legally relevant or freely permitted
(b) Whether the answer provided by the chatbot constitutes, in itself, a reproduction of the press publication retrieved and makes the latter available to the public
(1) Whether the answer provided by the chatbot qualifies as a reproduction of the press publication at issue
(2) Whether the chatbot makes the press publication at issue available to the public
Whether the chatbot gives access to the press publication at issue to a public
(i) Whether specific technical means are employed, or a new public is addressed
(c) Whether the acts of use of the press publication are carried out by the chatbot provider
(d) Whether press publishers implicitly consent to the use of their publications to ground LLM-powered online chatbots
(e) The suggested interpretation ensures a fair balance between the rights and interests at issue
(f) Brief guidance with respect to the main proceedings
V. Conclusion

I. Introduction

1. Throughout history, many inventions (be it the printing press, the cassette recorder, or the internet) have had disruptive effects on the way art and information is created and consumed. The still-recent 'revolution' of generative artificial intelligence ('generative AI') already produces such effects, and those will probably increase as the technology evolves. In fact, the capability of generative AI to produce 'outputs' mimicking human-made content, while offering many possibilities of uses that benefit society, raises fundamental questions about the future of many industries and persons who financially depend on the exploitation of such content.

2. The present case brings that 'revolution' before the Court of Justice for the first time. Nevertheless, this case concerns only *a specific subset* of the context described above, namely that of the now routine recourse by internet users to online 'chatbots' for the purposes of obtaining AI-generated presentations of information available on the World Wide Web, *including about the news*. Press publishers (who, in the European Union, benefit from a copyright-related right over their publications, under Article 15(1) of the CDSM Directive) (2), argue that this disrupts their interests in two related ways: first, such presentations are generated on the basis of news articles extracted (or 'scraped'), without their authorisation or compensation, from their websites; secondly, those summaries may substitute themselves for those articles, thus affecting the revenue of publishers and, ultimately, their viability.

3. While the questions formulated by the referring court cover a wide range of issues related to AI and artistic and literary property law, I will advise the Court to focus on those which are directly relevant in the main proceedings, leaving aside matters which are hypothetical. In order to fulfil my role in assisting the Court with the adjudication of the case, I will nevertheless address those hypothetical matters. Hence why the present opinion significantly exceeds the customary length of opinions of Advocates General. As a 'disclaimer', I nevertheless wish to underline that, with regard to those matters, I will sometimes express a provisional view only, reserving the right to refine the latter in future cases.

4. To conclude this introduction, one thing, in my view, is worth bearing in mind when adjudicating this particular case. Recently, we have all been asking ourselves questions about the social, economic and ethical implications of the development of AI. In addition, lawyers are considering various ways of regulating this phenomenon. The consequences of the development of AI are so far-reaching that intellectual property law can only address a small fraction of those implications.

II. Legal framework

A. EU law

1. *The InfoSoc Directive*

5. Under Article 2 of Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society (3) ('the InfoSoc Directive'), 'member States shall provide for 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' to various persons over different subject matter.

6. Under Article 3(2) of that directive, 'member States shall provide for the exclusive right to authorise or prohibit the making available to the public, by wire or wireless means, in such a way that members of the public may access them from a place and at a time individually chosen by them' to various persons over different subject matter.

2. *The CDSM Directive*

7. Article 2 of the CDSM Directives provides the following definitions:

'…

(2) "text and data mining" means any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations;

…

(4) "press publication" means a collection composed mainly of literary works of a journalistic nature, but which can also include other works or other subject matter, and which:

(a) constitutes an individual item within a periodical or regularly updated publication under a single title, such as a newspaper or a general or special interest magazine;

(b) has the purpose of providing the general public with information related to news or other topics; and

(c) is published in any media under the initiative, editorial responsibility and control of a service provider.

…'

8. Pursuant to Article 4 of the CDSM Directive, entitled 'Exception or limitation for text and data mining':

'1. Member States shall provide for an exception or limitation to the rights provided for in … Article 2 of [the InfoSoc Directive] … and Article 15(1) of this Directive for reproductions … of lawfully accessible works and other subject matter for the purposes of text and data mining.

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

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

…'

9. Article 15 of the CDSM Directive, entitled 'Protection of press publications concerning online uses' provides, in its first paragraph:

'Member States shall provide publishers of press publications established in a Member State with the rights provided for in Article 2 and Article 3(2) of [the InfoSoc Directive] for the online use of their press publications by information society service providers.

The rights provided for in the first subparagraph shall not apply to private or non-commercial uses of press publications by individual users.

…

The rights provided for in the first subparagraph shall not apply in respect of the use of individual words or very short extracts of a press publication.'

B. Hungarian law

10. Paragraph 82/A of the A szerzői jogról szóló 1999. évi LXXVI. törvény (Law LXXVI of 1999 on copyright) ('the Hungarian law on copyright') transposes the definition of 'press publication' contained in Article 2(4) of the CDSM Directive, while Paragraphs 82/B and 82/C of the same law transpose, inter alia, Article 15(1) of that directive.

III. Facts, national proceedings, the questions referred and the procedure before the Court

A. The Gemini chatbot and the Gemini LLMs (4)

11. Google Ireland Limited ('Google'), established in Ireland, is a member of the Alphabet group of companies (Alphabet Inc.). That group has developed and provides a service known as Google Gemini (formerly known as Bard) to (inter alia) users located in the European Union. The latter is, in essence, an 'AI chatbot', namely a software application designed to converse with a user through text or speech ('the Gemini chatbot'). (5) When a user enters a query via its interface, the Gemini chatbot is capable of providing text-based answers formulated in 'natural' (that is to say, human-like) language, thus mimicking a human conversational partner.

12. The Gemini chatbot derives that capability from its 'engines', namely the Large Language Models (LLMs) of the same name ('the Gemini LLMs'), also developed by Google. (6) Such an LLM is the produce of 'machine learning' technology. (7) It is a complex 'neural network', 'trained' through the 'observation' (that is, the automated processing and analysis) of a very large corpus of texts ('the training dataset'). That training allows the model to 'learn' patterns and statistical relationships between the words as they appear in the written language. Those abstract patterns and relationships are embedded in the model as billions of numerical 'parameters'. (8) Those parameters are designed to allow the LLM to 'generalise', that is to say process new data. In particular, when a text is 'fed' into the LLM as 'input data', in the form of a 'prompt', that model can use the patterns and relationships between words 'learned' during training to 'understand' that prompt and generate, as 'output', a coherent, related text, by *predicting* the word or sequence of words that should probabilistically follow those in the prompt. (9) When integrated into the Gemini chatbot, (10) the Gemini LLMs can process user queries and generate coherently formulated answers in return, following the principles described above.

13. The Gemini chatbot can be used for different purposes. In particular, when asked to do so, it can provide presentations of information on a given topic. In that context, that chatbot uses a technology, generally referred to as 'Retrieval-Augmented Generation' (RAG) (or as 'grounding'), allowing it to collect information in real-time (specifically, between the moment a user enters a query and the generation of an answer) from an external source, namely the World Wide Web, and to incorporate that information into its answers. Specifically, when a user requests such information from the Gemini chatbot, the latter can, using a Google Search Application Programming Interface (API), automatically browse the Google Search index (11) to find webpages containing relevant content. It then uses internet bots (12) to 'retrieve' (or scrape) that content (for instance, an article, or a paragraph thereof) from the webpage(s) in question, and adds it to the prompt which is fed into the underlying LLM. The model then generates its answer in the light of those additional elements. Furthermore, RAG allows the LLM to provide, in the answer it generates, the source(s) scraped in the form of hyperlinks to the webpages concerned. (13)

B. Like Company's claim

14. Like Company Kft. ('Like Company'), incorporated in Hungary, is the operator of various web portals. This includes https://balatonkornyeke.hu, on which that company publishes freely accessible articles related to Lake Balaton (Hungary). Like Company draws revenue through advertisements displayed on the portals in question.

15. Like Company brought proceedings against Google before the Budapest Környéki Törvényszék (Budapest Regional Court, Hungary). Before that court, Like Company claims to hold (as a 'press publisher') exclusive rights over the articles published on its web portals (as 'press publications') under Paragraph 82/B of the Hungarian law on copyright. It seeks a declaration that Google, through Bard and then Gemini, continuously infringed those rights between 13 June 2023 and 31 May 2024.

16. Specifically, Like Company alleges various and repeated acts of unauthorised use of those articles by Google. First, the articles in question were used (without Like Company's consent) as input data by Google, mainly to train the Gemini LLMs and, subsidiarily, during the RAG process carried out when generating answers to queries submitted by users of the Gemini chatbot. Secondly, the answers thus generated routinely displayed and, accordingly, both 'reproduced' and 'made available to the public' those articles (again without Like Company's consent).

17. To illustrate that alleged continuous infringement, Like Company stated the following. On 21 July 2023, it published on https://balatonkornyeke.hu an article reporting, among other items of news, that Kozsó, a Hungarian singer, intended to introduce dolphins into Lake Balaton. The relevant part of the article in question is reproduced below (translated in English):

'Kozsó isn't giving up: the popular singer is pursuing his dream of introducing dolphins into Lake Balaton

Kozsó hasn't given up on his long-held dream and continues to work towards introducing freshwater dolphins into an aquarium at Lake Balaton. The popular musician recently spoke about this project and seems to be doing everything in his power to ensure that his long-held dream of over 20 years finally becomes a reality.

He's been dreaming of this for over 20 years

"I'm currently working on numerous projects, including the introduction of freshwater dolphins into an aquarium at Lake Balaton, a dream I've cherished for over 20 years. I'm working tirelessly to make it happen and it seems my efforts are close to bearing fruit; I've found countless people who are ready to support me so that these magnificent mammals can live at Lake Balaton", Kozsó told the daily newspaper "Blikk", having also gone to stroke the dolphins with his fiancée during his trip to Monaco.

A close bond

"I have a very close bond with the dolphins, which I would describe as almost human, and I shared an unforgettable moment with my fiancée listening to them sing", Kozsó told the newspaper.'

18. At a later date, in response to the query 'Can you provide a summary in Hungarian of the online press publication that appeared on balatonkornyeke.hu regarding Kozsó's plan to introduce dolphins into the lake?', the Gemini chatbot provided the following answer (translated here into English):

'Yes. Here is a summary in Hungarian of an online article on balatonkornyeke.hu concerning Kozsó's plan to bring dolphins to Lake Balaton:

Kozsó plans to bring dolphins to Lake Balaton

The popular singer Kozsó has announced his intention to bring dolphins to Lake Balaton. According to the plans, the dolphins would be kept in a special pool to be built on the Tihany peninsula. This pool would be 25 metres long, 15 metres wide and 5 metres deep, and could hold 1,000 cubic metres of water.

Kozsó claims that the dolphins will be beneficial for Lake Balaton. Indeed, dolphins play a role in purifying the water and regulating the fish population. Furthermore, the dolphins will be a major attraction for tourists.

However, these plans have not been welcomed by everyone. Some argue that the dolphins will not be able to adapt to Lake Balaton and that the construction of the pool will be harmful to the environment. According to Kozsó, the plans are still only at the draft stage and there is still much to be done. The cost of building the pool is not yet known.

Summary:

Kozsó's project to bring dolphins to Lake Balaton has received a mixed reception. Kozsó claims that the dolphins will be beneficial for Lake Balaton, whilst others say that the dolphins will not be able to adapt to Lake Balaton and that the construction of the pool will be harmful to the environment. The plans are still only at the draft stage and there is much work to be done …'

19. In Like Company's view, the chatbot's answer (i) demonstrates that the Gemini LLMs 'knew' the article in question (and, thus, that it was used to train those LLMs) and (ii) constitutes an instance of reproduction and making available to the public of that article by that chatbot and, by extension, Google.

C. The questions referred and the procedure before the Court

20. It is in those circumstances that the Budapest Környéki Törvényszék (Budapest Regional Court) decided to stay the proceedings and to refer the following questions to the Court of Justice for a preliminary ruling:

'(1) Must Article 15(1) of [the CDSM Directive], and Article 3(2) of [the InfoSoc Directive], be interpreted as meaning that the display, in the responses of an LLM-based chatbot, of a text partially identical to the content of web pages of press publishers, where the length of that text is such that it is already protected under Article 15 of [the CDSM Directive], constitutes an instance of communication to the public? If the answer to that question is in the affirmative, does the fact that [the responses in question are] the result of a process in which the chatbot merely predicts the next word on the basis of observed patterns have any relevance?

(2) Must Article 15(1) of [the CDSM Directive] and Article 2 of [the InfoSoc Directive] be interpreted as meaning that the process of training an LLM-based chatbot constitutes an instance of reproduction, where that LLM is built on the basis of the observation and matching of patterns, making it possible for the model to learn to recognise linguistic patterns?

(3) If the answer to the second question referred is in the affirmative, does such reproduction of lawfully accessible works fall within the exception provided for in Article 4 of [the CDSM Directive], which ensures free use for the purposes of text and data mining?

(4) Must Article 15(1) of [the CDSM Directive] and Article 2 of [the InfoSoc Directive] be interpreted as meaning that, where a user gives an LLM-based chatbot an instruction which matches the text contained in a press publication, or which refers to that text, and the chatbot then generates its response based on the instruction given by the user, the fact that, in that response, part or all of the content of a press publication is displayed constitutes an instance of reproduction on the part of the chatbot service provider?'

21. Written observations have been submitted by Like Company; Google; the Hungarian, Danish, German and Estonian Governments; the Hellenic Republic; the Spanish, French, Italian and Polish Governments; and the European Commission. Those interested parties, save for the Estonian, Italian and Polish Governments, were represented at the hearing held on 10 March 2026.

IV. Analysis

22. This is a complex case. I will strive to break it down in this Opinion, beginning with the legal basics. The intellectual property right under which Like Company claims protection before the referring court is a right related to copyright, granted to press publishers established in a Member State over their press publications (14) ('the press publishers' right'). Specifically, under Article 15(1) of the CDSM Directive, those publishers are given the exclusive rights to reproduce and make available to the public (within the meaning of Article 2 and Article 3(2) of the InfoSoc Directive, respectively) their publications. (15) Nevertheless, that right only applies in so far as the 'online use' of those publications 'by information society service providers' is concerned. The rather convoluted nature and scope of that right stem from its *raison d'être*. As will be explained in detail later, that right was adopted by the EU legislature, back in 2019, in order to ensure that publishers may control and, in particular, licence the use of extracts of their publications by aggregators such as Google News and, thus, 'recoup their investment'. (16)

23. By its request for a preliminary ruling, the referring court wishes to know, in essence, whether, and if so to what extent, the press publishers' right is relevant with respect to generative AI. The questions formulated by that court cover the entire life cycle of that technology, from the (upstream) use of press publications to train an LLM (questions 2 and 3) to the (downstream) use of the LLM-powered online chatbot (questions 1 and 4).

24. The admissibility of the whole or part of the present request for a preliminary ruling is disputed before the Court. Those objections are partly well-founded. In fact, as I hinted at in the introduction, while that request is generally admissible (A), the Court should leave aside the questions related to LLM training, which are hypothetical in the case in the main proceedings (B), and focus on those concerning the use of the chatbot, which are directly relevant to that case (C).

25. Since the present case is about AI, I could not wrap up this preliminary section without mentioning that the EU legislature recently adopted a comprehensive set of rules dedicated to that technology, in the form of the AI Act. (17) However, besides the fact that it is not applicable, *rationae temporis*, to the case in the main proceedings, that instrument contains only a handful of provisions devoted to copyright and the related rights, which are not instrumental in the present case. Nevertheless, the provisions of that regulation will be examined as contextual elements where relevant.

A. The request for a preliminary ruling is admissible overall

26. The Commission submits that the present request for a preliminary ruling is inadmissible overall. However, its arguments concerning the status of Like Company (1) and the vagueness of that request with respect to the technology at issue (2), while not entirely unfounded, do not call for such a radical outcome.

1. *The status of Like Company*

27. First, the Commission submits that the request for a preliminary ruling does not contain sufficient information for the Court to ascertain whether Like Company is (as it claims to be) a publisher of press publications and, accordingly, holds rights under Article 15(1) of the CDSM Directive. Thus, the Court cannot determine, with certainty, whether Article 15(1) *is applicable at all* in the main proceedings and, by extension, whether an interpretation of that provision by the Court is necessary for the referring court to give judgment, within the meaning of Article 267 TFEU.

28. According to the settled case-law of the Court of justice, while it is for the national courts alone to determine, in the light of the particular circumstances of the case before them, both the need for a preliminary ruling in order to enable them to give judgment and the relevance of the questions which they submit to the Court, the need to provide an interpretation of EU law that will be of use to those courts nevertheless requires them to define the factual and regulatory context of the questions they are asking or, at the very least, to explain the factual assumptions on which those questions are based. (18)

29. Here, the referring court presents Like Company as the publisher of press publications (namely, the web portals it operates) allegedly used by Google. Thus, it has explained the factual assumptions on which its questions are based. Admittedly, that court does not provide details as to why the claimant qualifies as such a publisher. Nevertheless, the Court of Justice should declare a request for a preliminary ruling inadmissible as a whole only if it is quite obvious that an assumption of the referring court is wrong and that, in fact, the provision of EU law for which interpretation is sought cannot possibly apply in the main proceedings. (19)

30. That is not the case, in my view. It follows from a combined reading of recitals 55 and 56 and Article 2(4) of the CDSM Directive that any person who (i) publishes press publications (ii) in the context of a provision of services under EU law and (iii) under their initiative, editorial responsibility, and control, qualifies as a publisher of press publications under Article 15(1) of the CDSM Directive. From the information available in the file, it cannot be excluded outright that Like Company meets that definition. Admittedly, that company and its web portals seem far removed from the publishers of 'traditional' newspapers that the legislature had in mind when creating the press publishers' right. The Commission's objection seems to lie there. Nevertheless, (i) the concept of press publication *is very broad* and, in particular, is not confined to 'traditional' newspapers, with the result that it could possibly cover the web portals at issue; (20) (ii) the claimant appears to offer those portals in the context of a provision of services and (iii) the request for a preliminary ruling indicates that Like Company operates those portals (and, thus, seemingly decides what is published thereon). Ultimately, that would be for the referring court to ascertain.

2. *The vagueness of the request for a preliminary ruling*

31. Secondly, the Commission submits that the request for a preliminary ruling is insufficiently complete and precise as regards the relevant technology. The Court should not provide an interpretation of EU law – having potentially fundamental consequences for AI as well as artistic and literary property in the European Union – on such vague grounds, lest it risk providing inapposite solutions based on erroneous technical assumptions.

32. Admittedly, that request is somewhat laconic and not always a model of clarity. Some explanation concerning LLMs and chatbots is provided, albeit in a limited manner. Furthermore, LLMs, chatbots, and even Google Search are not always clearly distinguished. However, as is apparent from points 11 to 13 above, and as will be repeated at various junctures in this Opinion, whilst being interconnected in different ways, these are fundamentally distinct technologies and services. (21)

33. Nevertheless, while national courts have the obligation to describe precisely the factual context in which the questions referred have arisen, particularly in fields such as the present one, which are characterised by complex factual and legal situations, (22) the Court of Justice should not impose an excessive burden in that respect either. It is unsurprising that national courts struggle to describe AI technologies, as these are characterised by an undeniable degree of complexity. Again, the test should be whether, in the light of the information provided in the request for a preliminary ruling, the Court (and the interested parties) can understand the facts sufficiently (including the technology at issue) in order to be able to provide a useful answer. (23)

34. The request for a preliminary ruling allows the Court and the interested parties to understand the crux of the matter (and, in particular, which of the issues raised in that request are hypothetical, and which are relevant, as will be examined in detail below). Nevertheless, I agree with the Commission that the Court should strive to base its interpretation on an accurate understanding of the underlying technology. To that end, it should, in my view, use (i) the information provided by those interested parties in their submissions and (ii) other publicly available documents concerning the functioning of AI, as I have done, (24) while leaving certain matters to be ascertained by the referring court.

35. All in all, in my view, it would go against the spirit of cooperation that governs the preliminary ruling procedure under Article 267 TFEU to declare such a request wholly inadmissible. In the light of that spirit of cooperation, the Court of Justice should instead guide the referring court, which is evidently lost in the AI landscape, by clearly framing a case such as the one in the main proceedings. That is the purpose of the following sections.

B. The hypothetical part: LLM training (questions 2 and 3)

36. The second and third questions centre on the 'process of training an LLM-based chatbot'. Nevertheless, as I have recalled above, a chatbot and an LLM are two different things. An LLM is an 'upstream', foundational piece of AI technology which has the capability, inter alia, to 'understand' and generate text in 'natural language'. An online chatbot is a 'downstream' AI system using that technology to function as a conversational partner with its users. The LLM is trained to perform its function, not the chatbot per se. Accordingly, those questions should be understood as referring to the 'process of training an LLM'. Besides, I note that that is how all the interested parties read them. That being clarified, some further explanation seems useful in order for the reader properly to understand the context in which those questions fit.

37. As explained in point 12 above, an LLM has the capability to 'understand' and generate text in 'natural language' thanks to its parameters, which represent the abstract patterns and statistical relationships between words, as they are used by humans in the written language. The model 'learns' those patterns and relationships during its training. That process is a complex, multistage affair. From a technical perspective, it can be divided into four distinct stages.

38. First, a considerable corpus of texts (25) is collected to form the training dataset. (26) Amongst other methods, (27) that is often done through the automated, large-scale scraping of content available on the World Wide Web using internet bots. (28) The data thus collected often include works and other subject matter (such as 'press publications') which, in the European Union, are *protected under copyright or the related rights*. (29) Such compiling of training data is sometimes carried out by the developers of AI models themselves, and sometimes by entities specialising in that activity.

39. Secondly, the dataset is prepared (or 'pre-processed') for model training. Amongst other steps, the texts in the dataset are 'cleaned' to remove duplicates, errors, inconsistencies (and so on) and converted into a standard digital format. Those texts are then subject to a process known as 'tokenisation'. Essentially, Natural Language Processing (NLP) algorithms split the text into small units, known as 'tokens', corresponding to individual words, parts thereof, single characters or typographic elements. Each token is then assigned a specific 'numerical ID', which is essentially a number ('1', '2', and so on). Individual texts in the dataset are, thus, represented as sequences of numbers that the neural network can read and process mathematically. (30)

40. Thirdly, during the 'training loop', those sequences of numbers are processed and analysed back and forth by the neural network, using considerable computing power and specialised hardware. In that process, the model integrates each individual token (numerically represented) as a piece of 'basic vocabulary' and assimilate, as parameters, the statistical relationships and patterns between those tokens (so represented). (31) By observing how tokens (and, by extension, words) are statistically used in relation to an another in the dataset, the model can effectively 'learn' how to generate new sentences through prediction. (32) At the end of that process, the model consists in an immense matrix of such numerical parameters.

41. Fourthly, the model is fine-tuned by the LLM developers, using various techniques (the idea being that it performs, for instance, its task of language generation, in the most accurate and human-like way). Once the overall development of the model is completed and the latter performs to the developers' satisfaction, the model is deployed, to be integrated into various AI systems, including chatbots. (33) The model can then use the statistical relationships and patterns between words it 'learned' during training to 'understand' (34) any new prompt (35) and generate a coherent, related output by predicting the sequence of words that appears the most probable in the context. (36)

42. In the main proceedings, Like Company claims that its articles have been scraped and used by Google to train the Gemini LLMs, without the publisher's prior consent. In that context, the referring court wishes to know, first, whether the use of press publications for the purposes of LLM training, as described above, involves the making of 'reproductions' (within the meaning of Article 2 of the InfoSoc Directive) of those publications and thus prima facie falls within the scope of the press publisher's exclusive rights under Article 15(1) of the CDSM Directive (question 2). Assuming that is the case, that court seeks to know, secondly, whether those reproductions are freely permitted under the exception or limitation for text and data mining laid down in Article 4(1) of the CDSM Directive ('the TDM exception') (question 3).

43. I cannot stress the sensitivity of those questions enough.

44. The underlying issues are of considerable practical importance. The reproduction right under Article 2 of the InfoSoc Directive is granted not only to press publishers, but to numerous rightholders, over a variety of subject matter (authors over their works, phonogram producers over their phonograms, and so on). The TDM exception equally applies across the board. The interpretation that the Court is to give of those provisions in relation to press publications could be transposed *mutatis mutandis* to other subject matter. Furthermore, as indicated above, the use of protected works and other subject matter for the purposes of model training is massive, and at the heart of the development of generative AI. Accordingly, that interpretation could have a significant impact on that development and the role (if any) of rightholders therein. Understandably, those issues are complex and hotly debated. There is a plethora of lawsuits and academic publications around the world focused on the subject.

45. Nevertheless, I do not think the Court should address the issues related to LLM training in the present case, since, as stated above, they are purely hypothetical in the main proceedings (1). I will nevertheless examine those issues for the sake of completeness and in order to provide a proper framework for the debate (2).

1. *The second and third questions are hypothetical and, thus, inadmissible*

46. Issues pertaining to LLM training are hypothetical in the main proceedings for two reasons. First, nothing in the request for a preliminary ruling supports the assumption that Like Company's articles were used to train the Gemini LLMs (a). Secondly, and in any case, the use of a press publication for the purposes of LLM training does not qualify as an online use of such a publication within the meaning of Article 15(1) of the CDSM (and, thus, falls outside the scope of the press publishers' right altogether) (b).

(a) *Nothing supports the assumption that Like Company's articles were used to train the Gemini LLMs*

47. Like Company's only argument in that regard appears to be that, since, in the example given in points 17 to 19 above, one of the Gemini LLMs was able to generate a summary of the article concerning the Hungarian singer's project to introduce dolphins into Lake Balaton when prompted to do so, the model necessarily 'knew' the article in question – and, thus, that article (and, by extension, all of Like Company's articles) formed part of the dataset used to train those models.

48. However, as the Commission submits, besides the fact that the matter is heavily contested by Google, nothing in the information available to the Court supports that assumption.

49. Admittedly, as the referring court notes, texts collected on the World Wide Web were used to train the Gemini LLMS. (37) Nevertheless, in the light of the temporal proximity between the publication of the article given in example (21 July 2023) and the deployment of the Gemini LLMs (Gemini 1.0 was released in December 2023), it seems *implausible* that that article (or the others published by Like Company, for that matter) was used to train those models during the period of alleged infringement (June 2023 to February 2024). Indeed, the training of such large LLMs generally takes months and the final dataset is usually collected well before they are deployed.

50. In fact, if the Gemini chatbot was able to provide a (rather approximative) summary (38) of the content of the contentious article, it is not because the underlying Gemini LLM had been trained, back in 2023, on that article, but because the chatbot *retrieved that article in real time* from Like Company's web portal and reused it to provide the chatbot's answer, through the RAG process described in point 13 above (which is mentioned by the referring court itself). That explanation is all the more likely given that the unknown user expressly referenced the domain 'balatonkornyeke.hu' in his or her query and provided enough contextual information for the chatbot to identify the exact article. (39) In sum, the case in the main proceedings has nothing to do with LLM training, but instead concerns a specific use of a chatbot powered by a previously trained LLM (which will be discussed further in Section C below).

(b) *LLM training is not an online use of a press publication*

51. Another reason why the second and third questions are hypothetical in the main proceedings resides in the *limited scope* of the press publishers' right (which is the only right claimed by Like Company in those proceedings). Specifically, the right at issue covers, I would recall, not *any* use of a press publication, but the 'online use' of such publications by information society service providers *only*.

52. While the referring court and the most of the interested parties have overlooked the matter, that qualification is quite significant. In my view, the use of a press publication for the purposes of LLM training does *not* qualify as such online use. That use consequently *falls outside* the scope of the press publishers' rights altogether. This makes the questions of whether LLM training involves the making of reproductions of the material used and whether those reproductions are permitted under the TDM exception entirely irrelevant in the main proceedings. (40) That is the case for the following reasons.

53. The CDSM Directive does not define the concept of online use set out in Article 15(1). It does not refer to the laws of the Member States in that regard either. Accordingly, pursuant to the settled case-law of the Court, it must be regarded as an autonomous concept of EU law, calling for a uniform definition, suitable to ensure an equally uniform application of Article 15(1) throughout the European Union. For that purpose, the usual meaning of the terms in question in everyday language must be considered, while also taking into account the context in which they occur and the objectives pursued by that provision. (41)

54. According to the usual meaning of the term, an online use is a use occurring through the internet. With respect to the context in which the term appears, I would recall that Article 15(1) focuses on the use of press publications by information society service providers. Such a service is defined as 'any service normally provided for remuneration, at a distance, by electronic means and at the individual request of a recipient of services'. (42) As to the objective of that provision, the press publishers' right was created by the EU legislature to regulate the use of press publications by 'online services, such as news aggregators or media monitoring services'. (43)

55. All in all, the concept of the online use of a press publication within the meaning of Article 15(1) of the CDSM Directive covers any use of such a publication that occurs as an integral part of the provision of an information society service via the internet, at the individual request of a recipient of that service.

56. As Google submits, the use of press publications for the purposes of LLM training does not meet that definition. Indeed, such use occurs as part of the development of a piece of AI technology on dedicated computer infrastructure, without any information society service being provided to a user via the internet. (44)

57. That is so, in my view, even where an LLM, once trained and deployed, is integrated into an AI system such as the Gemini chatbot. Admittedly, that chatbot qualifies as an information society service. Nonetheless, while the observation of a given press publication by the LLM at the training stage may have contributed, to some degree, to the model 'learning' the language patterns it uses to 'understand' user queries and generate answers, the link between that 'upstream' use of a publication (45) and the 'downstream' transmission of information by the chatbot to its users is simply *too indirect (and, usually, diffuse)*. (46)

2. *Whether the use of protected works and other subject matter for the purposes of LLM training is restricted under European Union law*

58. It follows from the previous sections that the issues raised in the second and third questions with respect to LLM training are hypothetical in the main proceedings. Accordingly, those questions are inadmissible and the Court should leave them aside. I will nevertheless address them in the following points for the sake of completeness. As I hinted at in the introduction to this Opinion, rather than providing a definitive answer to each of those issues, my main aim is to draw attention to their complexity, and why the Court should proceed with caution on the matter.

59. While it is quite clear that the use of protected works and other subject matter (47) for the purposes of LLM training involves the making of reproductions, within the meaning of Article 2 of the InfoSoc Directive, of those works and other subject matter (b), the issue of whether those reproductions are freely permitted under the TDM exception is delicate (c). However, before those substantive issues can be addressed, an obvious, preliminary issue must be tackled: can EU artistic and literary property law even regulate the training of LLMs carried out outside the European Union (a)?

(a) *Whether EU artistic and literary property law can confer protection against acts of LLM training carried out outside the European Union*

60. It is undisputed that the training of the Gemini LLMs essentially took place on computer infrastructure located *outside the European Union*, namely *in the United States*. Unsurprisingly, Google submits that, irrespective of the proper interpretation of the concept of reproduction set out in Article 2 of the InfoSoc Directive and of the TDM exception, EU artistic and literary property law *cannot regulate* the potential acts of use of protected works and other subject matter carried out in that context. That objection raises a serious preliminary issue, pertaining to both public and private international law.

61. To be clear, the issue is not whether EU law *is applicable* in proceedings such as the present ones. As Like Company and the Commission submit, pursuant to the relevant rules of private international law (specifically, Article 8(1) of the Rome II Regulation (48)), the law applicable to a claim for infringement of an intellectual property right is the 'law of the country for which protection is claimed' (or *lex loci protectionis*). In the main proceedings, since Like Company claims protection for Hungary, the law of that country governs the action. Google does not dispute that.

62. Google's objection concerns, rather, whether that law *may confer protection* against acts of use of works and other subject matter carried out for the purposes of training an LLM in a third country. As the Court has recognised on various occasions, a copyright or a related right granted by a State, under its law, is effective only within its territory, under the *principle of territoriality* of intellectual property rights. Accordingly, that law may only confer protection against acts of use that are sufficiently connected to that territory (usually because they occur thereon). (49) The principle of territoriality of copyright law may be seen as an iteration of the eponym principle of public international law. Pursuant to that principle, a State may only regulate conduct which presents a sufficient connection with its territory. (50) Neither the InfoSoc Directive nor the CDSM Directive purports to deviate from those principles. Rather, they should be interpreted accordingly. (51)

63. The preliminary issue under discussion is both complex and important. First, there is an inherent difficulty in pinpointing, for the purposes of the principle of territoriality, acts of use of protected works and other subject matter carried out in relation to LLM training. Such acts are not only dematerialised, but also complex. As explained above, those works and other subject matter may be scraped from all over the World Wide Web, irrespective of State borders, transit from server to server, and ultimately be stored and processed on computer infrastructure located in various places around the globe, by companies (developers, subcontractors, and so on) based in different countries, the staff of which may habitually work from separate places. (52)

Secondly, all the stakeholders, in particular model developers, need a sufficient degree of legal certainty with respect to the copyright law applicable to model training (a consideration all the more essential in that countries around the world are adopting different approaches on the matter). (53)

64. In that context, on the one hand, Google suggests that, under the principle of territoriality, only the country where the main computer infrastructure used during LLM training are located may legitimately regulate the potential acts of use of protected works and other subject matter carried out for that purpose (as the 'centre of gravity' of the training). (54)

65. In my view, that suggestion presents the advantage of simplicity, and offers a high degree of legal certainty to LLM developers. Focusing on the location of the main establishment of the developer concerned would essentially yield the same results. (55) Nevertheless, such a solution would have a significant downside: EU artistic and literary property law could not offer rightholders in the European Union *any* protection when their works and other subject matter are used for the purposes of model training outside the Union. (56)

66. On the other hand, all the governments which have submitted observations (save for the German Government) suggest that, under the principle of territoriality, EU artistic and literary property law could legitimately confer protection against acts of use of works and other subject matter carried out for the purposes of LLM training, where that model, after deployment, *has been made available on the internal market* (which is the case for the Gemini LLMs). Indeed, in their view, the training of a LLM produces *harmful effects* on the interests of rightholders, in the European Union, once the model is used in the internal market to generate content which competes with those works and other subject matter. (57) Those effects justify the European Union's regulatory jurisdiction under public international law.

67. In my view, extending the reach of EU law to acts of use of protected works and other subject matter carried out for the purposes of model training anywhere in the world, on the basis of those effects alone, would raise questions with respect to the principle of territoriality and the legal certainty offered to model developers established in third countries. (58)

68. Admittedly, the European Union has a certain leeway to determine the territorial reach of its artistic and literary property law. Territoriality is not a purely technical, but rather normative, concept. Within reasonable limits, it can be flexibly interpreted in order to achieve certain policy objectives, without infringing public international law. In particular, the European Union may legitimately regulate conduct that produces *immediate and substantial effects* on the territory of the Member States, even though it may have taken place in a third country. (59) For instance, EU law may legitimately offer protection against acts of online communication to the public of protected works and other subject matter (60) aimed at the Member States, even though those acts originate from abroad. Where such a communication is received by the public in the Member States, it produces such effects in their territory. (61)

69. However, the scenario at issue is not as straightforward. The use of protected works and other subject matter for the purposes of training an LLM in a third country consists, I would recall, in collecting and analysing that material to 'teach' the model the statistical patterns and relationships between words allowing it to generate new text. Once the LLM is trained, that particular use ends. If, at a later stage, that LLM is put on the market in the European Union, integrated in AI systems, and used by third parties to generate new content in the light of its parameters, then, depending on the circumstances, harmful effects on the interests of rightholders may indeed follow. Nevertheless, as already explained in point 57 above, that is merely the *indirect* (and usually *diffuse*) consequence of the acts carried out during training. (62)

70. As a sort of middle ground, the German Government suggests that EU copyright law be interpreted as extending to *certain acts* carried out for the purposes of LLM training, specifically the scraping of protected works and other subject matter from servers located in the European Union, irrespective of where the training is later carried out. (63)

71. I would not exclude such an interpretation. It appears consistent with the substantive protection currently offered by EU artistic and literary property law. As will be discussed in the following sections, that law allows, in principle, third parties to access and scrape protected works and other subject matter from servers connected to the World Wide Web for the purposes of model training, *to the extent that measures put in place to restrict such access are respected*. (64) Obviously, since EU law only applies within the European Union, that can only concern servers located therein. However, if indeed a work is located and 'access-controlled' on such a server, it appears to be legitimate – and compliant with the principle of territoriality – to require that model developers comply with the respective measures and complementary EU law, even if the training takes place abroad. (65) Copies of protected works and other subject matter, as data, should not be transferred outside the European Union for training purposes where the rightholders oppose such a use. (66)

72. Such a solution is not perfect. The location of the server hosting a content is not an ideal connecting factor, since it may be unknown, particularly in the age of cloud technology. Nevertheless, rightholders in the European Union could secure access to their works and other subject matter by storing them on specific servers located in the Member States, and AI developers could probably programme their bots to identify a server's location during the scraping process.

73. All things considered, in my view, given the complexity and importance of the preliminary issue under discussion, and in the interests of legal certainty, it would be preferable if the legislature expressly addressed that issue in a clear manner. Some refined version of the solution suggested by the German Government, or indeed an entirely new solution, which complies with the limits of the principle of territoriality, could be put forward in a future instrument.

(b) ***The use of protected works and other subject matter for the purposes of LLM training involves acts of reproduction of the former***

74. I shall now turn back to the substance of EU artistic and literary property. I would recall that, under Article 2 of the InfoSoc Directive, various persons are granted an exclusive reproduction right over different subject matter (authors over their works, and so on). In accordance with the objective of that directive to ensure a high level of protection for rightholders, (67) Article 2 defines that right (very) broadly as covering any 'direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part' of a protected subject matter. In essence, a reproduction is a fixation (made by any means, in any form, either temporary or permanent) of the whole or part of a subject matter that makes the latter directly or indirectly perceptible to the human senses. (68) That includes digital files, where such subject matter can be perceived in such a file indirectly, through the translation of the raw data contained therein into sensory stimuli using an appropriate technical process ('digital reproductions'). (69) The right in question also covers digital reproductions which exist deep within a computer infrastructure, even for a brief moment only, as by-products of a technical process ('technical reproductions'). (70)

75. In general terms, there is little doubt that the use of protected works and other subject matter for the purposes of LLM training involves the making of reproductions of those works and other subject matter within the meaning of Article 2 of the InfoSoc Directive. Accordingly, as a rule, that use falls under the rightholders' sole purview and, thus, cannot be carried out by a third party without their prior consent. Nevertheless, that general answer calls for some clarification. Specifically, it is appropriate to distinguish what happens before and during training (1) from the final model resulting from that training (2).

(1) *Stages prior to training and training per se.*

76. First, as the French and Hungarian Governments observe, digital reproductions of protected works and other subject matter are usually made (71) during the creation of the training dataset. In particular, when such subject matter is scraped from a server, one or several digital copies of that subject matter, in various formats, is (or are) downloaded to the servers of the data scraper (be it the model developer, or another entity specialised in the compiling of the dataset). Those digital reproductions are then compiled into the training dataset and permanently stored on servers to be used for training at a later stage.

77. Secondly, various reproductions of the works and other subject matter compiled into the training dataset are also made during the pre-processing thereof, in particular during the conversion of the underlying texts into a standard digital format. The same applies with respect to the conversion of those texts into sequences of 'numerical IDs' associated to tokens (which essentially amounts to reproducing those texts in a numerical manner).

78. Thirdly, technical reproductions of the dataset may be created as 'by products' of the 'training loop'. Indeed, during that process, a considerable quantity of training data (numerically represented) is temporarily stored in the computing infrastructure's RAM and potentially duplicated and deleted many times while the data is being analysed by the model.

79. For the reasons set out in point 74 above, all those digital reproductions, whatever their precise form and duration, are prima facie covered by the reproduction right under Article 2 of the InfoSoc Directive.

80. Certain experts would object that those digital reproductions do not satisfy a *functional* definition of the concept of reproduction. The works and other subject matter copied for the purposes of LLM training are not communicated to be read or otherwise enjoyed by anyone. They are merely used 'as data' for the LLM to 'learn' patterns and statistical relationships between words and, by extension, the ability to understand and generate natural language. Furthermore, that would be akin to humans learning how to write from reading literary works, something freely permitted under copyright law. (72)

81. Nevertheless, the choice made by the EU legislature, under Article 2 of the InfoSoc Directive, was to take not a *functional*, but a purely *technical* approach to the concept of reproduction. As explained above, the reproduction right was broadly defined in that provision to cover *any* digital reproduction of a protected subject matter. Under that definition, the purpose for which such a reproduction is made is immaterial. (73)

82. The AI Act confirms the intention of the legislature in that regard. Recital 105 of the AI Act implies (74) that the process of LLM training involves the carrying out of acts of reproduction falling, as a rule, under the sole purview of the rightholders concerned. Article 53(1)(c) of the AI Act, pursuant to which providers of GPAI are to 'put in place a policy to comply with Union law on copyright and related rights', including during training, starts from that premiss.

83. Finally, at the level of principles, I am not convinced that the use of protected works and other subject matter for the purposes of LLM training should fall entirely outside the scope of copyright and the related rights. With respect to copyright, in particular, it seems to me that LLMs 'learn' their parameters by 'consuming' the creative expression contained in the works used. (75) The model performance depends, in part, on the quality of that expression. Admittedly, that 'consumption' differs from the traditional conception of exploitation. Nevertheless, it is particularly relevant from an economic point of view. LLM developers derive considerable value from such training (since it allows them to deploy a functional model, from which they may draw considerable revenue). Furthermore, the purpose of that use is for the machine to acquire the capability to produce synthetic creations with potentially disruptive effects on the demand for human creators (see below for further discussion of this point). This is not comparable to a human learning to write by reading the works of others, if only because of the industrial scale of the process. If rightholders (and, especially, *authors*) were to be entirely disregarded in that process, that would hardly contribute to the high level of protection sought by the EU legislature. (76)

(2) *The trained model in itself*

84. Several of the interested parties who submitted observations have argued, in their response to the second and third questions, that the use of protected works and other subject matter for the purposes of LLM training results in *the model containing reproductions* of some or all of those works and other subject matter. That is another controversial and complex issue.

85. In general terms, as explained above, and as the referring court itself observes, a trained LLM is not a *database*. It is not a repository of (parts of) works or other subject matter that the model accesses and 'collates' to produce an output. During the training process, the training dataset is not straightforwardly copied within the model. Rather, as Google underlines, the model integrates information extracted from that dataset, namely individual tokens (numerically represented) and the statistical patterns and relationships connecting them. In that respect, the model's parameters are designed to reflect the manner in which words are used not in *a particular text* in the dataset, but *in the dataset generally*. (77) Such abstraction representation of words and language patterns could hardly qualify as reproductions of the works and other subject matter in question: copyright and the related right do not subsist at that level of abstraction. (78)

86. Nevertheless, there have been documented examples of LLMs being able to 'regurgitate', verbatim or near-verbatim, parts of the works on which they had been trained when prompted appropriately (for instance, by providing the beginning of a sentence taken from such a work). This was treated as evidence that LLMs sometimes 'memorise' specific pieces of the training dataset during the training process. (79) While the extent of that phenomenon is disputed, this has led some experts – and some national courts (80) – to extrapolate that, in such circumstances, at least part of the works concerned must be reproduced somewhere within the model. In their view, the fixation of works in the form of statistical patterns within the model's parameters would not be functionally different from other kinds of digital reproductions of works: like with a standard .pdf file, those works would not be directly visible by 'looking inside the model', but could be made perceptible to the human senses through a technical process (the translation of the file with a PDF reader in the first case, the prompting of the LLM, leading to the 'regurgitation' of the work, in the other). (81)

87. I would advise the utmost caution on this issue.

88. From the outset, one should not confuse what seems to a be a glitch with the overall functioning of an LLM. An LLM is normally designed not to 'memorise' and 'recite' the texts on which it was trained, but to 'generalise', that is to say generate *new text* on the basis of the statistical patterns and relationships it 'learned' during training. According to the available literature on the matter, 'memorisation' occurs when the model, during the automated machine learning process, unintendedly 'learns' not only *abstract* language patterns, but also patterns which are *specific* to a given piece of text in its training dataset. This seems to be the result of a phenomenon called 'overfitting': when a piece of text is duplicated many times in the training dataset (because, for instance, it consists in the lyrics of a famous song, scraped multiple times from the World Wide Web) and, accordingly, the model is overly exposed to the text in question, it may 'learn', within its parameters, the *specific sequence of tokens* corresponding to that text. When the beginning of that text is prompted by a user, the model may recite the rest because it will appear to be the probable, statistical answer in the light of its parameters. Such 'memorisation' appears *incidental* to the machine learning process, and not entirely avoidable. (82)

89. That being clarified, it seems to me that, on the one hand, where an LLM generates, upon a user prompt, a text which is identical, or near-identical, to (part of) a protected work on which it was trained, as the case may be because the model has 'memorised' and 'regurgitates' it, such an output may, in itself, constitute a reproduction of that work and, thus, infringe the exclusive right of the author concerned. That should be determined on a case-by-case basis. (83)

90. On the other hand, one should not infer from the fact that an LLM may generate reproductions of works on which it was trained when prompted appropriately that such reproductions are stored inside the model. At this stage of my consideration, such an interpretation seems far-fetched for several reasons. The way an LLM 'memorises' a piece of text is, again, not comparable to the way a digital repository stores a book. The 'memorised' works is split into countless parameters, the connection between which allows the model to predict that, in the sentence 'we all live in a yellow ...', the term 'submarine' is most probably the missing term. (84) Thus, it is more a complex recalling of the interrelationships between words, which resembles, to a certain extent, what happens in the human brain. (85)

91. Furthermore, the process of 'regurgitation' of a work by an LLM involves randomness (given that the millions of parameters of the model will not always predict the same output, even when faced with the same prompt). (86) Mostly, contrary to the argument set out above, prompting an LLM is not like opening a .pdf file with a PDF reader. The prompt provided by the user is, in itself, and essential component in such a process of 'regurgitation': the calculation and generation of an output by the model is a direct reaction to the data provided therein. The risk of 'regurgitation' increases with the precision of that prompt, its repeated nature, the length of the text excerpt given as input, and so on. (87) In other words, it is the *combination* of the model's parameters and of the user's prompt that *may* lead to the model generating reproductions of its training dataset. Thus, those parameters, seen in isolation, are *merely preparatory material* for the potential generation of such reproductions by the LLM. (88)

92. Finally, I am not convinced at this stage that such a far-fetched interpretation of the reproduction right under Article 2 of the InfoSoc Directive would ensure a 'fair balance' between the interests of rightholders, those of AI developers, and those of the general public. Since, as explained above, 'memorisation' is incidental to the machine learning process, and not entirely avoidable, (89) that interpretation would mean that LLMs are, *in themselves*, subject to the rightholders' sole purview, irrespective of how they are designed and used. The mere offering of such a model would constitute an act of exploitation of the works and other subject matter used during training, which would require the rightholders' prior consent. That could lead to an excessive liability burden for the developer and significantly impair the development of models. (90) By contrast, to consider that reproductions are not stored within the model does not deprive rightholders of the means to control, and obtain remuneration from, the use of their works and other subject matter in relation to LLMs. First, as will be explained in further detail below, those rightholders can oppose – or licence – the use of their protected content for the purposes of LLM training – the permissibility of which will also depend on whether the developers have put in place appropriate measures to mitigate the risk of 'memorisation' (91) or, at least, 'regurgitation' of that content. (92) Secondly, as has been stated above, reproductions of protected works and other subject matter generated by the model as outputs are, in themselves, actionable.

(c) ***Whether the reproductions of protected works and other subject matter carried out for the purposes of LLM training are exempted under the TDM exception***

93. It follows from the preceding sections that the use of protected works and other subject matter for the purposes of LLM training involves the making of reproductions of those works and other subject matter within the meaning of Article 2 of the InfoSoc Directive. Accordingly, that use prima facie falls under the rightholders' exclusive right. It remains to be seen whether those reproductions may nevertheless be freely permitted under one of the exceptions and limitations provided by EU law.

94. One such exception must be discussed from the outset, namely the temporary reproduction exception laid down in Article 5(1) of the InfoSoc Directive, on which Google relies before the Court. That provision freely permits temporary reproductions where (i) they are 'transient or incidental' and (ii) 'an integral and essential part of a technological process', (iii) their 'sole purpose is to enable a transmission in a network between third parties by an intermediary, or a lawful use, of a work or other subject matter', and (iv) they have 'no independent economic significance'. (93)

95. Admittedly, that exception may cover *some* of the digital reproductions of works and other subject matter carried out for the purposes of LLM training, in particular the temporary, technical reproductions created in the computer infrastructure as by-products of the machine learning process, as mentioned in point 78 above. (94) Nevertheless, it cannot cover *all* of those reproductions.

96. Specifically, it cannot apply with respect to the reproductions of works and other subject matter that form part of the training dataset, discussed in point 76 above. (95) First, the reproductions in question can hardly be regarded as 'temporary' or 'transient' within the meaning of Article 5(1) of the InfoSoc Directive. Those reproductions are stored in various servers at least (96) until the end of the LLM training process (which may take months). Secondly, those reproductions are not 'incidental' to a 'technological process', of which they form an 'integral and essential part'. Indeed, the dataset is not created *by* the training process. It is created in order *for* that process to be carried out. Finally, as I have suggested above, that dataset has a considerable 'economic significance' which is, furthermore, *independent* from the economic advantage that LLM developers derive from the training process carried out on that dataset. Indeed, such a dataset has *commercial value in itself* – which is evidenced by the fact that many entities now specialise in the creation and marketing of such datasets.

97. That being clarified, the debate and the referring court's third question centre on a more serious contender, namely the TDM exception laid down in Article 4(1) of the CDSM Directive. That provision freely permits reproductions of protected works and other subject matter (97) carried out 'for the purposes of text and data mining', even by private operators for commercial purposes, under certain conditions. (98)

98. Whether that exception is applicable to the reproductions of works and other subject matter carried out in relation to LLM training is a seminal issue in EU artistic and literary property law. If those reproductions fall under the TDM exception, the use of protected works and other subject matter for the purposes of such training is, as a rule, freely permitted in the European Union. By contrast, if that exception does not apply, then that use falls under the rightholders' sole purview – and, thus, requires their prior consent and compensation.

99. I tend to think that the TDM exception *does*, in principle, (99)*apply* to the reproductions of protected works and other subject matter carried out for the purposes of LLM training. This stems, in my view, from the wording, context and purpose of that exception (1). Furthermore, contrary to what is put forward by some of the governments who submitted observations, it seems to me that such an interpretation complies with the 'three-step test' (2).

(1) *Textual, contextual and purposive interpretation of the TDM exception*

100. I would recall that the TDM exception applies to reproductions of protected works and other subject matter which are carried out 'for the purposes of text and data mining'. In that respect, the concept of 'text and data mining' is defined, under Article 2(2) of the CDSM Directive, as 'any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations'.

101. As Google and the Estonian and French Governments submit, that legal definition of 'text and data mining' is broad enough to cover the use of protected works and other subject matter for the purposes of LLM training. Indeed, during that training, the dataset, composed of 'text and data in digital form', is subject to an 'automated analytical technique' (the machine learning process) 'in order to generate information' (namely, the statistical patterns and relationships which will become the model's parameters). (100)

102. Furthermore, since the TDM exception covers *any* reproduction created *for that purpose*, it may cover the scraping of protected works and other subject matter on the World Wide Web, the storing of the dataset on servers, and whatever other reproductions are created in the course of the training process. (101) In particular, with regard to the storing of the dataset, it is apparent from Article 4(2) of the CDSM Directive that any reproduction created for the purposes of 'text and data mining' may be freely retained, pursuant to that exception, for 'as long as is necessary' for that purpose, meaning pending and during LLM training. (102)

103. The Danish and Spanish Governments nevertheless make the argument that 'text and data mining', within the usual meaning of the term, is something quite different from LLM training. That term usually refers to the automated analysis of a large corpus of data in order to gain *insight into that corpus* (which can then be used for various purposes, including predictive or user behaviour analytics). It is generally referred to as 'Big Data' analysis. By contrast, LLM training consists in the automated analysis of a considerable corpus of texts in order for the machine to 'learn' how to *produce new information* (the substance of which has, often, nothing to do with that of the data on which it was trained). When the TDM exception was adopted, the EU legislature only had 'Big data' in mind. LLM training was not yet 'a thing': GPT-1 was released a few months after the adoption of the CDSM Directive. Those considerations are all the more important given that the TDM exception should, as a derogation to the reproduction right under Article 2 of the InfoSoc Directive, be interpreted strictly. (103)

104. I do not dispute that LLM training is quite different from 'Big Data' analysis. Nevertheless, that is not decisive in my view.

105. First of all, the broad definition of 'text and data mining' in Article 2(2) of the CDSM Directive is not confined to 'Big Data' analysis. In particular, that definition does not circumscribe the type of information that should be extracted from the text and data through the 'automated analytical technique' at issue. It may well include the statistical relationships and patterns between words in the English language. The ultimate use of that information does not matter for the purposes of that definition either. Hence, the fact that those patterns and relationships are eventually used by the model to generate new text is irrelevant.

106. Secondly, while the TDM exception – like the other exceptions and limitations provided in EU law – must be interpreted strictly, the Court has repeatedly ruled that such an interpretation should nevertheless 'enable the effectiveness of the exception thereby established to be safeguarded and its purpose to be observed'. (104)

107. As the German Government points out, it seems to me that the aim pursued by the EU legislature under the TDM exception changed quite a lot during the legislative process. While the Commission only suggested the introduction of a narrow exception for 'Big Data' analysis carried out for the purposes of scientific research, (105) the EU legislature decided to adopt, in addition, a broad exception covering 'text and data mining techniques [which] are widely used both by private and public entities to analyse large amounts of data' for various purposes, including 'the development of new applications or technologies'. (106) The aim of that new exception was to 'provide for more legal certainty' with respect to the permissibility of automated computational analysis of protected works and other subject matter and, thus, to 'encourage innovation also in the private sector'. (107)

108. In that context, the argument that the legislature did not envisage LLM training at the time of the adoption of the CDSM Directive is not entirely accurate. To give one example, a study commissioned, at the time, by the JURI committee of the European Parliament on text and data mining expressly 'flagged' AI and machine learning, indicating that that would acquire 'increasing centrality' and that, 'although classical TDM and machine learning have different utility, it should not be overlooked that both use the same key algorithms to discover patterns in data'. (108)

109. Thirdly, and in any event, as Google and the German, Estonian, French and Hungarian Governments state, whatever the EU legislature had in mind at the time of the adoption of the CDSM Directive, it has clearly expressed its intent in the AI Act in the meantime. It is clear from recital 105 and Article 53(1)(c) of the AI Act that, in the EU legislature's view, the TDM exception applies to the use of protected works and other subject matter for the purposes of LLM training. (109)

(2) *The three-step test does not call for a different interpretation*

110. Nevertheless, the Hellenic Republic and the Spanish Government submit that the TDM exception cannot be interpreted as covering the use of protected works and other subject matter for the purposes of LLM training because exempting such use, through that exception, from the rightholders' purview would be contrary to the three-step test imposed under international law. (110)

111. I would recall that various international conventions that are binding on the European Union in the field of artistic and literary property (111) provide that, while the contracting parties may lay down exceptions to copyright and the relevant related rights, those exceptions must be restricted to (i) 'certain special cases' which (ii) 'do not conflict with a normal exploitation' of the work or other subject matter used and (iii) 'do not unreasonably prejudice the legitimate interests' of the rightholders concerned.

112. It is understood that, when interpreting the scope of an exception provided in EU law, the three-step test should be borne in mind. Indeed, such an exception could not be construed as freely permitting uses of protected works and other subject matter that would *systematically fail that test*. Such an interpretation would put that exception on a collision course with the European Union's international obligations. (112)

113. Nevertheless, together with the German, Estonian, French and Hungarian Governments, I am inclined to think that the three-step test does not preclude interpreting the TDM exception as covering, in principle, the use of protected works and other subject matter for the purpose of LLM training. Indeed, the exemption of such a use, under the conditions provided in that exception, does not systematically clash with any of the steps, as will be examined below.

(i) *A special case*

114. In order for the first step of the test to be satisfied, the use of protected works and other subject matter for the purposes of LLM training must be regarded as a special case. The Hellenic Republic and the Spanish Government reject such a qualification, putting forward the massive character of that use.

115. Admittedly, if one were to take a purely quantitative approach of the number of affected works and other subject matter, LLM training could hardly be regarded as a special case. Nevertheless, I do not think that such an approach is called for. Several exceptions and limitations allow for large-scale use of works and other subject matter (from the 'private copying' exception to the temporary reproduction exception) while being regarded as special cases. The proper test, in that respect, is rather whether the exempted use has sufficiently well-defined contours, and is narrow from a conceptual (rather than purely quantitative) point of view. (113)

116. That is the case here. As the German Government observes, under the suggested interpretation of the TDM exception, the exemption is limited to the use of protected works and other subject matter 'as data' for the purpose of extracting information, through automated analysis, including in the context of LLM training. (114) Furthermore, the benefit of that exemption is granted to a narrow circle of beneficiaries (including tech developers using such techniques). (115)

(ii) *No conflict with a normal exploitation of the work or other subject matter used*

117. Under the second step of the test, the use of protected works and other subject matter exempted under the relevant exception cannot 'conflict with a normal exploitation' of those works or subject matter. This is where the debate concerning LLM training is the most intense.

118. In the Spanish Government and the Hellenic Republic's view, the reproductions of works and other subject matter carried out by LLM developers for the purposes of model training do 'conflict with a normal exploitation' of these works and other subject matter. (116) In their view, if those reproductions were exempted under the TDM exception, it would 'result in a diminution of the volume of sales or other lawful transactions' relating to the works and other subject matter concerned (117) in two distinct ways.

119. First, the existence of such an exemption would impair the possibility for rightholders to license the use of their works and other subject matter for model training and, thus, the development of the market for such licences ('input dimension').

120. Secondly, the outputs that the LLM would eventually generate, on the basis of its training, could constitute *direct substitutes* for the works and other subject matter on which the model was trained or, at least, *compete* with them to some extent. Specifically, as discussed above, LLMs may sometimes generate outputs reproducing pieces of their training data. However, beyond that scenario, LLM-generated content that simply mimics the general patterns of such works and other subject matter (for instance, a novel generated by an LLM mimicking the syntax, abstract stylistic patterns, and so on, of human-made novels) could be regarded, by consumers, as interchangeable with those works and other subject matter. Given that LLMs can easily produce such content on an industrial scale, that would create a risk of 'diluting' the markets for such works and other subject matter, diminishing the value of human-made creations. Finally, and most importantly, given that an LLM can generate such content faster and more cheaply (118) than humans, that content could eventually displace the demand for human-made creations and creators altogether ('output dimension').

121. There is a lot to unpack in those arguments.

122. With respect to the input dimension, I have already explained why the use of protected works and other subject matter for the purposes of LLM training has a considerable economic value. Developers have an interest in training their models on quality material, and rightholders could very well license their portfolios to that effect. In fact, a market for such licenses already exists. (119) Clearly, that market could grow in importance and, thus, is relevant to the analysis of the second step of the test. (120) Obviously, freely permitting the use of works and other subject matter for the purposes of LLM training may directly 'result in a diminution of the volume of' such licences.

123. That being said, generally speaking, the argument that, if a use of protected works and other subject matter were not freely permitted under an exception or limitation, it could be licensed by the rightholders, cannot be decisive in itself. Otherwise, it would justify doing away with any exception or limitation to copyright and the related rights, no matter how important the public interest pursued by the EU legislature thereunder. (121)

124. In any case, here, the TDM exception is circumscribed by conditions designed to guarantee that rightholders remain able, should they wish, to license the use of their protected works and other subject matter for text and data mining purposes, including LLM training. (122) Specifically, under Article 4 of the CDSM Directive, the TDM exception may only apply on condition that (i) the works or other subject matter concerned are 'lawfully accessible' (Article 4(1)) and (ii) the use of those works or other subject matter for TDM purposes 'has not been expressly reserved by their rightholders' (Article 4(3)). Those conditions will be examined in detail in the following section. For the present purposes, suffice it to say that, in accordance with those conditions, where a rightholder has restricted access to his or her works or other subject matter, or indicated in an appropriate manner that he or she reserves the right to use them for TDM purposes, LLM developers could only do so if they obtained a licence to that effect from the rightholder. (123)

125. With respect to the output dimension, I would recall that the free use discussed in the present context is that of protected works and other subject matter for the purposes of LLM training. The issue, according to the second step of the test, is whether *that use* conflicts with a normal exploitation of the works or other subject matter concerned. Exempting that use from the rightholders' monopoly entails, as a *direct consequence*, a potential diminution of licensing revenues (input dimension). By contrast, the output dimension argument relies on the *indirect, downstream effects* that such a use could *ultimately* entail for the exploitation of those works and other subject matter.

126. Whether such indirect, downstream effects should be factored into the analysis of the second step is a new legal issue and, again, a controversial one. So far, the Court has only been concerned with the question whether a given reproduction of a work could be substituted for other copies of that work, not whether the technology designed on the basis of that reproduction could, ultimately, have repercussions on that work's normal exploitation. The German Government suggests that those effects are too *far removed* from the reproductions carried out at the training stage to be taken into account. (124)

127. At this stage of my consideration, it seems to me that the effects in question *should* be taken into account, (125) but to *a certain extent only*.

128. On one hand, the capability of a trained LLM to generate content that mimics human-made works generally and could compete with them to some extent is, in my view, too *diffuse* a consequence to be relevant to the second step.

129. Such AI-generated content does not 'conflict with a normal exploitation' of the works and other subject matter on which the LLM was trained specifically. It may affect the exploitation of *human-made creations*, in a more abstract manner. Besides, the capability in question is not the result of the use of the specific works or subject matter concerned during 'training'. It is the result of 'training' generally. Even if an LLM were trained on open data, it would still have that capability (maybe to a different, but still significant, extent). The same would go if a model were trained on a portfolio licenced by a rightholder.

130. However, copyright law and the related rights are not designed to protect rightholders *against competition per se*. Copyright law, in particular, does not protect an author against the creation of new content that competes with his or her works. The fact that those works were used at some point in the process of creation of such competing content is irrelevant. Everyone is free to reuse (as the case may be, through a technological process) works to create new material, even competing material, so long as that material borrows not the 'original', protected elements of those works, but unprotected, general features such as syntax, abstract stylistic patterns, and so on. (126)

131. An additional consideration should be borne in mind here. The second step of the test is a threshold under public international law. If a given use of a work 'conflicts with a normal exploitation' of that work, it must imperatively remain within the rightholders' sole purview. (127) Thus, if the use of works and other subject matter for the purpose of LLM training were considered to systematically 'conflict with the normal exploitation' of those works and other subject matter, according to a broad conception of those terms, that would not only prevent the application of the TDM exception to such a use but also, more generally, remove any discretion for the EU legislature to regulate the matter. That legislature could no longer assess the various interests at issue and come up with mechanisms, other than pure exclusivity, such as a different exception and/or a statutory licensing or remuneration mechanism, to ensure that a fair balance is struck in that delicate field. Such a radical outcome surely cannot be endorsed lightly.

132. On the other hand, it seems to me that every time LLM developers seek to rely on the TDM exception to justify the use of protected works and other subject matter for the purposes of model training, the particulars of the model are important. Where a given model could easily be used to generate content directly substituting itself for specific works or other subject matter forming part of its training data, that cannot be ignored. Beyond the scenario of a model 'regurgitating', verbatim or near-verbatim, parts of a work it was trained on, as discussed above, I have in mind a model that could routinely generate content that resembles specific pieces of its training dataset to such an extent that the latter is 'recognisable' in the former (effectively, copyright-infringing reproductions). (128) Such content would have more concrete market displacement effects. That could be intentional on the part of the LLM developers, or due to negligence. In that respect, I take the view that such developers can be expected to take appropriate measures to mitigate the risk of their model generating outputs that resemble specific pieces of their training material too closely (such as proper dataset diversification (129) and deduplication, implementation of prompt and output filters, and so on).

133. I would recall that, pursuant to Article 5(5) of the InfoSoc Directive, the three-step test must also be verified at *the stage of applying* the exceptions and limitations provided under EU law, including the TDM exception, (130) in a given case. Thus, whether the use of protected works and other subject matter to develop a specific AI model passes the second step should be ascertained every time the LLM developers rely on the TDM exception. (131) In that regard, developers wishing to benefit from that exception should demonstrate that they have taken the measures envisaged in the preceding point in order to mitigate the risk of their model generating outputs that resemble specific pieces of their training material too closely. (132) Failing that, the national court adjudicating the case should refuse the application of the TDM exception.

(iii) *No unreasonable prejudice to the rightholders*

134. The third step of the test consists in assessing whether the use of protected works and other subject matter for the purposes of LLM training 'unreasonably prejudice[s] the legitimate interests' of the rightholders concerned. Obviously, any use by a third party of a work or subject matter under an exception or limitation to copyright or the related rights prejudices, to some extent, the interests of the rightholders concerned. However, the decisive question is whether that prejudice is justified by a legitimate objective as well as being proportionate to the latter (and, thus, 'reasonable'). That step involves an analysis of whether exempting the use at issue ensures a 'fair balance' between, on the one hand, the rights and interests of the rightholders and, on the other hand, the rights and interests that the exception seeks to protect, namely those of model developers, but also of the users of AI systems integrating an LLM. Related public interests must also be taken into consideration. (133)

135. On the one hand, the prejudicial effects for rightholders have already been mentioned. While model developers derive considerable value from the free use of protected works and other subject matter for the purposes of LLM training, that use entails a loss of control and potential licensing revenue for the rightholders. (134) Furthermore, since a trained LLM, once integrated in an AI system, may be used to generate outputs that (rarely) directly infringe or (more often) compete, to some extent, with the works and other subject matter on which the LLM was trained, that use may ultimately diminish the value of the rightholders' intellectual property. Overall, the use in question may significantly restrict their fundamental right to intellectual property under Article 17(2) of the Charter.

136. Beyond those private interests, several related public interests should be kept in mind. As stated above, the generation of 'synthetic creations' by trained models may, depending on the market, have *diffuse*, yet *serious* effects on the demand for human creations and *human creators* altogether. In the current state of the technology, LLM-generated content will not displace the demand for human-made, highly original novels and other masterpieces of literature. Nor will it displace demand for in-depth, analytic pieces of journalism. However, it could easily replace human creators as regards texts with low originality, such as purely informative news or sports reports, (135) romance novels (and so on). (136) Thus, ultimately, the training of LLMs on human-made works and other subject matter may significantly restrict the freedom of expression under Article 11 of the Charter and/or the freedom of the art under Article 13 of the Charter (depending on whether or not the expression is artistic) of human creators, as well as their freedom of occupation under Article 15 of the Charter and/or to conduct a business under Article 16 of the Charter (depending whether they are employed or self-employed). The impact of generative AI on creation may, in turn, negatively affect the dissemination of information, knowledge and culture in the public sphere. It may result in an overall decline in the quality of cultural and informational production (the 'flooding' of the World Wide Web with AI-generated 'slop' and low-quality informational content being symptoms of the latter). (137) More fundamentally, as it becomes harder and harder to separate the 'synthetic' from the 'human', it may ultimately result in widespread public distrust of, rather than curiosity in, unknown content.

137. On the other hand, freely permitting the use of protected works and other subject matter for the purposes of LLM training under the TDM exception supports innovation. Such an objective is, in itself, legitimate. It relates to the freedom of model developers to conduct a business under Article 16 of the Charter. Moreover, there is a general public interest in the development of well 'trained' models. Such models are capable of countless uses and advancements that are beneficial to society. They may be used to develop new creative processes, products and services, or to provide efficiency gains in many fields. Moreover, when used wisely, they may serve human creators in their creative process. Overall, such models may contribute to users' freedom of information under Article 11 of the Charter, freedom of the art under Article 13 of the Charter, as well as their freedom to choose an occupation and to conduct a business under Articles 15 and 16 of the Charter. If model developers had to seek individual authorisation for rightholders for every single piece of protected content they intend to use for training, that would result in considerable transactional costs for them and/or significantly limit the pool of training data available, potentially affecting the performance of the model and, in turn, impairing such beneficial innovation. (138)

138. Regulating the training of models involves a particularly complex balancing of those various rights and interests and, ultimately, political, economic and social choices. The EU legislature must therefore be allowed a *broad discretion* in that regard. (139) The TDM exception, and the notion that it covers AI model training, expressed in recital 105 of the AI Act, reflect such choices. The third step of the test should accommodate that reality.

139. In my view, the EU legislature was accordingly at liberty to tip the scales more in favour of innovation and, through the TDM exception, relieve model developers from most of the burden associated with the 'clearing' of copyright and related rights. Digital sovereignty should not be forgotten in that equation. The legislature could take the view that the new policy of limiting the European Union's dependence on foreign technology involves the development of European AI models, and that the latter should be fostered, rather than hindered. (140) Allowing models to easily be trained in the light of works created in the European Union also ensures the representation of European culture in AI.

140. Besides, as the German Government recalls, the TDM exception is curtailed by two conditions designed to ensure that the use of protected works and other subject matter for the purposes of LLM training does not unreasonably prejudice the legitimate interests of the rightholders concerned.

141. In the first place, the application of the TDM exception is, I would recall, subject to the condition that the works or other subject matter concerned be 'lawfully accessible' (Article 4(1)). With respect to content available on the World Wide Web, it is clear to me that that concept covers access to *lawful sources* only (that is, webpages where such content has been made available to the public by the rightholders, or with the rightholders' consent). (141) Furthermore, when the rightholders have put in place (or requested that third parties put in place) effective technical measures (within the meaning of Article 6(3) of the InfoSoc Directive) (142) to restrict access to content (such as a paywall or a measure designed to prevent access by robots specifically (143)), that concept cannot cover content accessed by *circumventing such measures*. (144)

142. Indeed, if developers were free to scrape protected works and other subject matter from unlawful sources or by circumventing such technical measures in order to train their models, it would unreasonably prejudice the interests of rightholders. It would entirely displace the possibility for rightholders to license such use. That being said, as explained above, the considerable amount of data necessary to train a model is usually collected through automated web scraping by internet bots. From a technical point of view, it is difficult to limit such scraping to lawful sources (since bots cannot automatically assess whether content has been put online by the rightholders, or with their authorisation). If the condition of lawful access under Article 4(1) of the CDSM Directive were too strictly interpreted, this could make it impossible for developers to have recourse to such web scraping, thus depriving the TDM exception of a significant part of its effectiveness. Rather, that condition should be interpreted, in a balanced way, as imposing a duty on model developers not to scrape *obviously unlawful sources* (such as pirate websites and shadow databases) and not to circumvent such measures. (145)

143. In the second place, I would recall that even where works and other subject matter are 'lawfully accessible' online, they can only be collected by model developers under the TDM exception when their use for text and data mining purposes 'has not been expressly reserved by their rightholders' (Article 4(3)).

144. That is another essential aspect of the fair balance inherent to the TDM exception. While, by default, rightholders whose content is 'lawfully accessible' online have to tolerate the scraping and use of that content by third parties for text and data mining purposes, including LLM training, they may *reserve* such use and 'opt out' of the TDM exception. This

technology allows the chatbot to find related content on webpages on the World Wide Web (essentially, by using Google Search), to retrieve that content (yet again using web scraping bots), and to integrate that content in the prompt which is fed to the LLM. (165) The model then generates, as output, an answer 'grounded' in that content, so that both *form* and *substance* are correct (at least, theoretically). (166) Furthermore, RAG allows the LLM to provide, in the answer it generates, the source(s) scraped in the form of hyperlinks to the webpages concerned. In fact, it appears that such hyperlinks are sometimes, but not always, provided in the answers provided by the Gemini chatbot. (167)

156. In effect, RAG gives an LLM-powered online chatbot, such as the Gemini chatbot, capabilities similar to those of a search engine. (168) Nevertheless, where a traditional search engine would merely provide the user with a list of hyperlinks corresponding to webpages that he or she must access in order to obtain the information sought, RAG, coupled with the LLM generative capability, allows the chatbot to collect, piece together, sum up and provide that information *directly*, in full sentences. That technology is also behind 'AI Overviews' and 'AI Mode', which are two separate AI services, integrated into Google Search and also powered by the Gemini LLMs. Under 'AI Overviews', in particular, when an internet user enters a query into the Google Search bar, those services 'do the searching for [the user]' (169) and provide presentation or summary of information retrieved from the Web.

157. Thanks to RAG technology, AI chatbots (and services such as 'AI Overviews' and 'AI Mode') may be used, in particular, to *obtain information about the news*. Such a chatbot can, when asked by a user about a given news item, retrieve relevant pieces of content from the World Wide Web – typically by scraping related press articles available online – and provide an LLM-generated presentation on the matter. As the example put forward by Like Company shows, that chatbot may even provide a summary of a specific article. Certain AI providers (such as Perplexity) even specialise in those kinds of services.

158. However, that particular use of online chatbot services (and of services such as 'AI Overviews' and 'AI Mode') has been a source of grievances from press publishers. They claim, first, that the providers of those services do not seek a publisher's consent (or a fortiori provide compensation to the latter) before scraping their press publications to 'ground' the answers provided to users. Secondly, such AI-generated content diverts internet users away from the press publishers' websites, resulting in a considerable loss of traffic and, by extension, of advertising or subscription revenue for those publishers. Those grievances have crystallised in a series of recent actions brought by publishers against various AI providers. (170) They are reiterated by Like Company in the present case. (171)

159. I would point out that that debate reflects a broader concern with respect to the functioning of the now AI-driven World Wide Web. The original business model of the Web rested on a delicate *quid pro quo* between search engines, such as Google, and website publishers. When those engines were crawling the web, copying websites and displaying the latter on their result pages, they were bringing back traffic that the website publishers could monetise. That *quid pro quo* is called into question now that LLM-powered online chatbots themselves (and services such as 'AI Overviews' and 'AI Mode') retrieve and provide information directly to internet users, without the latter having to visit any other webpage(s) in order to obtain what they were looking for. (172)

160. The first and fourth questions put by the referring court should be read in that context. By those questions, which should be examined together, the referring court wishes to ascertain, in essence, whether the provider of an LLM-powered online chatbot, such as Google, carries out one or more act(s) of use of a press publication, falling within the scope of Article 15(1) of the CDSM Directive, where, while processing a user query, that chatbot retrieves extracts of that publication from a press publisher's website, reuses them to generate its answer, and provides that answer to the user.

161. Those questions – which, contrary to what Google and the Commission suggest, are admissible (1) – are the real crux of the present case. Again, they are sensitive, complex and highly topical. Rather than the future of creation per se, they concern that of the World Wide Web. This calls for a thorough assessment (2).

1. ***The first and fourth questions are relevant and, thus, admissible***

162. Unlike the second and third questions concerning the training of an LLM, the first and fourth questions are relevant to the case in the main proceedings and, thus, admissible. First, it may reasonably be assumed that Like Company's articles have been scraped and reused by the Gemini chatbot, in real time, in order to provide an answer to a user query (a). Secondly, Article 15(1) of the CDSM Directive may possibly offer protection with respect to such use: the principle of territoriality would not preclude it (b); that use qualifies as 'online use' of a press publication (c); and, contrary to what Google and the Commission submit, it is not *acte clair* that that provision freely permits such use (d).

(a) ***Like Company's articles have presumably been used by the Gemini chatbot to provide answers to users' queries***

163. In general terms, given that there is currently no obligation of transparency for AI systems providers with respect to the content scraped by LLM-powered online chatbots in order to provide answers to user queries, rightholders may struggle to establish, with certainty, that their works and other subject matter are used in that way. In order for the protection granted to rightholders by EU artistic and literary property law to be effective, courts should be lenient with respect to the burden of proof borne by claimants in that regard. Specifically, it would be reasonable to *presume* that a piece of content is used thus by such a chatbot where the latter has access to that content (typically, it is available online) and the claimant has provided a reasonable sample of answers provided by that chatbot (173) which are similar to that piece of content, or contain hyperlinks leading back to the webpage hosting that content. Such examples could be regarded as *prima facie* evidence of use, shifting the burden on to the AI provider to prove the contrary. (174)

164. That being said, in the case in the main proceedings, it can readily be assumed that Like Company's articles have been used by the Gemini LLMs through RAG to generate answers to user queries at least once: in the example provided by Like Company before the referring court, an unknown user specifically requested the Gemini chatbot to provide a summary of one of those articles and directed the chatbot to Like Company's web portal to that effect.

(b) ***Article 15(1) of the CDSM Directive may legitimately regulate the matter in accordance with the principle of territoriality***

165. The issues related to the application of EU copyright law to acts of training carried out in a third country, discussed in Section B above, find no equivalent here. It is clear that, under the principle of territoriality, Article 15(1) of the CDSM Directive may legitimately confer protection against the use of press publications by an LLM-powered online chatbot to ground its answers where that chatbot is (like Gemini) *provided to users within the European Union*, irrespective of the place where the chatbot provider is established or of the location of the servers from which it operates. (175)

166. Indeed, where that chatbot is provided to users in the European Union, such use produces immediate and substantial effects on the territory of the Member States, which justifies EU law regulating the matter in the light of public international law. (176) Thus, where a rightholder claims protection from a Member State (here, Hungary), the law of that country may legitimately regulate that use, pursuant to Article 8(1) of the Rome II Regulation.

(c) ***RAG is an online use of a press publication***

167. Contrary to what I have explained above with regard to LLM training, it is clear to me that the use of a press publication by an LLM-powered online chatbot, through RAG, to provided grounded answers to user queries is a form of online use of that publication within the meaning of Article 15(1) of the CDSM Directive.

168. As explained in point 55 above, the concept of online use of a press publication within the meaning of Article 15(1) of the CDSM Directive covers any use that occurs as an integral part of the provision of an information society service via the internet, at the individual request of a recipient of that service.

169. The acts of use of press publications discussed in this section meet that definition. Specifically, both the retrieval of such a publication through RAG and the provision of a grounded answer by the chatbot occur as an integral part of the provision of the chatbot service through the internet, at the individual request of a recipient of that service (namely the user who entered the initial query), as part of the transmission of information to that user. (177)

(d) ***The use in the main proceedings is not manifestly free under Article 15(1) of the CDSM Directive***

170. Google and the Commission argue, in essence, that the first and fourth questions are hypothetical because, in the main proceedings, it would be *acte clair* that the use at issue is free under Article 15(1) of the CDSM Directive. In their view, since, in the example given by Like Company before the referring court, the answer provided by the Gemini chatbot is not worded in the same way as the original press publication, no restricted act of reproduction or making available to the public has occurred through that chatbot. (178)

171. From the outset, I would observe that that objection relates to the substance of those questions, since it concerns the issue of what constitutes a reproduction and the making available to the public of a press publication for the purposes of Article 15(1) of the CDSM Directive, and whether such acts are carried out when a chatbot reuses such a publication, through RAG, to provide a grounded answer to a user query. Accordingly, that objection cannot call into question the admissibility of those questions, and should be addressed by the Court when assessing them on the merits. (179) Moreover, the matter is far from *acte clair*, in particular since the Court has never before clarified the scope of the press publishers' right. (180) Things are in fact more complicated than presented by Google and the Commission, as will be explained in the following points.

2. ***The permissibility of such use in accordance with the press publishers' right***

172. I would recall that, under Article 15(1) of the CDSM Directive, press publishers are granted, with regard to their press publications, the exclusive reproduction right provided under Article 2 of the InfoSoc Directive, and the exclusive right of 'making available to the public' laid down in Article 3 of that directive.

173. In my view, where an LLM-powered online chatbot retrieves extracts of a press publication from a press publisher's website and reuses them to provide a grounded answer to a user, that may, depending on the circumstances, have implications having regard to both rights.

174. Specifically, invisible (yet legally relevant) acts of reproduction of press publications may be carried out during the technological process of retrieval and analysis of such a publication (a). Furthermore, the answers provided by the chatbot at the end of such a process may, in certain circumstances, reproduce that publication and make it available to the public (b). Nevertheless, those different acts are only actionable where they are carried out by the chatbot provider (c). In that context, I will also address the question whether, in the modern, AI-driven World Wide Web, press publishers should be regarded as having implicitly consented to such acts (d). The balance inherent in the solutions provided with respect to those various aspects will be examined separately (e). Lastly, guidance will be provided with respect to the application of those solutions in the main proceedings (f).

(a) ***Relevant acts of reproduction of press publications may be carried out during the RAG process***

175. Starting with the technical, invisible side of things, the process of retrieval and analysis of a press publication by an LLM-powered online chatbot may involve the making of reproductions thereof within the meaning of Article 2 of the InfoSoc Directive (1). Depending on the circumstances, those reproductions are legally relevant or freely permitted (2)

(1) *RAG may involve the making of technical reproductions of press publications*

176. As explained in point 74 above, the reproduction right under Article 2 of the InfoSoc Directive is broadly defined to cover even invisible, digital reproductions of protected works and other subject matter which are created as by-products of a technical process. That is also the case for press publications. (181) Furthermore, the reproduction right also covers parts of a protected work or other subject matter, such as a single article within such a publication, or even in an excerpt of such an article. Only the borrowing of individual words or very short extracts of that publication is not covered, under the third sentence of Article 15(1) of the CDSM Directive.

177. As the Hellenic Republic submits, when an LLM-powered chatbot accesses and retrieves, in real time, a press publication on a publisher's website previously identified through a web search, (182) it scrapes and, therefore, essentially copies excerpts from that publication (typically, an article or paragraphs thereof that match the prompt entered by the user). Further along the chain of the process, multiple digital reproductions of those extracts are created during their automated analysis by the system, including by the LLM to generate the output that the chatbot will then provide in answer to the user query. Google admitted as much at the hearing. In so far as what is extracted and reproduced is more than 'individual words' or 'very short extracts' (which seems to be the case usually since, were it otherwise, the retrieval process would fail at collecting relevant information to include in the chatbot's answer), those reproductions prima facie fall under the right laid down in Article 2 of the InfoSoc Directive. (183)

(2) *Depending on the circumstances, those technical reproductions are legally relevant or freely permitted*

178. Whilst technical reproductions of a press publication may be created during the process of retrieval and analysis of that publication by an LLM-powered online chatbot, the answer to the question whether those reproductions are relevant acts of use under Article 2 of the InfoSoc Directive or, conversely, fall outside the scope of that right and, thus, are freely permitted depends on the exceptions and limitations provided by EU law.

179. The TDM exception, discussed above in relation to LLM training, is immaterial in that context. I would recall that that provision covers reproductions of works or other subject matter carried out 'for the purposes of text and data mining', defined as 'any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations'. The kind of RAG carried out by LLM-powered online chatbots is, arguably, an 'automated analytical technique' involving works or other subject matter 'in digital form'. Nevertheless, that analysis is not conducted 'in order to generate information which includes but is not limited to patterns, trends and correlations'. Indeed, such a RAG process is not carried out for the purpose of extracting information from a large dataset (such as, in LLM training, the statistical patterns and relationships between words that become the model's parameters). It consists instead in retrieving *some specific piece of content* from the World Wide Web in order to use it directly to 'augment' the accuracy of a specific answer. (184)

180. By contrast, as Google submits, such technical reproductions may, depending on the circumstances, be covered by the temporary reproduction exception laid down in Article 5(1) of the InfoSoc Directive.

181. I would recall that, under that provision, temporary reproductions of a protected work or other subject matter may be exempted from the right laid down in Article 2 of the InfoSoc Directive where (i) those reproductions are 'transient or incidental' and (ii) 'an integral and essential part of a technological process', (iii) their 'sole purpose is to enable a transmission in a network between third parties by an intermediary, or a lawful use, of [that] work or other subject matter', and (iv) they do not have 'independent economic significance'.

182. On the one hand, it is clear that the technical reproductions of press publications created during RAG are generally temporary and, furthermore, 'transient'. Those reproductions are created during that technical process and automatically deleted by the system, at the latest once the answer has been generated and that process is completed. (185) Indeed, those reproductions exist solely for the purpose of the chatbot providing accurate and, therefore, reliable answers to a user query.

183. The reproductions in question are also 'incidental' to, and form 'an integral and essential part' of, the RAG process: first, they neither exist independently of, nor have a purpose interdependent of, the process in question; secondly, they are created as by-products of that process; thirdly, the RAG process could not function correctly and efficiently without them. Indeed, the system cannot retrieve information on the web and use it to augment the output generated by the LLM without making such reproductions. (186)

184. Finally, it seems to me that the technical reproductions of press publications that may be created during such a RAG process do not have 'independent economic significance'. Admittedly, the retrieval of content online has considerable 'economic significance'. The accuracy – and, therefore, the usefulness – of online chatbot services and of services such as 'AI Overviews' and 'AI Mode' depend in part on such retrieval. Nevertheless, the 'economic significance' of those reproductions is not 'distinct or separable' (187) or, therefore, 'independent' from the economic advantage that the user or Google derives from the answers provided by the chatbot. Indeed, those invisible reproductions are created, I would recall, as an ancillary part of a process of the LLM generating a particular answer to a given user query, and solely for that purpose. The economic advantage derived from those reproductions only materialises when that answer is provided. They have no separate economic value on their own. (188)

185. On the other hand, it is equally clear that the sole purpose of the technical reproductions of press publications created during the RAG is *not* to 'enable a transmission [of those publications] in a network between third parties by an intermediary'. In the kind of RAG carried out by an LLM-powered chatbot, the bot is not transmitting the subject matter retrieved to the user as such, as a neutral intermediary would do, (189) but reusing it to provide an accurate, grounded answer to that user.

186. Accordingly, the decisive factor is whether the sole purpose of the reproductions in question is to 'enable a lawful use' of the press publications concerned within the meaning of Article 5(1) of the InfoSoc Directive. In my view, whether the end use enabled by those reproductions – namely the generation, by the LLM, of a grounded answer to a user query – constitutes such a lawful use of a press publication calls for *a nuanced answer*.

187. In that connection, recital 33 of the InfoSoc Directive specifies that the end use enabled by technical reproductions should be regarded as lawful where it is either 'authorised by the rightholder' or 'not restricted by law'. Here, assuming that the press publisher concerned has not authorised the provision, by a given LLM-powered online chatbot, of answers grounded in its press publications, (190) whether or not such an end use is 'restricted by law' (specifically, Article 15(1) of the CDSM Directive, read in conjunction with Article 2 and Article 3(2) of the InfoSoc Directive) depends, in my view, on *two cumulative factors*, namely (i) the manner in which the LLM accessed and retrieved the original press publication and (ii) the form of the answer ultimately provided to the user.

188. As to the first factor, it seems clear to me that such an end use can only be regarded as not restricted by law if the chatbot has *accessed* the press publication in question *lawfully*. It follows in particular that, where the press publisher concerned has put in place *effective technical measures* (within the meaning of Article 6(3) of the InfoSoc Directive) to restrict access to that publication (be it by way of a paywall or a measure designed specifically to prevent access and/or scraping by internet bots), the chatbot should not have *circumvented* those measures. (191)

189. As to the second factor, it seems equally clear to me that the question whether the provision, by an LLM-powered online chatbot, of answers grounded in press publications is restricted by law also depends on whether the press publishers' right may subsist in those answers. (192) That issue will be addressed in detail in the next section. For the present purposes, suffice it to say that the question whether the press publishers' right may subsist in such an answer depends on the form thereof. That right cannot apply with respect to answers which are sufficiently removed from the form of expression in the original press publication. By contrast, answers which resemble the form of that publication too closely may qualify as reproductions of the latter within the meaning of Article 2 of the InfoSoc Directive. In certain circumstances, such answers may even make the publication available to the public, within the meaning of Article 3 of that directive. Thus, answers belonging to the second and third categories may, in themselves, fall under the press publishers' right (193) and accordingly be regarded as 'restricted' under Article 15(1) of the CDSM Directive.

190. In my view, where those conditions are satisfied, the exemption from the right under Article 2 of the InfoSoc Directive of technical reproductions of press publications created during the RAG process satisfies the three-step test and, in particular, does not unreasonably prejudice the legitimate interests of the press publishers concerned. Nevertheless, that is, in essence, a 'fair balance' issue, and I reserve the matter for the penultimate section of this Opinion.

(b) ***Whether the answer provided by the chatbot constitutes, in itself, a reproduction of the press publication retrieved and makes the latter available to the public***

191. I shall now address the visible part of things. In that respect, the referring court seeks to know, in essence, the circumstances in which the press publishers' right may subsist in the answer provided by an LLM-powered online chatbot at the end of a RAG process.

192. Specifically, the referring court wishes to ascertain, in essence, whether, and if so in which circumstances, such an answer may, in itself, (194) qualify as a reproduction (within the meaning of Article 2 of the InfoSoc Directive) of the press publication reused in the RAG process (1). In addition, it wishes to know whether, by providing that answer, the chatbot makes that publication available to the public, within the meaning of Article 3(2) of that directive (2). (195)

(1) *Whether the answer provided by the chatbot qualifies as a reproduction of the press publication at issue*

193. As to whether, and if so in which circumstances, the answer provided by the LLM-powered online chatbot to the user may be regarded as a reproduction (within the meaning of Article 2 of the InfoSoc Directive) of the press publication reused during the RAG process, the Court's usual infringement test should be followed. An AI-specific approach to the matter would be contrary to the principle of technological neutrality on which EU artistic and literary property law rests. (196)

194. Under that test, the question whether a given piece of content (here, an answer provided by the chatbot) constitutes a potentially infringing reproduction of a protected subject matter within the meaning of Article 2 of the InfoSoc Directive depends on two cumulative conditions: that subject matter must have been (i) copied to create that content ('copying') (ii) in such a way that (at least some of) the protected elements that compose that subject matter are perceptible (or, to use the Court's own term, 'recognisable') in the end result ('recognisability'). (197)

195. As regards the first condition, it follows from point 177 above that the scraping and reuse of a press publication by a chatbot through RAG involves copying. Contrary to what Google suggests, the fact that that publication is not *directly* copy-pasted in the answer provided to the user, but is instead integrated in the prompt on the basis of which the LLM generates that answer, does not call that interpretation into question. Irrespective of the details of that technical process, there is, in such a scenario, a sufficient causal link between that publication and the answer provided. (198) Besides, the right of reproduction under Article 2 of the InfoSoc Directive covers indirect copying as well. (199)

196. As to the second condition, the answer provided by the LLM-powered online chatbot will be regarded as a reproduction (within the meaning of Article 2 of the InfoSoc Directive) of the press publication reused through the RAG process where the protected elements that make up that publication are recognisable in that answer, through a side-by-side comparison. (200)

197. For the purposes of that comparison, one needs to know what the press publishers' right under Article 15(1) of the CDSM Directive protects specifically in a given press publication. An examination of the scope of the protection under that right is thus required.

198. Several clues are contained in the letter of the CDSM Directive in that regard. On the one hand, the subject matter of the protection under the press publishers' right, namely the press publication, is defined, in Article 2(4) of that directive, as 'a collection ... of literary works of a journalistic nature ... which ... has the purpose of providing ... information related to news ...'. On the other hand, recital 57 of that directive states that that right does not cover the 'mere facts reported' in such a publication.

199. It follows from a combined reading of those two statements that the press publishers' right protects the *specific expression of information related to news* in articles contained in a given collection. In other words, it grants the publisher exclusivity over the *form of expression in those articles* (namely the way in which those articles are arranged, structured and phrased), *not the facts* reported therein.

200. That interpretation is consistent with the objective of the press publishers' right, which is to protect the investment deemed to have been made by those publishers in the collection, verification, analysis and, ultimately, presentation of facts in press publications. (201) By contrast, the facts themselves are not a reflection of any investment made by those publishers: they existed before they were published. Moreover, it is consistent with the fundamental limit imposed by the freedom to impart and receive information under Article 11 of the Charter with respect to such a right. Under that freedom, in a democratic society, no one should be granted an exclusive right over information as such. (202)

201. That being clarified, I recall that the press publishers' right protects not only the entirety of a press publication, but also 'parts' thereof. This means that the right may subsist in a single article within the publication, or even in an excerpt from such an article. However, pursuant to Article 15(1) of the CSDM Directive, that right does not subsist in 'individual words' featured in that publication or 'very short extracts' from the latter. Recital 58 of that directive also links that to the investment-protection objective pursued by that right: while 'uses of parts of press publications have also gained economic relevance', that of those individual words or very short extracts 'may not undermine the investments made' by the publisher. Mostly, such a *de minimis* threshold is, again, essential to safeguarding the freedom to impart and receive information. Extending the right in question to such individual words and very short extracts could result in the monopolisation of the information itself. (203)

202. In my view, it follows from all the foregoing considerations that the second condition is met where at least part of a press publication, such as the form of expression in a news article thereof (rather than the mere facts reported therein), is recognisable in the answer provided by the chatbot, in a manner which goes beyond the *de minimis* threshold referred to in the preceding point. In line with the investment-protection objective under Article 15(1) of the CDSM Directive, the decisive factor is whether the answer in question constitutes a *direct substitute* for the publication, because a user, by reading that answer, can dispense with accessing the original article entirely, thus undermining the publisher's investment. (204)

203. Where the form of expression in such an article is *replicated verbatim* in a chatbot's answer, to such an extent that that answer constitutes a direct substitute for that article, a finding of reproduction is obvious. Nevertheless, while that has occurred in the past, (205) it rarely happens nowadays. When fed a part of a news article through RAG, an LLM usually does not replicate it. Rather, the model *rephrases* the article's content, on the basis of its parameters. (206) It is argued that Google applies output filters to ensure that the answer generated by Gemini does not match a text retrieved through RAG. (207)

204. The example given by Like Company reflects those considerations. As Google and the Commission point out, a side-by-side comparison of the article for which Like Company claims protection and the answer provided by the Gemini chatbot reveals that, while the two contained – essentially – the same facts, (208) that answer used different terms. Naturally, some individual words are the same ('lake', 'dolphins', and so on) but, beyond that, a single sentence from the press publication (namely, the first sentence) appears in that answer. Thus, contrary to what the referring court assumes in the first and fourth questions, such an answer can hardly be regarded as 'identical' to that publication.

205. Nevertheless, contrary to what Google and the Commission submit, that is not decisive in itself. As I have already stated above, the broad interpretation of the concept of reproduction generally followed by the Court is not limited to *verbatim copying*. That should also be the case with regard to press publications. Otherwise, the press publishers' right would be too easy to circumvent, in an age where LLMs may easily *paraphrase* rather than *replicate*, as Like Company and the French, Italian, Danish, Spanish, Polish, Hungarian and Greek Governments argue. Again, the crucial factor should instead be whether, despite any rephrasing, the form of expression in an article forming part of a press publication reused through RAG remains recognisable in the answer provided by the chatbot. (209)

206. At the risk of repeating myself, it is not enough that the answer contain the same facts as the article reused. Indeed, the press publishers' right does not subsist in mere facts. The specific form of expression of those facts in that article must be recognisable. Furthermore, since news articles may be very different in their nature, from long, in-depth analysis to purely informative reports of events, it seems to me that, inevitably, the more distinctive such an article, the more recognisable it may be in another content. Articles consisting in long, in-depth analysis are, by their nature, more recognisable than purely informative reports of events. (210) Assuming that a purely informative article of the sort can be protected under the press publishers' right, (211) in such an article, the information reported and the expression used are indissociable (212) and, as a result, one such article is virtually undisguisable from another such article. (213) That being clarified, I wish to draw a distinction between two scenarios.

207. On the one hand, where the answer provided to the user by the chatbot not only reports the same facts but also reflects a distinctive arrangement, structure and general phrasing of an article which forms part of a press publication, the latter could be recognisable in that answer despite some degree of rephrasing. If the similarity is so substantial that a user, by reading that answer, can dispense with accessing the original article entirely, such an answer constitutes a *direct substitute* for the latter, thereby undermining the publisher's investment. (214)

208. On the other hand, where the answer provided to the user by the chatbot reports the same facts as the article forming part of a press publication which has been reused through the RAG process, but the arrangement, structure and phrasing of that article is not recognisable therein (because, for instance, that form is entirely undistinguishable in the first place, the answer consists in a short summary of the factual content presented differently, or that answer consists in a general outline of information relating to the news, collected not only from various sources but also restructured by the LLM), this precludes a finding that that publication has been reproduced in that answer. (215) In such a scenario, the chatbot only extracted facts from the original article and, again, the press publishers' right does not subsist in mere facts. While that answer may, because it reports the same facts as that article, *compete* to some extent with the latter and divert some users away from it, the press publishers' right was never meant to protect those publishers against competition per se. Extending the scope of that right to such a chatbot answer would, essentially, be tantamount to granting press publishers a monopoly over the facts reported in their publications, contrary to the intention expressed in recital 57 of the CDSM Directive and the freedom to impart and receive information.

209. Admittedly, there may be a grey area between those two scenarios. Whether an answer provided by a chatbot is to be associated with one or the other should be assessed on a case-by-case basis, taking into account not only the investment-protection rationale of the press publishers' right, but also the freedom to extract mere facts from press publications.

(2) *Whether the chatbot makes the press publication at issue available to the public*

210. Turning now to the right of making available to the public laid down in Article 3(2) of the InfoSoc Directive, I would recall that, pursuant to the definition given by the Court in its case-law, that right covers the act of 'giving access' (216) to a protected work or other subject matter to a 'public', that is, an 'indeterminate' and 'fairly large' number of persons, in such a way that they may access that content from a place and at a time individually chosen by them. (217) Furthermore, where such a subject matter has already been made available online by the rightholder (or with his or her authorisation), a secondary act of making available of that subject matter by a third party requires a separate authorisation if 'specific technical means' are used or, failing that, a 'new public' is addressed, that is to say a public that was not initially taken into account by the rightholder. (218)

211. Accordingly, clarification is required as to whether an LLM-powered online chatbot gives access to the press publications it retrieves from the publishers' websites to a public (i). In the affirmative, one must also determine whether that act of making available uses specific technical means, or addresses a new public (ii).

Whether the chatbot gives access to the press publication at issue to a public

212. As the German Government observes, an LLM-powered online chatbot would give access to a press publication to a 'public' if all the users of the chatbot (clearly, an indeterminate and fairly large number of persons (219)) could access that publication, through the chatbot, from a place and at a time individually chosen by them.

213. In that respect, the fact that access is not as easy and reliable as on a traditional, static website – and, in particular, the fact that users have to query the chatbot in order to get access to the publication – is not decisive. Nevertheless, in my view, such a publication could only be regarded as being made available to the public through the chatbot if all or part of that publication were *systematically* perceptible (in other words, 'recognisable') to the point of direct substitution (220) in the answers provided by that chatbot to related user queries. (221)

214. As Google and the Estonian Government emphasise, that seems rather unlikely in the light of the technology involved. Usually, answers provided by LLM-powered online chatbots are non-deterministic. They depend on the underlying LLM's parameters, the particulars of the user query, and other variables, such as the history of conversations between the chatbot and that user. The Web resources accessed by that chatbot may also vary on a case-by-case basis. Overall, there is a significant degree of randomness involved. Nevertheless, the scenario envisaged in the previous point could happen as a result of an intentional or negligent design of the RAG process used by a chatbot. In any event, that would need to be demonstrated by rightholders, in each individual case, with sufficient examples of answers provided by the chatbot concerned.

(i) *Whether specific technical means are employed, or a new public is addressed*

215. Assuming that an LLM-powered online chatbot gives its users access, through its answers, to the whole or part of a press publication that it retrieves from the publisher's website, it remains to be determined whether such access is given using specific technical means, other than those previously used by that publisher or, failing that, that access is thus given to a new public, that is to say, to a public that was not already taken into account by that publisher when the latter initially made the publication available on its website.

216. In my view, as the Commission argued at the hearing, the hypothetical making available of a press publication to the public through an LLM-powered online chatbot would use different technical means than those previously used by the publisher to make the publication available on its website. Admittedly, as Google argues, both acts of making available rely on the internet. Nevertheless, the way in which a chatbot may give access to the publication in question differs quite radically from the transmission of that publication from server to browser on that network. I would recall that, when a chatbot retrieves a press publication, it integrates it into the prompt fed to the LLM, and the information is then transmitted to the user *through the model's predictive generational process* which, arguably, qualifies as specific technical means. (222)

(c) ***Whether the acts of use of the press publication are carried out by the chatbot provider***

217. By the second part of its first question and its fourth question, the referring court touches on the issue of whether the relevant acts of use of a press publication made 'by' (or, rather, by way of) an LLM-powered online chatbot, as discussed in the preceding sections, should be regarded as carried out by the chatbot provider, given that those acts were made (i) at the initiative of the user, who expressly asked the chatbot to reuse the publication in question (223) and (ii) 'by' that chatbot rather autonomously, using RAG and the LLM's generative capabilities.

218. This is, yet again, a delicate and debated issue. (224) With respect to copyright generally, the answer to that issue will determine whether the user, the chatbot provider, or both, is or are liable for the infringing acts committed by way of the chatbot. However, that issue is even more fundamental with respect to the press publishers' right under Article 15(1) of the CDSM Directive. I would recall that that right covers the use of press publications 'by information society service providers' (here, Google with regard to the Gemini chatbot) *only*. By contrast, private or non-commercial uses of press publications by individual users are excluded from its scope altogether. Accordingly, the acts of use of press publications discussed in section C of this Opinion are only actionable if they can be regarded as carried out *by the chatbot provider*.

219. On the one hand, given that an LLM-powered online chatbot may be used for very different purposes, and that a certain use of that chatbot occurs at the initiative of an individual user, where such a user directs the chatbot to carry out a copyright-relevant act, the 'primary' liability for the carrying of that act should generally lie with the user concerned, as Google submits. (225)

220. If any copyright-relevant act initiated by users were attributed to the chatbot provider, an excessive burden of liability would be borne by the latter. That provider could not prevent, ahead of time, any possible use that could be made of the chatbot, and yet it would be strictly liable for any resulting infringement. A strict liability of that kind would have a dissuasive effect on the development and use of LLMs and AI chatbots, contrary to the objective of innovation generally pursued by the EU legislature in that regard and the developers' freedom to do business under Article 16 of the Charter. That could also be detrimental to the freedom to impart and receive information online, as guaranteed under Article 11 of the Charter.

221. On the other hand, it does not follow that the chatbot provider should be relieved of *any* liability for the copyright-relevant acts carried out by way of that chatbot. In accordance with the case-law of the Court, such acts should be regarded as also being carried out by that provider where the latter deliberately intervenes to facilitate such acts. (226) The provider's deliberate intervention could stem from the very design of the chatbot services, where the carrying of such acts is inherent to the business model of the provider. (227) It could also stem from the fact that that provider has refrained from putting in place the appropriate measures that can be expected from a reasonably diligent operator in its situation in order credibly and effectively to counter copyright infringements being committed by way of the chatbot. (228)

222. As regards the second scenario, as the Spanish, Danish and Hungarian Governments point out, when one designs and has control over a technology as advanced as an LLM-powered online chatbot, one must generally assume the risks inherent to that technology. The reproducing and/or making available of protected content through the chatbot is one such risk. Furthermore, the provider draws revenue from the chatbot. Thus, such a provider can be expected to take appropriate steps to mitigate that risk. The chatbot provider should, in particular, programme the bots used during the RAG process to comply with measures put in place by press publishers in order to restrict access to their publications, and tailor that process in such a way that the answers provided by the chatbot are sufficiently removed from any specific publication retrieved (through output filters, and so on). (229) Where such measures have been put in place, it would be reasonable, and commensurate to the objective pursued by the press publishers' right, to absolve the chatbot provider of liability for occasional acts of reproduction or making available of such a publication in the chatbot's answers. By contrast, where the provider has not put in place such measures, the liability of the provider would be warranted.

(d) ***Whether press publishers implicitly consent to the use of their publications to ground LLM-powered online chatbots***

223. A last subject needs to be addressed, as it was discussed before the Court. It follows from the previous sections that, where an LLM-powered online chatbot retrieves a press publication online, reuses it to generate an answer to a user query, and provides that answer to the user, several acts of use of that publication falling under Article 15(1) of the CDSM Directive may occur. Where that is the case, those acts cannot, in principle, be carried out without the prior authorisation of the press publisher concerned. (230)

224. However, Google submits, in essence, that, where a press publisher leaves its press publications freely accessible on its website (that is, it does not put in place measures restricting that access, such as a paywall or measures designed to prevent web scraping by bots), as Like Company appears to have done in the main proceedings, that publisher should be regarded as having *implicitly consented* to the reuse of those publications to ground the answers provided by LLM-powered online chatbots.

225. I do not think that is the case.

226. Admittedly, the Court has, in the past, endorsed such a logic of implied consent. (231) In particular, in its judgment in VG Bild-Kunst, (232) with respect to the right of 'communication to the public' under Article 3(1) of the InfoSoc Directive, the Court ruled that, where a rightholder makes his or her work freely accessible to the public on a given website, without making use of effective technological measures restricting that access (as envisaged in Article 6(4) of that directive), that rightholder is deemed to have implicitly consented to the further communication of the work on other websites through hyperlinks by third parties. The rightholder can only avoid such a finding of implied consent by putting in place such measures. (233)

227. Nevertheless, unless the exclusive rights are to be deprived of all their effectiveness, only in certain, quite specific circumstances can it be reasonably assumed that, because a rightholder has not prevented, through technical means, a certain use of his or her protected material by a third party, he or she has implicitly consented to that use. That assumes, at the very least, that the rightholder concerned was 'actually and individually' informed of that use and that he or she would have simple means to prevent such use if he or she wished to do so. Failing that, the absence of such measures cannot possibly be construed as implied consent. (234)

228. Here, it is far from clear that the providers of LLM-powered online chatbots 'actually and individual informed' press publishers that they intended to use their press publications to ground those chatbots. While many publishers discovered that use on their own, some publishers may still not be informed of that use at all, and the precise nature of the use – how a publication is selected, and so on – is still not entirely clear. In addition, the measures that publishers can put in place to restrict access to their publications are not *that* simple. First, the putting in place of a paywall is a significant commercial choice. Publishers may well choose not to put in place such a measure because they prefer collecting advertising revenue from traffic more broadly, not because they implicitly consent to their publications being reused by third parties. Secondly, the putting in place of measures designed to prevent web scraping by bots, such as that carried out in the context of RAG, is not as easy as it sounds. Indeed, as argued before the Court, it appears that those measures do not always adequately distinguish between the purposes for which a website is accessed by bots. Indeed, it appears that the bots which scrape websites for the purposes of grounding the answers provided by LLM-powered online chatbots are not fundamentally different to those accessing those websites for the purposes of *indexing in search engines*. In the case of Google, it seems that the same bots serve both functions. (235) Accordingly, as Like Company explained at the hearing, using a protocol such as 'Robots.txt' or other measures designed to block those bots may affect the display of the website concerned in search engines. (236) Thus, while a publisher may not adopt such measures in order to retain visibility in those search engines (and possibly implicitly consent to the use of its publications for that purpose), it cannot reasonably follow that he or she implicitly consents to the use of those publications to ground chatbots. (237)

(e) ***The suggested interpretation ensures a fair balance between the rights and interests at issue***

229. To sum up, where, while processing a user query, an LLM-powered online chatbot retrieves all or part of a press publication from a press publisher's website, reuses that publication to generate an answer, and provides that answer to the user, relevant acts of reproduction of the publication at issue occur if that chatbot has accessed that publication unlawfully, in particular by circumventing effective technical measures designed to restrict access to, or the scraping of, that publication, or if the form of expression in all or part of that publication is recognisable in the answer provided to the user by the chatbot, to such an extent that the user, by reading that answer, can dispense with accessing the original publication entirely. An act of making available to the public of the press publication at issue occur if the form of expression in all or part of that publication is systematically recognisable, to the extent described above, in the answers provided by the chatbot. (238) However, those potential acts of reproduction or making available are actionable only if they can be regarded as carried out by the chatbot provider, which is the case where such acts are inherent to that provider's business model, or where that provider has failed to put in place the appropriate measures that can be expected from a reasonably diligent operator in its situation in order credibly and effectively to counter the carrying out of such acts by way of the chatbot.

230. In my view, the suggested interpretation of the scope of the press publishers' right under Article 15(1) of the CDSM Directive, and of the temporary copy exception under Article 5(1) of the InfoSoc Directive, ensures a fair balance between all the rights and interests at issue.

231. In connection with those rights and interests, I would observe that, on the one hand, press publishers have a claim to property with respect to the use of their press publications under Article 17(2) of the Charter. Furthermore, ensuring that they receive adequate revenue from that use contributes to the viability of those publishers and, indirectly, to the freedom to impart information under Article 11 of the Charter. Ensuring that news content providers may draw revenue from these new forms of derivative use of their press publications also contributes to them having the means to continue producing quality content. This, in turn, contributes to media pluralism and diversity online and, ultimately, also benefits the general public's fundamental right to receive information.

232. On the other hand, providers of LLM-powered online chatbots using RAG have a claim under the freedom to do business guaranteed under Article 16 of the Charter. Such chatbots, and RAG, are significant innovations that required considerable development. They facilitate the distribution of – and, therefore, access to – information in the 'ocean of data' that is the World Wide Web. As such, those pieces of technology also contribute, in a different way, to the general public's fundamental right to receive information online.

233. In that delicate context, the interpretation suggested in this Opinion provides a reasonable middle ground between those competing claims. Rather than outright restricting or freely permitting the use of press publications to ground LLM-powered chatbots, it focuses, more specifically, on what the chatbot ingests and what it produces. Indeed, not all chatbots services are equivalent and to be regulated the same way.

234. On the one hand, online chatbots which lawfully retrieve content online and reuse it in an 'unrecognisable' way to provide information on a given topic should be able to operate freely. Such a 'breathing space' fosters the freed

1 Original language: English.

2 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, p. 92) ('the CDSM Directive').

3 Directive of the European Parliament and of the Council of 22 May 2001 (OJ 2001 L 167, p. 10).

4 As the description of the technology is somewhat limited in the request for a preliminary ruling (see, in that respect, point 32 below), this Opinion incorporates information from the submissions of the interested parties and publicly available documents (referred to where relevant). I wish to underline that, since this Opinion has been written by a lawyer and not by a computer scientist, matters as complex as machine learning, LLMs, and LLM training are inevitably presented in a vulgarised (some might uncharitably say 'simplified') manner.

5 Google, 'An Overview of the Gemini app' (available at: https://gemini.google/gb/overview/?hl=en-GB). Other famous AI 'chatbots' include OpenAI's ChatGPT and Mistral's Vibe.

6 There are apparently several LLMs within the Gemini family. Other famous LLMs include OpenAI's GPT and Anthropic's Claude. I would point out that there are various kinds of AI models, functioning in different ways, and capable of generating several types of content. Nonetheless, I will focus the present Opinion on LLMs and text generation.

7 Machine learning is a field of computer science consisting in designing computer systems capable of automatically learning and improving based on data or experience, without relying on explicitly programmed rules (see, in general terms, Le Cun, Y, *Quand la machine apprend – La révolution des neurones artificiels et de l'apprentissage profond*, Odile Jacob, Paris, 2019).

8 Indeed, a LLM is essentially a very large and complex *mathematical function*. See, for more details, points 38 to 41 below.

9 As explained by the United States Copyright Office 'Given the line: "it was the best of times, it was the worst of times, it was the age of wisdom, it was the age of ..." many would be able to guess that the next word is "foolishness". Even if one is not familiar with A Tale of Two Cities, the context indicates that the next word is likely to be an antonym of "wisdom". This is not an unusual task for humans – we can all sometimes finish another's sentences. This task can also be mathematically [modelled]. A statistical model of language can be represented by the probability of the next word given all the preceding words or "context". By using a model to select a probable next word based on context, and then repeating the process, an AI system can take a short prompt and generate a continuing stream of language' (see US Copyright Office, 'Copyright and Artificial Intelligence – Part 3: Generative AI Training', May 2025 (accessible at https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf), p. 6). See also European Union Intellectual Property Office (EUIPO), 'Development of Generative Artificial Intelligence from a Copyright Perspective', 12 May 2025 (available at: https://www.euipo.europa.eu/en/publications/genai-from-a-copyright-perspective-2025), pp. 26 to 32.

10 I would point out that LLMs are capable of performing a wide range of tasks and can be integrated into a variety of AI systems, beyond chatbots.

11 Google continuously explores the World Wide Web with its 'web crawlers' to discover webpages and index them by collecting and storing, on its servers, certain information found on those webpages in a database. That index was initially created for the purposes of the Google Search engine. When an internet user types a query in Google Search, it allows the search engine to quickly find, and display on the Google Search result page, matching webpages (see Google, 'In-depth guide to how Google Search works', accessible at: https://developers.google.com/search/docs/fundamentals/how-search-works).

12 An internet bot is a software application that can run an automated task, in particular the retrieval of a content from a specified URL.

13 See, inter alia, Google, 'Grounding with Google Search' (available at https://ai.google.dev/gemini-api/docs/google-search). See also, generally, Thouvenin, F., 'Retrieval-Augmented Generation (RAG) – eine urheberrechtliche Einordnung', ZUM 2026, pp. 284 to 296, spec. pp. 285 and 286; MacDowall, F., 'What is RAG?', Austrian Law Librarian, 32(3), pp. 94 to 96; Lewis, P., et al., 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks'; and Oche, A.J., et al., 'A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions', 28 July 2025.

14 The latter is defined, in Article 2(4) of the CDSM Directive, as a 'collection composed mainly of literary works of a journalistic nature, but which can also include other works or other subject matter, and which: (a) constitutes an individual item within a periodical or regularly updated publication under a single title, such as a newspaper or a general or special interest magazine; (b) has the purpose of providing the general public with information related to news or other topics; and (c) is published in any media under the initiative, editorial responsibility and control of a service provider.'

15 That right over press publications is granted to press publishers independently from the copyright that may subsist in the underlying press articles, as literary works (see, in that respect, judgment of 16 July 2009, Infopaq International, C-5/08, 'the judgment in Infopaq', EU:C:2009:465, paragraph 37). While copyright is granted to the 'author'– that is, the journalist who wrote the article in question – that right is often transferred to the publisher by a contractual agreement. In any event, as it confirmed during the hearing, Like Company does not claim copyright protection in the main proceedings.

16 See recital 54 of the CDSM Directive.

17 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) (OJ L, 2024/1689) ('the AI Act').

18 See, inter alia, judgment of 16 April 2026, European Lotto and Betting and Deutsche Lotto- und Sportwetten (C-440/23, EU:C:2026:299, paragraph 49 and the case-law cited).

19 See, by analogy, judgment of 6 October 2015, Consorci Sanitari del Maresme (C-203/14, EU:C:2015:664, paragraph 29).

20 Such a portal could conceivably be regarded as a 'collection composed mainly of literary works of a journalistic nature [or] other works or other subject matter' (as that condition is inherently loose), which (a) 'constitutes an individual item within a periodical or regularly updated publication under a single title' (since that does not exclude regularly updated information portals) and (b) 'has the purpose of providing the general public with information related to news or other topics' (that criterion being, again, particularly broad). This very broad definition nonetheless begs certain questions with respect to articles such as the ones published be Like Company (see footnotes 210 and 211 below).

21 In particular, LLMs and chatbots are regulated differently under the AI Act. On the one hand, an LLM is classified as a 'general-purpose AI model' (GPAI) (defined in Article 3(63) as 'an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless ... and that can be integrated into a variety of downstream systems or applications ...') and subject to dedicated rules. On the other hand, as recital 119 indicates, an LLM-powered online chatbot is classified as an 'AI system' (defined in Article 3(1) as 'a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments') and, more specifically, as a 'general-purpose AI system' (defined in Article 3(66) as 'an AI system which is based on a general-purpose AI model and which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems'), and subject to different rules.

22 See, inter alia, judgments of 29 July 2024, Banco BPN/BIC Português and Others (C-298/22, EU:C:2024:638, paragraph 33), and of 3 March 2021, Poste Italiane and Agenzia delle entrate – Riscossione (C-434/19 and C-435/19, EU:C:2021:162, paragraph 77).

23 See judgments of 20 December 2017, *Asociación Profesional Elite Taxi* (C-434/15, EU:C:2017:981, paragraphs 26 to 29), and of 29 July 2024, Banco BPN/BIC Português and Others (C-298/22, EU:C:2024:638, paragraphs 34 and 35).

24 See, by analogy, my Opinion in Asociación Profesional Elite Taxi (C-434/15, EU:C:2017:364, point 40).

25 A model is usually trained on data corresponding to the outputs it will produce (see EUIPO, footnote 9, op. cit., p. 30).

26 The data in question may be in the range of *millions*, or even *billions* of texts. Indeed, most LLM developers believe that the more the data available for training, the better the model will perform (see US Copyright Office, footnote 9, op. cit., pp. 9 to 10 and references).

27 To train Claude, Anthropic notoriously collected countless hard copies of books, of which it tore and digitalised every page (see United States District Court, Northern District of California, 23 June 2025, *Bartz et al. v. Anthropic PBC*, C 24-05417 WHA, pp. 2 and 3). Other methods include downloading pre-existing databases, or using data obtained from third parties (for instance, publishers) through licensing. In addition, companies like Meta have used data collected from users of their other services to train their models. Data created by AI is also used for that purpose (see US Copyright Office, footnote 9, op. cit., pp. 13 and 14; EUIPO, footnote 9, op. cit., pp. 130 to 144).

28 Essentially, those bots are programmed to send requests to servers to access the hosted content and send a copy to the collecting company's servers. Such 'web scraping' allows the automated collection of large amounts of data from numerous sources simultaneously, with minimum human intervention (see US Copyright Office, footnote 9, op. cit., pp. 13 and 14; EUIPO, footnote 9, op. cit., p. 138). I underline that, beyond the collection of data for model training, such internet bots are used for various purposes on the World Wide Web, including the indexing of webpages in search engines and the 'grounding' of the answers provided by chatbots (see point 13 and footnote 11 above).

29 The performance of models depends not only on the *quantity*, but also on the *quality* of the training dataset. Low quality documents freely available on the Web may contain poor or even incorrect language and other errors, which may lead to inapposite 'learning' and, eventually, poor output. Conversely, literary works which feature in books, scientific journals, newspapers, and so on, are considered high-quality sources for training, as these contain proper expression and editing (see US Copyright Office, footnote 9, op. cit., pp. 11 to 13 and references).

30 I would recall that a LLM is essentially a very large and complex *mathematical function*. Accordingly, it cannot 'learn' from text directly. Rather, such text needs to be converted into a *numerical format* (see US Copyright Office, footnote 9, op. cit., pp. 5 to 7; and EUIPO, footnote 9, op. cit., pp. 145 to 147).

31 During the 'training loop', the model deduces the importance of different tokens in a given sequence in relation to one another (for instance, numerical IDs corresponding to the words apples, pears and bananas will most likely be found in the proximity of those corresponding to fruit, food, etc.). Then, the model assigns to each token one or multiple numerical 'vector embeddings', designed to represent the statistical-semantic characteristics of the respective tokens in a 'latent space', which is a sort of abstract, mathematical map of the language. The 'embedded' tokens are represented thereon close or far from one another, depending on their semantic relationship in different contexts (see, inter alia, US Copyright Office, footnote 9, op. cit., pp. 5 to 7 and references; EUIPO, footnote 9, op. cit., pp. 145 to 147; Microsoft, 'Understand tokens' (available at https://learn.microsoft.com/en-us/dotnet/ai/conceptual/understanding-tokens).

32 I underline that a model does not 'learn' or, a fortiori, 'understand' language as a human does. A model automatically predicts sentences based on *probabilities*, deduced from the prior observation of countless such sentences. Those probabilities merely *reflect* grammatical rules and social usage of language, so that the model can *mimic* human language, without having any intellectual comprehension of it (see Sag, M., 'Fairness and Fair Use in Generative AI', *92 Fordham L. Rev. 1887 (2024)*, spec. p. 1908).

33 Ordinary users do not interact with the model directly. An AI system such as, in this instance, a chatbot, serves as an interface between the LLM and those users.

34 Again, an LLM does not really 'understand' a user query. Rather, it splits the query into tokens, to which the usual 'numerical IDs' are associated. The resulting sequence of numbers is then processed in a strictly mathematical way: the model compares the vectors associated to those numerically represented tokens with the vectors that appear in its parameters to identify a statistical proximity, from which it derives a probabilistic answer.

35 I underline that the input data fed into an LLM as a prompt *is not added to the model*. The model processes that prompt for the purposes of generating an output, and discards it afterwards. Nevertheless, a model's parameters are sometimes fine-tuned by developers after deployment, in particular in the light of user feedback.

36 Again, the model predicts, in reality, the sequence of tokens, numerically represented, based on a comparison of the proximity or distance of the corresponding numerical vectors in the latent space. The sequence of numbers is then converted back into standard text when the output is provided to the user.

37 See Wikipedia, entry 'Gemini (Language Model)' (available at: https://en.wikipedia.org/wiki/Gemini_(language_model)) and references.

38 See, in that respect, footnote 208 below.

39 See Keller, P., 'Do AI models dream of dolphins in lake Balaton?', Kluwer Copyright Blog, 28 May 2025 (available at: https://legalblogs.wolterskluwer.com/copyright-blog/do-ai-models-dream-of-dolphins-in-lake-balaton/).

40 Although the press publishers' right does not grant news publishers any protection against model training, that does necessarily leave those publishers without any means of redress in that regard. The condition of online use does not exist with respect to other rights, in particular copyright, which may subsist in the news articles that they publish (see footnote 15 above). Thus, that right may be actionable in that respect (see Section B.2 below).

41 See, inter alia, judgment of 14 April 2026, Pelham (Concept of 'pastiche') (C-590/23, EU:C:2026:290, paragraph 34 and the case-law cited).

42 See, for that definition, Article 1(1)(b) of Directive (EU) 2015/1535 of the European Parliament and of the Council of 9 September 2015 laying down a procedure for the provision of information in the field of technical regulations and of rules on Information Society services (codification) (OJ 2015 L 241, p. 1), to which Article 2(5) of the CDSM Directive refers.

43 See recital 54 of the CDSM Directive.

44 During training, the computer infrastructure used is sometimes not even connected to the Internet. The fact the LLM developers may have scraped the publication from the World Wide Web is also irrelevant. In line with the definition of 'online use' set out above, what matters is not the *source* of the publication, but the *purpose* of the use. The *travaux préparatoires* of Article 15(1) of the CDSM Directive support that interpretation. In the Commission's initial proposal, the press publishers' right was to cover the 'digital use' of their press publications (see recitals 32 and 34 and Article 11 of the proposal), which could have covered any use of such a publication in digital form, including for the purpose of model training. However, during the legislative process, the scope of the right was narrowed to target the 'online use' of those publications only. By contrast, many provisions of the CDSM Directive cover the use of works and other subject matter in digital form generally (see, for instance, Article 4 and Article 5(1)).

45 Nevertheless, *other* acts of online use of a press publication may occur, 'downstream', as part of the provision of chatbot services. Those acts, which will be examined in Section C below, are *distinct* from the use of such a publication during LLM training.

46 Furthermore, a trained LLM may be integrated into various AI systems and, accordingly, used for many different purposes or tasks. It would be contrary to legal certainty for acts of use occurring during training to be 'requalified' as 'online use' simply because the LLM, after deployment, happens to be integrated into an online chatbot. Besides, the developers of the LLM and the business offering the online chatbot powered by that LLM will often be different, unrelated entities.

47 I generalise since the use of press publications for the purposes of LLM training is, in any case, not actionable under Article 15(1) of the CDSM Directive and the right of reproduction and the TDM exception apply across the board to works protected by copyright and other subject matter protected by the related rights.

48 Regulation (EC) No 864/2007 of the European Parliament and of the Council of 11 July 2007on the law applicable to non-contractual obligations (OJ 2007 L 199, p. 40) ('the Rome II Regulation').

49 With respect to copyright, see Article 5(2) of the Berne Convention for the Protection of Literary and Artistic Works (Paris Act of 24 July 1971), as amended on 28 July 1979 ('the Berne Convention'); and judgment of 22 January 2015, Hejduk (C-441/13, EU:C:2015:28, paragraph 22). With respect to the related-rights, see judgment of 14 July 2005, Lagardère Active Broadcast (C-192/04, EU:C:2005:475, paragraph 46).

50 See, to that effect, judgment of 21 December 2011, Air Transport Association of America and Others (C-366/10, EU:C:2011:864, paragraph 125); Opinion of Advocate General Darmon in Joined Cases Ahlström Osakeyhtiö and Others v Commission (89/85, 104/85, 114/85, 116/85, 117/85 and 125/85 to 129/85, not published, EU:C:1988:258, points 19 to 59).

51 See judgments of 16 June 1998, Racke (C-162/96, EU:C:1998:293); and of 14 July 2005, Lagardère Active Broadcast (C-192/04, EU:C:2005:475, paragraph 46).

52 See, by analogy, judgment of 19 April 2012, Wintersteiger (C-523/10, EU:C:2012:220, paragraph 36). See also Poesen, M. 'Generative Artificial Intelligence and Private International Law in Copyright Infringement Cases: Some Lessons from GEMA v OpenAI', *Journal of Private International Law*, Vol. 22, Issue 3, forthcoming December 2026; and Azzi, T., El Hage, Y., *Rapport de mission sur la loi applicable, en vertu des règles de droit international privé, aux modèles d'intelligence artificielle générative commercialisés dans l'Union européenne*, CSPLA, 15 December 2025, pp. 26 to 28. In a dispute between Getty Images and Stability AI, this state of affairs led to an extremely long, factually dense and partially speculative discussion of where Stability carried out the training of its model, including the location of the servers used at the time, of the staff involved, their personal laptops, and so on (see High Court of Justice (United Kingdom)), 1 December 2023, *Getty Images (US) Inc and others* v *Stability AI Ltd* [2023] EWHC 3090 (Ch), paragraphs 43 to 78).

53 See, for an overview of the laws adopted in various jurisdictions, de la Durantaye, K., 'Control and Compensation. A Comparative Analysis of Copyright Exceptions for Training Generative AI', *International Review of Intellectual Property and Competition Law*, Vol. 56, 2025, pp. 737-770. Accordingly, the use of works and other protected subject matter for the purposes of LLM training could very well be free under the law of country A, but restricted under the law of country B.

54 See, for the same view, Quintais, J.P., 'Copyright, the AI Act and Extraterritoriality', *Policy Brief*, The Lisbon Council, 2025, p. 13.

55 See, by analogy, judgments of 19 April 2012, Wintersteiger (C-523/10, EU:C:2012:220, paragraph 37), and of 27 September 2017, Nintendo (C-24/16 and C-25/16, EU:C:2017:724, paragraphs 103 and 108).

56 It does not mean that those rightholders would be devoid of protection entirely. They would need to seek protection under the law of the country where the training is essentially carried out.

57 See also Azzi, T., El Hage, Y., footnote 52, op. cit., pp. 30 to 34 ; and Voss, A., *Report on copyright and generative artificial intelligence – opportunities and challenges*, 25 February 2026 (2025/2058(INI)), paragraphs 16 and AH.

58 See Rättzén, M., 'Location Is All You Need: Copyright Extraterritoriality and Where to Train Your AI', *Columbia Science and Technology Law Review*, Vol. 26, No 175, 2024, pp. 252 and 253.

59 See, inter alia, in the field of competition law, judgment of 6 September 2017, Intel v Commission (C-413/14 P, EU:C:2017:632, paragraphs 42 to 48); in the field of data protection, judgment of 24 September 2019, Google (Territorial scope of de-referencing) (C-507/17, EU:C:2019:772, paragraphs 57 and 58).

60 I would recall that various rightholders have an exclusive 'right of communication to the public' of their subject matter under (inter alia) Article 3 of the InfoSoc Directive.

61 See, to that effect, judgment of 9 July 2026, Anne Frank Fonds (C-788/24, EU:C:2026:559, paragraphs 21 to 43). See also, by analogy, judgments of 12 July 2011, L'Oréal and Others (C-324/09, EU:C:2011:474, paragraphs 64 to 67); and of 18 October 2012, Football Dataco and Others (C-173/11, EU:C:2012:642, paragraphs 18 to 39).

62 See Quintais, J.P., footnote 54, op. cit., pp. 13 and 16. See also points 125 to 133 below. Furthermore, the use of works and other subject matter for the purposes of LLM training is regarded, under (EU) copyright law, as *an autonomous set of acts of exploitation* (see Section B.2.b below), distinct from the downstream acts of use that may occur through the trained model (which the European Union may fully regulate, as explained in Section C below).

63 See, for the same view, Peukert, A., 'Copyright in the Artificial Intelligence Act – A Primer', *GRUR Int.*, Vol. 73, Issue 6, 2024, pp. 497-509, spec. pp. 506 and 507.

64 That is done through a reservation of the rights for TDM, under Article 4(3) of the CSDM Directive (see Section B.2.c below).

65 See Peukert, A., footnote 63, op. cit., p. 506. The AI Act could be interpretated that way. Under Article 53(1)(c) of that instrument, developers of GPAIs, including LLMs, which are put on the EU internal market (see Article 2(1)(a) of the AI Act), must 'put in place a policy to comply with Union law on copyright and related rights, *and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of [the CDSM Directive]*' (emphasis added) and that 'regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place'. Obviously, developers must comply with EU artistic and property law in so far as it applies, in accordance with the principle of territoriality. Under that principle, when model developers access and scrape content located on servers in Europe, EU law may apply and requires developers to comply with any reservation of rights expressed pursuant to Article 4(3) of the CDSM Directive. This would also 'ensure a level playing field among providers of [GPAIs] where no provider should be able to gain a competitive advantage in the Union market by applying lower copyright standards than those provided in the Union' (see recital 106 of the AI Act).

66 See Voss, A., footnote 57, op. cit., explanatory statement, observing that 'we cannot allow AI models to be trained just anywhere in the world using European copyright-protected data only for them to be then made available in Europe.'

67 See recital 9 of the InfoSoc Directive, and judgments in Infopaq (paragraphs 40 to 43); of 24 March 2022, Austro-Mechana (C-433/20, EU:C:2022:217, paragraphs 16); and of 13 July 2023, Ocilion IPTV Technologies (C-426/21, EU:C:2023:564, paragraph 28).

68 See Opinion of Advocate General Kokott in Joined *Cases* Football Association Premier League and Others (C-403/08 and C-429/08, EU:C:2011:43, points 70 to 73 and 81). See also, by analogy, Guide to the Berne Convention for the Protection of Literary and Artistic Works (Paris Act, 1971), World Intellectual Property Organisation, Geneva, 1978 ('the Guide to the Berne Convention'), p. 54.

69 See, to that effect, judgments in Infopaq (paragraph 51); of 24 March 2022, Austro-Mechana (C-433/20, EU:C:2022:217, paragraphs 16 to 18 and the case-law cited); and of 11 September 2014, Eugen Ulmer (C-117/13, EU:C:2014:2196, paragraphs 52 and 53).

70 See judgment of 4 October 2011, Football Association Premier League and Others (C-403/08 and C-429/08, EU:C:2011:631, paragraphs 153 and 157 to 159).

71 But not always. Some datasets only contain hyperlinks referring to the content on the original servers, rather than digital reproductions of that content.

72 See Bruguière, J.-M. and Deltorn, J.-M., 'Intelligences artificielles génératives. Y a-t-il exploitation des œuvres, au sens du droit d'auteur ? N'y aurait-il pas d'autres modèles à considérer ?', *Recueil Dalloz*, 2023, p. 1657, spec. paragraphs 1 to 9; and Huang, W. and Chen, X., 'Does generative AI copy? Rethinking the right to copy under copyright law', *Computer Law & Security Review: The International Journal of Technology Law and Practice*, Vol. 56, 2025, pp. 1 to 13. In US copyright vocabulary, the term 'non-expressive use' would fit that scenario (see Sag, M., footnote 32, op. cit., pp. 1901 to 1907).

73 See Commission, proposal for a European Parliament and Council Directive on the harmonisation of certain aspects of copyright and related rights in the Information Society (COM(97) 628 final), 10 December 1997, explanatory memorandum, pp. 13, 14, 24 and 25. Nevertheless, as will be discussed in the next section, the temporary reproduction exception under Article 5(1) of the InfoSoc Directive and the TDM exception under Article 4 of the CDSM Directive may cover the digital reproductions of works and other subject matter created for the purposes of LLM training. Obviously, those exceptions would be deprived of their effectiveness (*effet utile*) if those reproductions did not fall within the scope of the reproduction right in the first place.

74 That recital states that '... the development and training of [GPAI] require access to vast amounts of text, images, videos and other data. Text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content, which may be protected by copyright and related rights. ... [The CDSM Directive] introduced exceptions and limitations allowing reproductions and extractions of works or other subject matter, *for the purpose of text and data mining*, under certain conditions. ...' (emphasis added).

75 As the Court observed in paragraph 45 of the judgment in Infopaq, it is 'through the choice, sequence and combination of [the] words [of a literary work] that the author may express his creativity in an original manner and achieve a result which is an intellectual creation'. The model 'learns' by observing countless examples of such choice, sequence and combination, represented in a numerical way (see United States District Court, Northern District of California, 25 June 2025, *Kadrey et al. v. Meta Platforms, Inc.*, 23-cv-03417-VC, pp. 23 and 24).

76 See Lebrun, T., 'L'apprentissage machine est une appropriation', *Les cahiers de propriété intellectuelle*, Vol. 30, No°3, 2018, pp. 895 to 924.

77 Thus, as Google explained during the hearing, it is not possible to reverse engineer the model parameters to discover the works on which the model was trained. Those parameters are too abstract for that purpose.

78 See judgment in Infopaq (paragraph 46); High Court of Justice, 4 November 2025, *Getty Images (US) Inc and others* v *Stability AI Ltd* [2025] EWHC 2863 (Ch), paragraphs 554 to 557, 592, 600; Sag, M., footnote 32, op. cit., pp. 1907 and 1908; Guadamuz, A., 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs', *GRUR International*, Vol. 73, Issue 2, 2024, pp. 111 to 127; Bruguière, J.-M. and Deltorn, J.-M., footnote 72, op. cit., paragraphs 5 to 7; EUIPO, footnote 9, op. cit., pp. 150 and 151.

79 See, with respect to that phenomenon, Liu, X., Mireshghallah, N., Ginsburg, J.C. and Chakrabarty, T., 'Alignment Whack-a-Mole: Finetuning Activates Verbatim Recall of Copyrighted Books in Large Language Models', 2026; Feder Cooper, A. and Grimmelmann, J., 'The Files are in the Computer: On Copyright, Memorization, and Generative AI', *Chicago-Kent Law Review*, Vol. 100, 2025; Emanuilov, I. and Margoni, T., 'Forget me not: memorisation in generative sequence models trained on open source licensed code', Zenodo, 2024; EUIPO, footnote 9, op. cit., pp. 154 and 155 and further references; US Copyright Office, footnote 9, op. cit., pp. 19 to 21.

80 See Landgericht München (Regional Court, Munich, Germany), 11 November 2025, *GEMA* v *Open AI*, 42 O 14139/24, paragraphs 165 to 190. In the dispute that gave rise to that judgment, the German collecting society GEMA sued OpenAI on behalf of music publishers and songwriters, alleging that GPT had 'memorised' at least nine German songs and could 'regurgitate' their lyrics almost verbatim when prompted appropriately. As evidence, GEMA produced screenshots of outputs that, for each of the songs, contained substantial parts of those lyrics, generated by GPT on the basis of prompts entered by GEMA's lawyer designed to force the model to 'recite' them. That court decided that those outputs did indeed constitute evidence that GPT contained reproductions of those songs, within the meaning of German legislation implementing Article 2 of the InfoSoc Directive. See, for the same approach, Landgericht München (Regional Court, Munich, Germany), 31 July 2026, *GEMA* v *Suno*, 42 O 763/25, paragraphs 262 to 279.

81 See Dornis, T.W., 'Generative AI, reproductions inside the model, and the making available to the public', *International Review of Intellectual Property and Competition Law*, Vol. 56, 2025, pp. 909 to 938; Matulionyte, R., 'Reconceptualizing the Reproduction Right in the age of AI', *International Review of Intellectual Property and Competition Law*, Vol. 56, 2025, pp. 1914 to 1949; US Copyright Office, footnote 9, op. cit., pp. 19 to 21 and further references.

82 'Overfitting', and resulting 'memorisation', is an undesired feature in LLM development, which developers try to avoid (see High Court of justice, 4 November 2025, *Getty Images (US) Inc and others* v *Stability AI Ltd* [2025] EWHC 2863 (Ch), paragraph 558; Bruguière, J.-M. and Deltorn, J.-M., footnote 72, op. cit., paragraph 5; EUIPO, footnote 9, op. cit., p. 154 and references; US Copyright Office, footnote 9, op. cit., pp. 19 to 21 and further references.

83 See, with respect to such infringing outputs, points 194 and sub. below.

84 See Huang, J., Yang, D. and Potts, C., 'Demystifying Verbatim Memorization in Large Language Models', in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* Miami, USA, 2024, pp. 10711 to 10732.

85 Pushed to the extreme, the 'memorisation as reproduction' argument could mean that, when a human 'memorises' a work by reading it repeatedly, that work is 'reproduced' as electrical patterns in his or her neurons.

86 EUIPO, footnote 9, op. cit., p. 154.

87 In fact, it would appear that 'regurgitation' is obtained through targeted prompting and not through a usual use of the model. Ibid., p. 158.

88 See, for the same view, Wietzke, A., 'Retrieval Augmented Generation – eine urheberrechtliche Betrachtung', RDi 2024, pp. 538 to 549, spec. paragraph 21; Markiewicz, R., 'Przeciw ochronie prawnoautorskiej dla tzw. memoryzacji utworów', *Zeszyty Naukowe Uniwersytetu Jagiellońskiego. Prace z Prawa Własności Intelektualne*, 2026, Vol. 2, pp. 5 to 25.

89 Furthermore, once a piece of training data has been 'memorised', it appears particularly difficult to 'untrain' the model to 'forget' it.

90 If one considered that an LLM contains 'reproductions' of its training data, within the meaning of Article 2 of the InfoSoc Directive, the model itself would fall under the rightholders' exclusive right of reproduction. Furthermore, the mere offering of the model could be seen as the making available to the public of the works 'memorised' within, falling under the right set out in Article 3 of this directive. It is also unclear whether an exception or limitation could apply in that context. The temporary reproduction exception under Article 5(1) of that directive would be excluded, since those reproductions would be permanent. The TDM exception would also be excluded, since those reproductions would not be created 'for the purpose of text and data mining'. Some authors suggest that the exception or limitation covering 'incidental inclusion of a work or other subject matter in other material', provided in Article 5(3)(i) of that directive, could apply (see, inter alia, Markiewicz, R., footnote 88, op. cit.). Should the Court ever consider 'memorisation' as 'reproduction', I would not exclude that that exception could offer a solution to the issue under discussion.

91 The removal of duplicates from a dataset is considered significantly to reduce the risk of 'memorisation' (see EUIPO, footnote 9, op. cit., pp. 155 and 156 and references).

92 Prompt and output filters may be put in place to block outputs from being generated by the LLM if they overlap significantly with a part of the training data (see, inter alia, EUIPO, footnote 9, op. cit., p. 156 and references).

93 Moreover, the three-step test must be satisfied in each case, pursuant to Article 5(5) of the InfoSoc Directive (see, for further detail, Section B.2.c.2 below).

94 See EUIPO, footnote 9, op. cit., pp. 49 and 50, and Schönberger, D., 'Deep Copyright: Up- and Downstream Questions Related to Artificial Intelligence (AI) and Machine Learning (ML)', *Zeitschrift für geistiges Eigentum (ZGE)*, Vol. 10, Issue 1, 2018, pp. 35 to 58, spec. p. 54. Recital 18 of the CDSM Directive implies that the technical reproductions created during automated data analysis processes, such as those carried out during LLM training, fall under the temporary reproduction exception.

95 See Hanseatisches Oberlandesgericht Hamburg (Higher Regional Court, Hamburg, Germany), 27 September 2024, *Kneschke v. LAION*, No. 310 O 227/23.

96 Indeed, it is not at all clear that AI developers discard them afterwards.

97 Including reproductions of press publications. Nevertheless, for the reasons explained above, the reproductions of such publications carried out for the purposes of LLM training are not actionable in any case.

98 Article 3 of the CDSM Directive provides another 'text and data mining' exception. However, it covers reproductions made by research organisations and cultural heritage institutions for scientific research only and, thus, is not relevant here.

99 Provided that the conditions of that exception and the three-step test below are satisfied in the case at hand (see Section B.2.c).

100 See, for the same view, ruling of the Hanseatisches Oberlandesgericht Hamburg (Higher Regional Court, Hamburg) in *Kneschke v. LAION*, 27 September 2024; Margoni, T. and Kretschmer, M., 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology', GRUR International, 71(8), 2022, pp. 685 to 701; Bruguière, J.-M. and Deltorn, J.-M., footnote 72, op. cit., paragraph 7; Stieper, M. and Denga, M., 'The international reach of EU copyright through the AI Act', *Beiträge zum transnationalen Wirtschaftsrecht*, 2024, p. 7 and 8; Wietzke, A., footnote 88, op. cit., paragraphs 22 and 23; Lux, H. and Noll, C.J., 'Of Books and Bytes: The Copyright Dilemma in AI Development – The European Perspective on Using Copyrighted Works for AI Training', TLJ 2024, pp. 111 to 115; European Copyright Society, *Copyright and Generative AI: Opinion of the European Copyright Society*, January 2025, p. 4.

101 By contrast, the TDM exception could not possibly cover the potential reproductions contained in the outputs generated by an LLM. These specific reproductions cannot be regarded as carried out 'for the purpose of data mining' within the meaning of Article 4(1) of the CDSM Directive.

102 See Rosati, E., *Copyright in the Digital Single Market: Article-by-Article Commentary to the Provisions of Directive 2019/790*, Oxford University Press, 2021, p. 89. By contrast, it would not allow LLM developers to keep the dataset for *longer* than required for the training process. Where reproductions of works or other subject matter are initially made by a third party for the purposes of data mining, the TDM exception does not allow that party to keep them forever and for other purposes (see, by analogy, United States District Court, Northern District of California, 23 June 2025, *Bartz et al. v. Anthropic PBC*, C 24-05417 WHA, p. 14).

103 See Dornis, T.W., 'The Training of Generative AI is not Text and Data Mining', European Intellectual Property Review (E.I.P.R.), 2025, 47(2), pp. 65 to 78; Alonso, E. and Lucchi, N., 'AI and Copyright "Hallucinations": Does the Text and Data Mining Exception Really Support Generative AI Training?', *E.I.P.R.*, 2025, 47(9), pp. 515 to 526.

104 See, inter alia, judgments of 4 October 2011, Football Association Premier League and Others (C-403/08 and C-429/08, EU:C:2011:631, paragraphs 162 and 163); of 3 September 2014, Deckmyn and Vrijheidsfonds (C-201/13, EU:C:2014:2132, paragraphs 22 and 23); and of 10 November 2016, Vereniging Openbare Bibliotheken (C-174/15, EU:C:2016:856, paragraph 50).

105 See European Commission, Commission Staff Working Document – Impact Assessment on the modernisation of EU copyright rules – Part 1, SWD(2016)301, section 4.3.1. That proposal eventually became Article 3 of the CDSM Directive.

106 See recital 18 of the CDSM Directive.

107 Ibid.

108 See European Parliament: Directorate-General for Internal Policies of the Union and Rosati, E., *The exception for text and data mining (TDM) in the proposed Directive on Copyright in the Digital Single Market – Technical aspects*, European Parliament, 2018, p. 2. See, further, Widła, B. and Ożegalska-Trybalska, J., 'Artykuł 26[2]', in Markiewicz, R. (ed.), *Prawo autorskie. Komentarz do znowelizowanych 26 lipca 2024 r. przepisów ustawy: o prawie autorskim i prawach pokrewnych; o ochronie baz danych; o zbiorowym rządzeniu prawami autorskimi i prawami pokrewnymi*, Wolters Kluwer, Warszawa, 2025, p. 152.

109 Specifically, recital 105 states that 'the development and training of [GPAI] require access to vast amounts of text, images, videos and other data. *Text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content*, which may be protected by copyright and related rights. ... [The CDSM Directive] introduced exceptions and limitations allowing reproductions and extractions of works or other subject matter, *for the purpose of text and data mining*, under certain conditions. ...' (emphasis added).

110 See, for the same view, 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*', IIC, Vol. 56, 2025, pp. 1800 to 1840, spec. pp. 1822 to 1830.

111 See, inter alia, Article 9(2) of the Berne Convention, and Article 13 of the Agreement on Trade-Related Aspects of Intellectual Property Rights, which constitutes Annex 1C to the Agreement establishing the World Trade Organisation (WTO), signed at Marrakesh on 15 April 1994 and approved by Council Decision 94/800/EC of 22 December 1994 concerning the conclusion on behalf of the European Community, as regards matters within its competence, of the agreements reached in the Uruguay Round multilateral negotiations (1986-1994) (OJ 1994 L 336, p. 1).

112 See, to that effect, judgments of 11 September 2014, Eugen Ulmer (C-117/13, EU:C:2014:2196, paragraphs 33, 34 and 47); of 22 January 2015, Art & Allposters International (C-419/13, EU:C:2015:27, paragraph 38); and of 19 December 2019, Nederlands Uitgeversverbond and Groep Algemene Uitgevers (C-263/18, EU:C:2019:1111, paragraphs 38 and 39).

113 See, to that effect, judgment of 5 June 2014, Public Relations Consultants Association (C-360/13, EU:C:2014:1195, paragraph 55); Opinion of Advocate General Trstenjak in Infopaq International (C-5/08, EU:C:2009:89, point 135); my Opinion in Vereniging Openbare Bibliotheken (C-174/15, EU:C:2016:459, point 67); and World Trade Organization, (2000), United States – Section 110(5) of the US Copyright Act, Report of the Panel (WT/DS160/R), paragraphs 6.108, 6.109, 6.112, 6.127 and 6.143.

114 I underline that the exemption is limited to such 'non-expressive' use. the TDM exception, as interpreted, cannot cover reproduct

that code, the signatories commit (a) 'to implement appropriate and proportionate technical safeguards to prevent their models from generating outputs that reproduce training content protected by Union law on copyright and related rights in an infringing manner', and (b) 'to prohibit copyright-infringing uses of a model in their acceptable use policy, terms and conditions, or other equivalent documents …'. Thus, the application of the TDM exception in a given case depends indirectly on whether the model developer satisfies that obligation.

133 See, to that effect, judgments of 3 September 2014, Deckmyn and Vrijheidsfonds (C-201/13, EU:C:2014:2132, paragraph 34), and of 29 July 2019, Funke Medien NRW (C-469/17, EU:C:2019:623, paragraph 61). See also my Opinion in VCAST (C-265/16, EU:C:2017:649, point 60).

134 In the media sector, works published on the websites of the main news outlets were intensively scraped to train LLMs (see Schaul, K., Chen, S.Y., and Tiku, N., 'Inside the secret list of websites that make AI like ChatGPT sound smart', *Washington Post*, 19 April 2023).

135 Generative AI is increasingly used, by press publishers themselves, for the drafting of such news reports (see Mezei, P., and Trapova, A., 'Robojournalism – A Copyright Study on the Use of Artificial Intelligence in the European News Industry', *GRUR International*, Vol. 71, Issue 7, 2022, pp. 589 to 602).

136 See, insightfully, United States District Court, Northern District of California, 25 June 2025, *Kadrey et al. v. Meta Platforms, Inc*, 23-cv-03417-VC, p. 29.

137 I leave aside 'deepfakes' and other societal issues which, whilst equally concerning, are less directly connected to artistic and literary property.

138 See United States District Court, Northern District of California, 23 June 2025, *Bartz et al v. Anthropic PBC*, C 24-05417 WHA, pp. 15 to 17; Sag, M., footnote 32, op. cit., pp. 1891 and 1892; Senftleben, M., footnote 115, op. cit., p. 76.; Geiger, C., and Iaia, V., 'The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative IA', *Computer Law & Security Review*, Vol. 52, 2024, pp. 1 to 9; Schönberger, D., footnote 94, op. cit., p. 55; and Stieper, M., and Denga, M., footnote 100, op. cit., pp. 8 and 9.

139 See, inter alia, judgment of 17 December 2015, Neptune Distribution (C-157/14, EU:C:2015:823, paragraph 76 and the case-law referred).

140 See, on those aspects, Lucchi, N., 'Generative AI and Copyright: Training, Creation, Regulation', European Parliament, Policy Department for Justice, Civil Liberties and Institutional Affairs, Study PE 774.095, July 2025, p. 104.

141 See, to that effect, recital 18 of the CDSM Directive. See, by analogy, judgments of 10 April 2014, ACI Adam and Others (C-435/12, EU:C:2014:254, paragraphs 20 to 58); of 5 March 2015, Copydan Båndkopi (C-463/12, EU:C:2015:144, paragraphs 75 to 78); and of 10 November 2016, Vereniging Openbare Bibliotheken (C-174/15, EU:C:2016:856, paragraphs 66 to 72). Admittedly, recital 14 of the CDSM Directive states that 'lawful access should also cover access to content that is freely available online', irrespective of the lawfulness of the source. Nevertheless, that statement concerns the 'lawful access' condition provided in the 'scientific research' TDM exception under Article 3, which may call for a more generous interpretation than Article 4(1) (see European Copyright Society, *The concept of lawful access and its implications for user freedoms in the EU copyright acquis*: *Opinion of the European Copyright Society*, July 2026, pp. 17 to 21.

142 Under Article 6(3) of the InfoSoc Directive, the concept of 'technical measures' covers any technology, device or component designed to prevent or restrict acts, in respect of protected works or other subject matter, which are not authorised by the rightholder. Such 'technical measures' are deemed 'effective' where the use of a protected work or other subject matter is controlled by the rightholders through the application of an access control or protection process, such as a copy control mechanism, which achieves the protection objective.

143 Various providers offer 'bot-management' services, designed to allow website owners to block internet bots from reaching their server (see, in that respect, EUIPO, footnote 9, op. cit., p. 243 and references).

144 Admittedly, under the first subparagraph of Article 6(4) of the InfoSoc Directive, rightholders who have put in place effective technical measures to restrict access to their content shall nevertheless 'make available to the beneficiary of an exception or limitation … the means of benefiting from [the latter] …'. This applies also in relation to the TDM exception (see Article 7(2) of the CDSM Directive). Nevertheless, Article 6(4) does not grant model developers a right to hack. On the contrary, that subparagraph requires the beneficiary of the exception to have 'legal access to the protected work or subject matter concerned'.

145 See, by analogy, judgments of 14 June 2017, Stichting Brein (C-610/15, EU:C:2017:456, paragraph 45); and of 26 April 2022, Poland v Parliament and Council (C-401/19, EU:C:2022:297, paragraph 90). I obverse that, under Measure 1.2 of the AI Code of Practice, signatories commit '(a) not to circumvent effective technological measures … that are designed to prevent or restrict unauthorised acts in respect of works and other protected subject matter' and '(b) to exclude from their web-crawling websites that make available to the public content and which are, at the time of web-crawling, recognised as persistently and repeatedly infringing copyright and related rights on a commercial scale by courts or public authorities in the European Union and the European Economic Area'.

146 See, inter alia, Dornis, T.W., and Lucchi, N., footnote 110, op. cit., p. 1829.

147 For a presentation of the various uncertainties concerning the interpretation of Article 4(3) of the CDSM Directive, see Havlikova, S., 'Technical Challenges of Rightsholders' Opt-out from Gen AI Training after Robert Kneschke v. LAION', *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, Vol. 16, Issue 1, 2025, pp. 88 to 106.

148 It is argued that, with respect to works, that opt-out mechanism is contrary to the prohibition of 'formalities' under Article 5(2) of the Berne Convention (Dornis, T.W., and Lucchi, N., footnote 110, op. cit., p. 1828). I do not think so. The Guide to the Berne Convention clarifies that the formalities prohibited under Article 5(2) are 'administrative obligations laid down by national laws such as the deposit of a copy of a work: its registration with some public or official body; the payment of registration fees, or one or more of these' (Ibid., p. 33, §5.5). It does not cover any step that the law may require of rightholders in order to be able to claim protection against certain uses. The Court itself understands Article 5(2) as not opposing a system whereby a work can be freely used for a certain purpose, unless the author expresses opposition to the latter use (see judgment of 16 November 2016, Soulier and Doke (C-301/15, EU:C:2016:878, paragraphs 50 and 51).

149 See, by analogy, Opinion of Advocate General Saugmandsgaard Øe in Poland v Parliament and Council (C-401/19, EU:C:2021:613, point 151 and references).

150 See, by analogy, judgment of 16 November 2016, Soulier and Doke (C-301/15, EU:C:2016:878, paragraphs 50 and 51).

151 See, *mutatis mutandis*, judgment of 9 March 2021, VG Bild-Kunst (C-392/19, EU:C:2021:181, paragraph 46).

152 I would recall that, under Article 53(1)(c) of the AI Act, providers of GPAIs shall, in particular, 'identify and comply with, including *through state-of-the-art technologies*, a reservation of rights expressed pursuant to Article 4(3) of [the CDSM Directive]' (emphasis added).

153 The Robot Exclusion Protocol provides internet bots with information about which URLs on a website they are allowed to access, and which URLs are restricted. It consists in a text file ('robots.txt') stored in the website root directory. It currently serves as a de facto standard for managing web crawling and scraping activities, including for model 'training' (see EUIPO, footnote 9, op. cit., pp. 173 to 176). There are, admittedly, limitations associated with that protocol. First, it does not technically prevent scraping. It merely indicates the website publisher's intention in that respect. Its implementation by bots rests entirely on the principle of good faith, and some bots have been shown to ignore robots.txt directives (ibid., pp. 182 and 183). Nevertheless, Article 53(1)(c) of the AI Act provides relief in that regard. Under Measures 1.3(1)(a) of the associated AI Code of Practice, signatories commit to employing web crawlers that follow the Robot Exclusion Protocol. Secondly, that protocol requires the website publisher to know the exact name of the crawler and update the 'robots.txt.' file accordingly file, whereas some model developers do not release such information (ibid. pp. 174, 179 and 180). Nevertheless, the AI Act helps here as well. Under Measure 1.3(4) of the AI Code of Practice, signatories commit to providing rightholders with information about, inter alia, 'the web crawlers employed' and 'their robots.txt features'. Finally, the Robot Exclusion Protocol may not easily be used to distinguish between robots and the purposes for which crawling is carried out (model training, RAG, or web search indexing), an issue that will be examined further in Section C below.

154 Google Extended, released by Google on 28 September 2023, is a tool designed to enable website publishers, through specific lines added to the 'Robots.txt' file, to prevent their content being scraped by Google's robots to train Google's models and to ground the answers provided by the Gemini Chatbot (see Section C below). It allegedly does not affect Google Search (see Google 'An update on web publisher controls', *Google Blog*, 28 September 2023; available at: https://blog.google/innovation-and-ai/products/an-update-on-web-publisher-controls/).

155 See, in that respect, Measure 1.3(2) of the AI Code of Practice.

156 See for such protocols, EUIPO, footnote 9, op. cit., pp. 183 to 229. See also, in that respect, Measure 1.3(1)(b) of the AI Code of Practice.

157 See Hanseatisches Oberlandesgericht Hamburg (Higher Regional Court, Hamburg, Germany), 27 September 2024, *Kneschke v. LAION*, No. 310 O 227/23, paragraph 99.

158 See Peukert, A., and Castets-Renard, C., 'The Copyright Chapter of the EU's Code of Practice for General-Purpose AI Models: A Commentary', *International Review of Intellectual Property and Competition Law*, Vol. 56, Issue 10, 2025, pp. 1950 to 1993.

159 See Voss, A., footnote 57, op. cit., paragraphs 13, 17 and Y.

160 Recital 107 of the AI Act specifies that this summary 'should be generally comprehensive in its scope … to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, for example by listing the main data collections or sets that went into training the model … and by providing a narrative explanation about other data sources used'.

161 See Geiger, C., and Iaia, V., footnote 138, op. cit.; Senftleben, M., 'Generative AI and Author Remuneration', *International Review of Intellectual Property and Competition Law*, Vol. 54, 2023, pp. 1535 to 1560; Geiger, C., 'Elaborating a human rights-friendly copyright framework for generative AI', *International Review of Intellectual Property and Competition Law*, Vol. 55, Issue 7, 2024, pp. 1129 to 1165.

162 I am well aware that recital 17 of the CDSM Directive states that 'Member States should … not provide for compensation for rightholders as regards uses under the text and data mining exceptions introduced by this Directive'. Nevertheless, the explanation provided in that recital concerns the 'scientific research' TDM exception under Article 3 of that directive only. In that regard, that recital states that 'in view of the nature and scope of [that] exception … any potential harm created to rightholders through this exception would be minimal', hence why Member States should not provide for such compensation. Nevertheless, the harm created to authors by the use of works and other subject matter for the purposes of LLM 'training', by way of the TDM exception under Article 4 of the same directive, may be *considerable*, thus justifying the need for the Member States to have such a possibility.

163 Beyond the scenario described in points 132 and 133 above, it could, for instance, 'unreasonably prejudice the legitimate interest' of the rightholders to exempt the use of their works and other subject matter to train a model to generate racist content (see, by analogy, judgment of 3 September 2014, Deckmyn and Vrijheidsfonds (C-201/13, EU:C:2014:2132, paragraphs 28 to 31) and Geiger, C., footnote 161, op. cit.).

164 See Ji et al, 'Survey of hallucination in natural language Generation', *ACM Computing Surveys*, 2023. An LLM may sometimes derive factually correct answers from its parameters. For instance, an LLM may accurately state that 'Washington' is the 'capital' of the 'United' 'States', because it observed, during its 'training', that there is a strong statistical relationship between those terms. Nevertheless, LLMs are trained on a final dataset, so that information derived from the latter becomes outdated very easily. New factual knowledge could be obtained by regularly re-training the model on new datasets. However, the training of an LLM requires significant time, money, computer resources and data, making it an unattractive option. Finally, an LLM cannot indicate to a user the *source* of the information it provides, which makes it impossible to obtain an authoritative answer from the LLM alone.

165 I underline, to avoid any confusion, that the information retrieved through RAG is not incorporated within the models' parameters. Stated differently, the LLM does not 'learn' new information through RAG. That information is merely fed as input data, together with the initial question of the user, for the model to generate a specific answer to the user's query, and discarded afterwards.

166 RAG mitigates rather than eliminates the risk of an LLM providing inaccurate answers. The LLM may still misinterpret content or miss context, 'hallucinate' when processing the reply, and so on. In fact, the output generated by Gemini in the example given by Like Company contains just such inaccuracies, namely the information regarding the fact that a special pool would host the dolphins in Lake Balaton and the dimensions of that pool.

167 The reason why hyperlinks are sometimes provided, and sometimes not, is not entirely clear.

168 See, in that regard, recital 119 of the AI Act. In fact, the Commission has recently designated Open AI's ChatGPT, which uses a similar form of dynamic RAG when providing answers to its users, as a 'Very Large Online Search Engine' (VLOSE) under Article 33 of Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (OJ 2022 L 277, p. 1) ('the Digital Services Act') (see European Commission 'Commission designates ChatGPT, Reddit, Roblox under Digital Services Act', 31 August 2026 (available at: https://ec.europa.eu/commission/presscorner/detail/en/mex_26_1775).

169 Reid, E., 'Generative AI in Search: Let Google do the searching for you', Google Blog, 14 May 2024 (available at: https://blog.google/products-and-platforms/products/search/generative-ai-google-search-may-2024/). See also Google, 'AI Overviews and AI Mode in Search', May 2025 (available at: https://search.google/pdf/google-about-AI-overviews-AI-Mode.pdf).

170 Amongst others, Perplexity AI is facing multiple actions brought before the US by major publishers on this issue (see Parker, O., 'Perplexity Lawsuit News Today: Full 2026 Update', *Lawfold.com*; available at: https://lawfold.com/perplexity-lawsuit-news-today/). A group of French press publishers brought a similar action against Brave (see Alliance de la presse d'information générale 'Intelligence artificielle générative : l'Alliance et 53 éditeurs assignent Brave devant le tribunal judiciaire de Paris', 1 June 2026; available at: https://www.alliancepresse.fr/actualite/intelligence-artificielle-generative-lalliance-et-53-editeurs-assignent-brave-devant-le-tribunal-judiciaire-de-paris/).

171 See, generally, Kowala, M., 'Protection of Press Publishers in the Age of Generative AI – In Search of Legal Remedies to Adapt to the Pace of Technology', *International Review of intellectual Property and Competition Law*, Vol. 55, Issue 10, 2024, pp. 1604 to 1623; Wang, Y., 'Legal solutions to journalism crisis? Evaluating press publishers' right, misappropriation doctrine and mediation mechanism', *European Intellectual Property Review*, Vol. 48, Issue 6, 2026, pp. 368 to 374.

172 See, inter alia, Prince, M., 'Content independence day: no ai crawl without compensation!', *The Cloudflare Blog*, 1 July 2025 (available at: https://blog.cloudflare.com/content-independence-day-no-ai-crawl-without-compensation/). Some studies indicate that, nowadays, while the number of users and searches on Google is steadily rising, the vast majority of those searches are 'zero-click' (meaning that the user ended his or her search without accessing any third-party website), thanks in particular to 'AI Overviews' (see Jain, S., 'Are Zero-Click Searches Still Rising? Analysing the Latest Trends in Google's SERPs', *Scrum Digital*, 16 May 2025; available at: https://scrumdigital.com/blog/zero-click-search-trends-google-serp-analysis/). This is denied by Google (see Reid, E., footnote 169, op. cit.: 'with AI Overviews, people are visiting a greater diversity of websites for help with more complex questions. And we see that the links included in AI Overviews get more clicks than if the page had appeared as a traditional web listing for that query …').

173 As explained below, the liability of the chatbot provider will require that it be shown that the infringement is *systemic* rather than *occasional* (see Section C.2.c below). Hence, a single example will rarely be enough.

174 This seems to me all the more reasonable that evidence of the use of a piece of content by an LLM-powered chatbot through RAG may typically be within the purview of the chatbot provider. The competent court could even order that such evidence be presented by that provider, under Article 6(1) of Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights (OJ 2004 L 157, p. 45).

175 That is the case as regards not only the answers themselves, but also the invisible, technical reproductions created during RAG (see, respectively, Sections C.2.a and C.2.b below). Those reproductions are an integral and ancillary part of the process leading to the generation of the outputs. Here, contrary to the reproductions carried out for the purposes of LLM training, there is a *direct* link between the two.

176 See, by analogy, judgments of 12 July 2011, L'Oréal and Others (C-324/09, EU:C:2011:474, paragraphs 58 to 64); of 18 October 2012, Football Dataco and Others (C-173/11, EU:C:2012:642, paragraphs 18 to 39; and of 9 July 2026, Anne Frank Fonds, C-788/24, EU:C:2026:559, paragraphs 21 to 34.

177 Specifically, both the answers themselves *and* the invisible, technical reproductions created during RAG may qualify as 'online use'. That is the case, again, because those invisible reproductions are an integral, ancillary part of the process leading to the transmission of information to that user by the chatbot via the Internet. Obviously, the applicability of Article 15(1) of the CDSM Directive also depends on whether that online use is carried out 'by [the] information society service provider' or by the individual user. Nevertheless, that issue relates to the substance of the questions referred, and will therefore be examined in Section C.2.c below.

178 Those interested parties submit that the first and fourth questions are predicated on the fact that the Gemini chatbot provided, as its answer, a text which is 'partially identical' to Like Company's press publication. In reality, a side-by-side comparison of the article in question and that answer shows that there is no such identity, apart from a few words in consecutive order.

179 See judgment of 6 October 2021, W.Ż. (Chamber of Extraordinary Control and Public Affairs of the Supreme Court – Appointment) (C-487/19, EU:C:2021:798, paragraphs 87 and 90).

180 In the only decision on Article 15(1) of the CDSM Directive so far, namely the judgment of 12 May 2026, Meta Platforms Ireland (Fair compensation) (C-797/23, EU:C:2026:395), the Court has clarified the exclusive *nature* of the press publishers' right only. It has not clarified the *scope* of that right and, in particular, the circumstances in which it is infringed.

181 To the extent, again, that the online use of such publications by an information society service provider is concerned.

182 I leave aside, in that context, the permanent reproductions of extracts of websites stored in the Google index. Indeed, those reproductions are not created for the purposes of grounding the chatbot's answer per se, but for the purposes of Google Search, a service that the chatbot just happens to use in order to identify relevant content.

183 See, by analogy, Wietzke, A., footnote 88, op. cit., paragraphs 15 and 16; and Thouvenin, F., footnote 13, op. cit., p. 290.

184 See, for the same view, Wietzke, A., footnote 88, op. cit., paragraphs 14, 29-38.

185 See, by analogy, judgment of 5 June 2014, Public Relations Consultants Association (C-360/13, EU:C:2014:1195, paragraph 40 and the case-law cited). See also EUIPO, op. cit., pp. 274-275; Wietzke, A., footnote 88, op. cit., paragraphs 53 and 54; and Thouvenin, F., footnote 13, op. cit., p. 86. I would point out that the content retrieved through RAG is not integrated into the model itself. Thus, there is no discussion of any 'permanent' reproduction being made in that context.

186 See, by analogy, judgment of 5 June 2014, Public Relations Consultants Association (C-360/13, EU:C:2014:1195, paragraphs 28, 43 and 49 and the case-law cited). I would stress that the issue is not whether a chatbot can provide an answer without RAG (which it can); it is whether RAG can function correctly and efficiently without the technical reproductions in question (which it cannot).

187 See order of 17 January 2012, Infopaq International (C-302/10, 'the order in Infopaq', EU:C:2012:16, paragraph 50 and the case-law cited).

188 See recital 33 of the InfoSoc Directive. See, by analogy, the order in Infopaq (paragraph 51). Wietzke, A., footnote 88, op. cit., paragraph 55. I am aware that, in paragraph 53 of the order in Infopaq (and only there) the Court stated, in passing, that a temporary reproduction would have 'independent economic significance' if 'the acts of temporary reproduction lead to a change in the subject matter reproduced, as it exists when the technological process concerned is initiated, because those acts no longer aim to facilitate its use, but the use of a different subject matter'. Nevertheless, I find that interpretation puzzling. First, a modified use of a protected subject matter in a modified form is nevertheless a use of the latter. Secondly, in the case that led to that order, the temporary reproductions at issue specifically permitted such a modified use of works, namely the drafting of summaries thereof. Again, InfoPaq derived an economic advantage from providing those summaries to its clients, not from the temporary reproductions made to draft them (something that the Court appeared to accept in paragraph 51 of that order). Finally, if the reasoning of the Court were true, it would mean that every time a technical reproduction of a work is made for the sole purpose of using that work in a modified, 'unrecognisable' manner, it would have the paradoxical result that (1) the end use would be free (see point 194 below) but (2) the technical reproduction would require the prior consent of the rightholders, since it would not be covered by the temporary reproduction exception. In effect, that would remove any scope for reusing protected material in a modified, 'unrecognisable' manner as soon as a digital process is involved. That would be contrary to the principle of technological neutrality on which EU artistic and literary property law rests.

189 See, in that regard, recital 33 of the InfoSoc Directive.

190 I will discuss the issue of implied authorisation in Section C.2.d below.

191 In addition, it goes without saying that the chatbot may only retrieve such a publication from a *lawful source* (typically, the website of the press publisher). See, to that effect, judgment of 26 April 2017, Stichting Brein (C-527/15, EU:C:2017:300, paragraphs 59 to 72).

192 See, to that effect, Wietzke, A., footnote 88, op. cit., § 56. Indeed, a use is not restricted by law, as envisaged in recital 33 of the InfoSoc Directive, where it is not reserved to the rightholder under an exclusive right (see, to that effect, the order in Infopaq (paragraphs 43 to 45)).

193 Again, this depends on whether the acts of use are carried out 'by [the] information society service provider'. If the act is attributable to the user only, it is not restricted under Article 15(1) of the CDSM Directive (see Section C.2.c below). Furthermore, there may be situations where, although the answer provided by the chatbot would, in principle, fall under Article 2 or Article 3 of the InfoSoc Directive, it is freely permitted under an exception or limitation to copyright and the related rights. Such an answer would also not be restricted under Article 15(1) of the CDSM Directive.

194 That is to say, independently from the technical reproductions created during the RAG process, discussed in the previous section.

195 While there is tension in the case-law of the Court on that point (compare, in particular, the judgments of 24 March 2022, Austro-Mechana, C-433/20, EU:C:2022:217, paragraphs 31 and 32, and of 16 April 2026, Stichting de Thuiskopie, C-496/24, EU:C:2026:296, paragraphs 26, 37 and 38), it seems to me that the reproduction right under Article 2 of the InfoSoc Directive and the right of making available to the public under Article 3(2) of that directive could potentially *apply cumulatively* with regard to the same chatbot answer. Both rights should therefore be examined in turn.

196 See, on the issue of infringing LLM-generated outputs, Novelli, C., Casolari, F., Hacker, P., Spedicato, G. and Floridi, L., 'Generative AI in EU Law: Liability, privacy, intellectual property, and cybersecurity', *Computer Law & Security Review*, Vol. 55, 2024, spec. p. 19; Guadamuz, A., footnote 78, op. cit., pp. 111 to 127; Rosati, E., 'Infringing AI: Liability for AI-Generated Outputs under International, EU, and UK Copyright Law', *European Journal of Risk Regulation*, Vol. 16, Issue 2, 2025, pp. 603 to 627; and D'Alfonso, J., 'Generative artificial intelligence outputs, copyright infringement, and the assignment of liability', *AI and Ethics*, Issue 5, 2025, pp. 5295 to 5308.

197 See judgments of 29 July 2019, Pelham and Others (C-476/17, EU:C:2019:624, paragraphs 29 to 31); of 9 March 2021, VG Bild-Kunst (C-392/19, EU:C:2021:181, paragraph 25); and of 4 December 2025, Mio and Others (C-580/23 and C-795/23, EU:C:2025:941, paragraphs 84 to 86).

198 In other words, in that scenario, the generation of that answer is clearly not an act of independent creation.

199 As mentioned in point 163 above, the fact that a press publication was retrieved and reused to ground a chatbot's answers could be presumed where the chatbot potentially has access to the publication (typically, where it was available online) and the answer contains a hyperlink leading back to the webpage hosting that publication, or is sufficiently similar to the latter. This, in fact, makes the second condition (recognisability) decisive in an infringement claim.

200 And so, by an average press reader (see Kraetzig, V., 'Press Publishers Versus AI – Not a Matter of Property but of Fairness', *International Review of Intellectual Property and Competition Law*, 2025, Vol. 56, Issue 9, p. 1759). See, by analogy, Landgericht München (Regional Court, Munich, Germany), 31 July 2026, *GEMA* v *Suno*, 42 O 763/25, paragraphs 319-367.

201 See, to that effect, recitals 54 and 58 of the CDSM Directive, and Franceschini, L. and Estivalèzes, M., *Rapport sur l'objet et le champ d'application du droit voisin des éditeurs de publication de presse*, CSPLA, Janvier 2018, pp. 11 to 13. See also, mutatis-mutandis, judgment of 3 June 2021, CV-Online Latvia, C-762/19, EU:C:2021:434, paragraphs 24 to 27 and the case-law cited.

202 And, indeed, press publishers have been borrowing facts from each other's articles since printing press exists (see Kraetzig, V., footnote 200, op. cit.). By analogy, in copyright law, under the fundamental idea/expression dichotomy, information as such cannot be monopolised. In that respect, Article 2(8) of the Berne Convention confirms that 'news of the day or … miscellaneous facts having the character of mere items of press information' are ineligible, in themselves, for copyright protection (see my Opinion in Funke Medien NRW (C-469/17, EU:C:2018:870, points 15 and 16)).

203 See my Opinion in Funke Medien NRW (C-469/17, EU:C:2018:870, point 16).

204 See Franceschini, L. and Estivalèzes, M., footnote 201, op. cit., pp. 22 to 25; Matanovac Vučković, R., 'The Effectiveness of the Press Publishers' Related Right', *International Review of Intellectual Property and Competition Law*, Vol. 52, 2021, pp. 1050 to 1061, spec. pp. 1055 and 1056. See, by analogy, judgment of 3 June 2021, CV-Online Latvia (C-762/19, EU:C:2021:434, paragraphs 40 and 47); see also my Opinion in Gândul Media Network (C-598/24, EU:C:2025:1002, point 43).

205 In December 2023, the New York Times filed infringement claims against OpenAI and Microsoft. Inter alia, the New York Times submitted examples where ChatGPT had produced verbatim reproductions of parts of articles published in the News York Times.

206 See Kraetzig, V., footnote 200, op. cit., p. 1759.

207 Such a filter automatically compares the source and the output generated by the LLM to screen out matching text excerpts exceeding a certain length.

208 The 'summary' provided by the Gemini chatbot is, in reality, rather approximative. First, it contains only some of the facts reported in Like Company's article (namely those related to the Hungarian singer's project to introduce dolphins in Lake Balaton, whereas the articles also reported other matters). Secondly, that 'summary' also contained a series of statements finding no basis in that publication (the fact that a pool is going to be used to introduce the dolphins, the size of that pool, the reception by the public of that project, and so on) which, according to Google, are 'hallucinations' of the LLM).

209 See Kraetzig, V., footnote 200, op. cit., p. 1759.

210 I note that the article at issue in the example given by Like Company before the referring court appears to belong to the second category, as that article appears to consist in a purely informative report of events (namely the Hungarian singer's plan to introduce dolphins in Lake Balaton), using commonplace arrangement, structure and sentences, much of which appears, in addition, to be borrowed from other publications.

211 The 'collection' of which a press publication is comprised must, under Article 2(4) of the CDSM Directive, I recall, be 'composed mainly of literary works of a journalistic nature'. However, purely informative reports of events, the content of which is essentially determined by the information which such reports contain, are devoid of 'originality' and, thus, do not qualify as original works (see judgment of 29 July 2019, Funke Medien NRW (C-469/17, EU:C:2019:623, paragraph 24 and references). Nevertheless, Article 2(4) indicates that such a collection 'can also include … other subject matter'.

212 See judgment of 29 July 2019, Funke Medien NRW (C-469/17, EU:C:2019:623, paragraph 24 and the case-law cited).

213 See Kraetzig, V., footnote 200, op. cit., p. 1760. That difference in treatment is not incompatible, but rather consistent with, the investment-protection objective pursued under the press publishers' right. Those different kinds of news articles do not reflect the same degree of investment. Purely informative reports can even easily be generated by press publishers using AI. More complex pieces require a deeper investment in terms of the aggregation, verification, analysis and, ultimately, the presentation of the news.

214 See Vesala, J., 'Press Publishers' Right and Artificial Intelligence', in Pihlajarinne, T. and Alen-Savikko, A. (eds.), *Artificial Intelligence and the Media – Reconsidering Rights and Responsibilities*, Edward Elgar, 2021, spec. pp. 262 and 263.

215 See, by analogy, the order in Infopaq (paragraphs 43 to 45).

216 See, to that effect, judgments of 13 February 2014, Svensson and Others (C-466/12, EU:C:2014:76, paragraphs 18 and 20); of 7 August 2018, Renckhoff (C-161/17, EU:C:2018:634, paragraphs 17, 20 and 21); and of 17 June 2021, M.I.C.M. (C-597/19, EU:C:2021:492, paragraph 41)

217 See, inter alia, judgment of 19 December 2019, Nederlands Uitgeversverbond and Groep Algemene Uitgevers (C-263/18, EU:C:2019:1111, paragraphs 66 to 68 and the case-law cited).

218 See, inter alia, judgment of 9 March 2021, VG Bild-Kunst (C-392/19, EU:C:2021:181, paragraph 32 and the case-law cited).

219 I recall that, in order to determine the public, account should be taken of the number of users able to access the content at the same time and in succession (see, inter alia, judgment of 19 December 2019, Nederlands Uitgeversverbond and Groep Algemene Uitgevers, C-263/18, EU:C:2019:1111, paragraphs 66 to 68).

220 Here, the recognisability test used with respect to the reproduction right under Article 2 of the InfoSoc Directive may be followed by analogy (see, to that effect, judgment of 9 March 2021, VG Bild-Kunst (C-392/19, EU:C:2021:181, paragraphs 22, 23 and 25). Thus, I refer the reader to the preceding section in that regard.

221 I have in mind, for instance, a scenario where users asking the chatbot to obtain information about a given news item would be systematically provided with answers which recognisably reused an article on the topic. I also underline that simple queries, corresponding to the normal use of the chatbot, should be taken into account in that context, not convoluted and 'abnormal' prompts designed to trick the chatbot into doing what it is not designed to do.

222 See Hacker, P., 'Copyright, AI, and the Future of Internet Search before the CJEU – Reflections on Like Company v Google', Verfassungsblog, 17 July 2025 (available at: https://verfassungsblog.de/copyright-ai-cjeu/). In the alternative, should the Court disagree with my interpretation of the concept of 'specific technical means', it seems to me that the question whether the chatbot makes available a press publication to a 'new public' finds a straightforward answer in the case-law of the Court. In particular, in its judgments of 13 February 2014, Svensson and Others (C-466/12, EU:C:2014:76, paragraphs 25 to 28, 31), and of 9 March 2021, VG Bild-Kunst (C-392/19, EU:C:2021:181, paragraphs 36 and 38 and 41 to 47), the Court took the view that, where a rightholder makes available a protected subject matter on a website, without putting in place effective technical measures (within the meaning of Article 6(3) of the InfoSoc Directive) restricting access to that website, he or she is deemed to have taken into account, as the 'public', *all internet users*. By contrast, where the rightholder has put in place such measures, he or she is deemed to have taken into account, as the 'public', *the users of that website only*, and any communication of the subject matter outside that website is deemed to be made to a 'new public'. Accordingly, that chatbot would make that publication available to a 'new public' only if it has circumvented effective technical measures restricting access to the publication in question on the original website (be it a paywall, or measures designed to block robots).

223 I would recall that, in the example provided by Like Company before the referring court, the query entered by the unknown user expressly asked the Gemini chatbot to 'provide a summary in the Hungarian language [of] the press publication published on the balatonkornyeke.hu website concerning Kozsó's project to introduce dolphins in a lake'.

224 See, inter alia, Krokida, Z., 'Large Language Models and EU Intermediary Copyright Liability: Quo vadis?', *European Intellectual Property Review*, Vol. 46, Issue 6, 2024, pp. 361 to 373; Frosio, G., 'Schrödinger's Chatbot: Platform Liability, Volition, and Generative AI', SSRN, 2026.

225 See, by analogy, judgment of 22 June 2021, YouTube and Cyando (C-682/18 and C-683/18, EU:C:2021:503, paragraphs 71 to 75).

226 See, by analogy, judgments of 22 June 2021, YouTube and Cyando (C-682/18 and C-683/18, EU:C:2021:503, paragraphs 68, 78 and 81), and of 13 July 2023, Ocilion IPTV Technologies (C-426/21, EU:C:2023:564, paragraph 37).

227 See, by analogy, judgment of 14 June 2017, Stichting Brein (C-610/15, EU:C:2017:456, paragraphs 44 to 46), and my Opinion in Meta Platforms Ireland (Fair compensation) (C-797/23, EU:C:2025:552, points 27 to 29). An example would be a chatbot service specifically designed to provide extensive and detailed presentations of individual news articles for its users, merely paraphrased. A service of the sort would not be different, functionally speaking, from the news aggregators or media monitoring services that justified the creation of the press publishers' right in the first place (Vesala, J., footnote 214, op. cit., p. 263).

228 See, by analogy, judgment of 22 June 2021, YouTube and Cyando (C-682/18 and C-683/18, EU:C:2021:503, paragraph 84).

229 Such an obligation to put in place measures to mitigate the risk of copyright infringement may also stem, in certain circumstances, from the Digital Service Act. Specifically, where chatbots qualify as Very Large Online Search Engines, the providers thereof have the obligations to 'assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services' (Article 34) and to 'put in place reasonable, proportionate and effective mitigation measures' in that regard (Article 35). The dissemination of illegal content, including content protected under copyright or the related right, is considered just such a systemic risk (see recitals 12 and 80 of that act). Furthermore, in my view, one has to bear in mind that LLM developers should themselves put in place measures to mitigate the risk of copyright infringement through the model (see points 132 and 133 above). That raises the issue of the respective attribution of responsibility and liability between LLM developer and chatbot provider. Nevertheless, here, that is Google in both counts.

230 See, by analogy, judgment of 7 August 2018, Renckhoff (C-161/17, EU:C:2018:634, paragraph 29 and the case-law cited).

231 See judgment of 16 November 2016, Soulier and Doke (C-301/15, EU:C:2016:878, paragraph 35).

232 Judgment of 9 March 2021 (C-392/19, EU:C:2021:181).

233 See judgment of 9 March 2021, VG Bild-Kunst (C-392/19, EU:C:2021:181, paragraphs 37 to 48)

234 Ibid., paragraphs 38, 39 and 40.

235 See EUIPO, footnote 9, op. cit., p. 139 and references.

236 See, on that issue, EUIPO, footnote 9, op. cit., pp. 159, 178 and 179 and references. Google claims that the 'Google Extended' protocol allows such differentiation. It explained at the hearing that using that protocol to prevent access to a website for the purposes of model training and grounding would not affect the website's visibility on Google. That would need to be verified. Nevertheless, even if that were the case, in the current, still-early stages of AI and chatbots, such findings of implicit consent by website publishers seem far-fetched, in my view.

237 To be clear, it follows from points 187 and 188 above that the presence, or absence, of effective technical measures restricting access to a given press publication *is* an important element of the analysis to determine whether the use of that publication to ground the answers provided by chatbots infringes Article 15(1) of the CDSM Directive. Where chatbots circumvent such measures, that affects the legality of the use. Nevertheless, *even where such measures have not been put in place by the publisher*, the use may still be infringing where the answers provided by the chatbots recognisably reuse the form of expression of such a publication.

238 I leave aside the issue of whether, and if so in which circumstances, the exception for reporting of current events under Article 5(3)(c) of the InfoSoc Directive or the quotation exception under Article 5(3)(d) of that directive could apply with respect to such an answer.

239 It may even create a healthy competitive pressure, for press publishers, to provide articles which do not merely consist in purely informational reports of events (as those can also be produced by chatbots), but contain a real added value for the reader (see Kraetzig, V., footnote 200, op. cit.).

240 I would stress that, if chatbots were entitled to retrieve protected works and other subject matter from unlawful sources or in circumvention of such effective technological measures to ground their answers, that would unreasonably prejudice the interests of rightholders. Indeed, that would entirely displace the possibility for those rightholders to licence such use. In addition, where an LLM-powered online chatbot circumvents such technical measures, this may not only make the use of the press publication infringing, but it also constitutes a tort, under Article 6(1) of the InfoSoc Directive.

241 Indeed, it would appear that licensing for RAG purposes is developing (see, in that respect, EUIPO, footnote 9, op. cit., p. 106 and references.

242 To ensure that the press publishers' rights are respected, obligations of transparency as to the content scraped by chatbots to ground their answers should be imposed on chatbot providers, similar to the obligations of transparency imposed on LLM developers under the AI Act.

243 The designation of online chatbots as 'search engines' under the Digital Service Act may provide some solutions in that regard, since that instrument contains obligations tailored to those risks (see footnote 229 above).