From Reproduction to Licensing: Applying Article 15 CDSMD to the Process of Generative AI Training

Klara Schinzler

ABSTRACT

As Generative AI becomes central to the digital landscape, its reliance on vast datasets – often sourced from publicly available press publications – raises pressing legal questions concerning intellectual property (IP) rights. This article examines whether the use of such content for training AI systems may infringe Article 15 of the Copyright in the Digital Single Market Directive (CDSMD), a provision originally intended to regulate unlicensed uses by news aggregators and search engines.[1] It explores the legal implications of using press content at scale for training purposes – typically without attribution, remuneration, or a clear legal basis – and identifies the specific stages of the training process where reproduction rights may be implicated. A central issue is the scope of the press publishers' right (PPR), particularly the distinction between protected editorial content and unprotected "mere facts". To support this analysis, the article develops a test for assessing whether a given use constitutes infringing reproduction under Article 15 CDSMD. It further argues that, where training uses qualify as infringing, collective licensing could offer a pragmatic solution – ensuring legal certainty for developers, fair compensation for publishers, and fostering a sustainable and pluralistic digital information ecosystem.

1. INTRODUCTION

Since the public emergence of Generative AI in late 2022, the technology has been described as an "earthquake in the creative sectors and in the field of copyright, of a magnitude not experienced since the emergence of the Internet".[2] These models rely on vast datasets – much of it scraped from publicly available sources without authorisation. Press content plays a central role in these datasets. Key LLM training datasets are disproportionately composed of high-quality content owned by commercial publishers of news and media websites.[3] This places press publishers in a paradoxical position: their content is indispensable for AI development, yet their rights are frequently ignored. As *Francesco Marconi* notes, media companies hold "some of the most valuable assets for AI development: text data for training models and ethical principles for creating reliable and trustworthy systems."[4] Unlike news aggregators, Generative AI does not link to or summarise content – it processes and internalises it in new forms, often bypassing attribution and user engagement entirely. As a recent *TollBit* report indicates, referral rates from AI chatbots to publishers' sites are 95.7% lower than from traditional search engines, with only 0.37% of users clicking through.[5]

In this context, Article 15 CDSMD emerges as a potentially significant legal tool. It was designed to grant press publishers control over certain uses of their content online. Yet its applicability to Generative AI training remains uncertain. The exclusion of "mere facts" and the absence of a clear threshold for protection create interpretative difficulties, especially in a sector where factual

1 Recitals 54 & 55 CDSMD; E., Treppoz., "The Past and Present of Press Publishers' Rights in the EU", (2023), 46 (3) Colum. J.L. & Arts, 276 <https://doi.org/10.52214/jla.v46i3.11228> last accessed 13.05.2025.

2 P. B., Hugenholtz, "Copyright and the Expression Engine: Idea and Expression in AI-Assisted Creations", (2024), Chicago-Kent Law Review, 3 <https://www.ivir.nl/publicaties/download/chicagokentlawreview2024.pdf> last accessed 13.05.2025.

3 G., Wukoson & J., Fortuna, "The Predominant Use of High-Authority Commercial Web Publisher Content to Train Leading LLM's", (2024), 1 referring to publishers in the United States, <https://www.ziffdavis.com/wp-content/uploads/2024/11/The-Predominant-Use-of-High-Authority-Commercial-Web-Publisher-Content-to-Train-Leading-LLMs.pdf> last accessed 13.05.2025.

4 M., Adami, "Is ChatGPT a threat or an opportunity for journalism? Five AI experts weigh in", (Reuters Institute 2023) <https://reutersinstitute.politics.ox.ac.uk/news/chatgpt-threat-or-opportunity-journalism-five-ai-experts-weigh> last accessed 13.05.2025.

5 EPC, "AI chatbots are killing publishers traffic – everyone loses out", (2025) <https://www.epceurope.eu/post/ai-chat-bots-are-killing-publishers-traffic-everyone-loses-out> last accessed 13.05.2025 citing TOLLBIT, "AI Scraping Is On The Rise. TollBit State of the Bots – Q42024", (2025) <https://tollbit.com/bots/24q4/> accessed last on 13.05.2025.

reporting is central. Meanwhile, growing concerns about IP infringement in AI development – now rated a top risk by McKinsey's 2024 Global AI Survey[6] – underscore the urgency of resolving these issues. In response, press publishers are exploring parallel strategies: suing and signing.[7] So, some are initiating lawsuits, while others advocate for agreements.

This article examines whether and how the PPR applies to AI training on press content and explores how lawful reuse could be enabled through licensing mechanisms. This analysis unfolds across three principal sections: defining the substantive scope of the right (including the exclusion of "mere facts"), applying this framework to the technical architecture of Generative AI training processes and evaluating licensing mechanisms – particularly collective licensing – as a mechanism to reconcile legal protection with innovation.

2. THE SUBSTANTIVE SCOPE OF ARTICLE 15 CDSMD AND WHEN INFRINGEMENT OCCURS

To determine whether Article 15 can be applied to the context of Generative AI training, it is essential to first clarify the general scope of the right and establish when acts of infringement arise.

2.1 Scope of Protection

The PPR grants press publishers protection for the online use of their publications by ISSP's (Information Society Service Providers). It creates a standalone related right – similar to those granted to other investors like broadcasters or phonogram and film producers.[8]

"Press publication" is defined in Article 2 (4) CDSMD as a collection primarily composed of literary works of a journalistic nature, which may also include other works or subject matter, and which cumulatively fulfil three conditions: (a) Constituting 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) having 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. Periodicals with scientific or academic aims are expressly excluded. Recital 56 of the CDSMD further clarifies that the concept covers media such as newspapers, subscription-based magazines, and news websites, but not blogs or non-editorial platforms. While other content types such as videos or photos are not excluded per se, the publication must still be primarily journalistic in nature. As with toher provisions in EU Directives that do not refer to Member States' laws, the concept of "press publication" is an autonomous notion of EU law requiring uniform application across the Union,[9] while its application to specific facts must be conducted on a case-by-case basis within the fixed legal framework, so taking into account all the cumulative requirements.[10]

Article 2 (5) CDSMD defines ISSPs in line with Article 1 (1) (b) Directive 2015/1535:[11] services must be remunerated, provided at a distance, by electronic means, and on individual request.[12] CJEU case law and Recital 18 of the Ecommerce Directive 2000/31 confirm that this definition covers a broad range of online economic activities.[13] ISSPs do not need to be established within the EU, but targeting EU users appears to be necessary, according to *Rosati* mere accessibility seems to be insufficient.[14]

The rights conferred in Article 15(1) mirror those in Articles 2 and 3(2) of the InfoSoc-Directive,[15] namely reproduction and communication to the public, including making available to the public. Article 2 defines reproduction broadly, including direct and indirect copying, temporary or permanent, whole or partial, leading to a high level of protection.[16] Though Article 15 does not clar-

6 R., Levy, "Navigating Copyright in the Age of Generative AI: Responsible AI Starts with Licensing", (2024) <https://www.copyright.com/blog/navigating-copyright-Generative-ai-responsible-ai-starts-with-licensing/> last accessed 13.05.2025 citing A., Singla and Others, "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value", (2024), Exhibit 7 <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024> last accessed 13.05.2025; for a newer version of the study s. A., Singla and Others, "The state of AI: How organisations are rewiring to capture value", (2025) <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/> last accessed 13.05.2025.

7 C., Tobitt, "Who's suing AI and who's signing: Ziff Davis sues OpenAI after Washington Post signs deal. 14 major publishers sue AI start-up Cohere Inc.", (2025) <https://pressgazette.co.uk/platforms/news-publisher-ai-deals-lawsuits-openai-google/> last accessed 13.05.2025.

8 Which was the original idea of the proposal; s. L., Bently and Others, "Strengthening the Position of Press Publishers and Authors and Performers in the Copyright Directive", (European Parliament, Policy Department for Citizens' Rights and Constitutional Affairs 2017), Study for the JURI Committee, 15 <https://www.europarl.europa.eu/RegData/etudes/STUD/2017/596810/IPOL_STU(2017)596810_EN.pdf> last accessed 13.05.2025.

9 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 260, for general cases in the field of copyright and related rights s. inter alia – Case C-5/08 *Infopaq International* ECLI:EU:C:2009:465 para 27–29 and C-128/11 UsedSoft ECLI:EU:C:2012:407 para 40 cited Ibid.

10 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 262.

11 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, OJ L 241, 17.9.2015, pp. 1–15.

12 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 262.

13 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 262 citing Case C-649/18 *A (Advertising and sale of medicinal products online)* EU:C:2020:764 para 31.

14 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 263. Deriving this statement from the fact that the CJEU, while no decision has been made yet in relation to the right of communication to the public, it has established this approach in relation to the right of distribution, the SGDR and in the trade mark field.

15 Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L 167/10, herein InfoSoc-Directive.

16 Case C-5/08 *Infopaq International* ECLI:EU:C:2009:465 para 42–42 & Case C-476/17 *Pelham* ECLI:EU:C:2019:624 para 30 cited in E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 264.

ify whether the PPR follows Article 2(a) (authors) or 2(b) – (e) (related rights holders) InfoSoc-Directive, Recitals 54–55 clarify that the protection is based on investment, aligning it with the latter.[17] However, the right does not apply to private or non-commercial use.[18] Article 15 (1) CDSMD also excludes hyperlinking, individual words, and "very short extracts", though they remain undefined,[19] leading to fragmentation in national implementation.[20]

A further exclusion – of "mere facts" – appears in Recital 57 CDSMD. It is prima facie based on a foundational principle of copyright, the ideas/expression dichotomy – which holds that protection is only granted to the expression of ideas rather than ideas themselves.[21] "Copyright protection may be granted to expressions, but not to ideas, procedures, methods of operation or mathematical concepts as such."[22] In the light of this premise, the notion of facts shall be intended to encompass ideas, procedures, methods of operation, or mathematical concepts as such[23] and "mere" refers to "nothing more than".[24] However, the PPR differs from copyright in that it does not require originality; it protects not the intellectual creation, but the organisational and financial investment made by the press publisher in producing press publications.[25]

Despite the centrality of this exclusion, it is noteworthy that it does not appear in the operative provision itself. This raises the question about its legal function. Recitals cannot create new rights or restrictions; however, they may clarify the meaning of provisions where consistent with the legislative text.[26] In this context, the "mere facts" exclusion is best understood not as an autonomous norm-setting device, but as a clarification that does not extend beyond the scope already implied by Article 15 (1) CDSMD. Facts, by nature, are discovered rather than created; they are the raw materials of journalism, not its protected product. The investment protected by Article 15 must go beyond the mere collection of factual content and reflect organisational or editorial effort.[27] This understanding finds further support in the Sui Generis Database Right (SGDR) under Directive 96/9/EC,[28] which protects substantial investment in obtaining or verifying data, but not in its creation.[29] Analogously, under Article 15 CDSMD, the mere effort of uncovering or recording facts does not suffice to trigger protection, unless those facts are presented in a way that reflects editorial or organisational input.

Unlike the SGDR,[30] Article 15 CDSMD does not require substantiality of investment.[31] Consequently, any demon-

17 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 267.

18 Those remain subject of already existing copyright rules, s. Recital 55 CDSMD.

19 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 274.

20 S. i.e. E., Rosati, "Is Harmonization Good if the End Result is Even More Fragmentation? The Case of Article 15 CDSM Directive and the Exclusion of 'Very Short Extracts'", (2023), forthcoming in M., Senftleben and Others (eds), The Cambridge Handbook on Media Law and Policy in Europe (CUP), Stockholm Faculty of Law Research Paper Series, no. 129 <https://ssrn.com/abstract=4519834> last accessed 13.05.2025.

21 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 286; Case C-310/07 *Levola Hengelo* EU:C:2018:899 para 39 referring to Case C-406/10 *SAS Institute* EU:C:2012:259 para 33.

22 Case C-310/07 *Levola Hengelo* EU:C:2018:899 para 39 referring to Case C-406/10 *SAS Institute* EU:C:2012:259 para 33.

23 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 286.

24 "mere" <https://dictionary.cambridge.org/dictionary/english/mere> last accessed 13.05.2025.

25 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 286.

26 T., Klimas & J. Vaičiukaitė, "The Law Of Recitals in European Community Legislation", (2009), 15(1) ILSA 63 <https://nsuworks.nova.edu/ilsajournal/vol15/iss1/6> last accessed 13.05.2025; Case C-173/99 *BECTU* ECLI:EU:C:2001:356 para 37–39 cited in M., Den Heijer, T. v. O. v. den Abeelen, & A., Maslyka, "On the Use and Misuse of Recitals in European Union Law", (2019), Amsterdam Law School Research Paper No. 2019-31, Amsterdam Center for International Law No. 2019-15, 5 <https://dx.doi.org/10.2139/ssrn.3445372> last accessed 13.05.2025.

27 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 286., Axel Springer SE, Written Comments in Response to the US Office's Pubilshers' Protection Study, (2021), 25 <https://www.copyright.gov/policy/publishersprotections/initial-comments/Axel%20Springer%20SE%20-%20Initial%20Comment.pdf> last accessed 13.05.2025.

28 Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the protection of databases [1996] OJ L77/20.

29 For the so-called creation/obtaining dichotomy s. Case C-762/19 *CV-Online Latvia* ECLI:EU:C:2021:434; Case C-338/02 *Fixtures-Svenska* ECLI:EU:C:2004:696; Case C-203/02 *British Horseracing* ECLI:EU:C:2004:695; Case C-46/02 *Fixtures-Oy* ECLI:EU:C:2004:694; Case C-444/02 *Fixtures-OPAP* ECLI:EU:C:2004:697 all as cited in P., Burdese, "AI-generated databases. Do the creation/obtaining Dichotomy and the Substantial Investment Requirement Exclude the Sui Generis Right Provided for under the EU Database Directive? Reflection and proposals.", (2020), WIPO academy, University of Turin and ITC-ILO, Master of Laws in IP, Research Papers Collection – 2019–2020, 5 <https://dx.doi.org/10.2139/ssrn.3850662> last accessed 13.05.2025.

30 Recitals 7, 39 and 40 of the Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the protection of databases [1996] OJ L77/20.

31 In a *a contrario* reading of the Directives, comparable to Case C-476/17 *Pelham* ECLI:EU:C:2019:624, Opinion AG Szpunar ECLI:EU:C:2018:1002 para 37, 38.

strable investment, however minimal, may attract protection – unless the content qualifies as "mere facts." As such, the exclusion of mere factual content becomes the primary threshold delimiting the scope of the PPR. Mirroring the copyrights idea/expression dichotomy, which requires originality from creative freedom to trigger protection,[32] analogies can be drawn from copyright case-law. According to the CJEU, content entirely determined by facts – where expression and information are indissociable – lacks originality.[33] *AG Szpunar* in *Funke Medien NRW* stressed that copyright must not be used to restrict access to information vital for democratic discourse and mechanisms like the idea/expression dichotomy must be given full effect in light of freedom of expression.[34] Analogously, under the PPR, when press content is wholly shaped by facts – i.e. simple headlines or statistical reports – protection does not arise unless distinct editorial investment is evident.

Concluding, this limitation ensures that the PPR does not devolve into a mechanism for monopolising public domain content but remains focused on its stated objective: securing a sustainable press sector by protecting investment in the editorial process. In the specific context of the news sector, the exclusion of "mere facts" is particularly significant. Information works are often constrained by limited expressive means, raising concerns under the idea/expression dichotomy.[35] Applying this argument to related rights requires caution, as the PPR protects press publications regardless of originality. While facts may be expressed in limited ways—and thus investment in presenting them is also limited—this does not unduly restrict the PPR's scope, especially since the exclusion of facts is the only explicit threshold under Article 15 CDSMD and serves to balance IP protection with fundamental rights under Article 17(2) ECFR.[36]

2.2 Determining Infringement: Towards a functional test

Having clarified the scope of Article 15 CDSMD through the "mere facts" exclusion, the next step is to assess when a specific use of protected content constitutes infringing reproduction (in part). This inquiry is central to determining whether acts such as Generative AI training may infringe the Press Publishers' Right (PPR). While content lacking financial or organisational investment falls outside the right's scope, use of protected content still requires assessment as to whether it triggers the reproduction right – especially in cases involving partial reuse.

Drawing from *Pelham*,[37] infringement occurs where reproduction interferes with the rightholder's ability to recoup investment. Although *Pelham* concerned phonogram producers, the CJEU's reasoning is applicable to Article 2(b)–(e) InfoSoc rights more broadly.[38] Given that Article 15 CDSMD shares this investment-based rationale, applying this interpretation and the underlying balancing approach is both appropriate and coherent. Article 17(2) of the Charter of Fundamental Rights of the European Union (ECFR) does not confer absolute IP protection; it must be balanced against competing rights, including freedom of expression under Article 11 ECFR. Therefore, the relevant question becomes whether the reproduction interferes with the economic return on investment, not merely whether a portion of content is taken.[39] This is the case when what has been reproduced, indirectly or directly, in whole or in part, reflects the investment made by the concerned publisher.[40]

To make this determination, the concept of "investment" must be understood. The meaning and scope of reproduction (in part) must be determined by considering their usual meaning in everyday language, while also taking into account the context in which they occur and the purpose of the rules of which they are part.[41] As this is tied to the concept of investment, the same goes for that determination. According to the *Cambridge English Dictionary* an investment is the act of putting money or effort into something to make a profit or achieve a result.[42] Financially, it refers to using capital in the present to increase an assets value over time.[43] Legally, the nature of protected investment is inherently dependent on the subject matter of the related right in question.[44]

Investment, as relevant to press publishers and taken from the definition of "press publication", stems from editorial initiative, responsibility, and control. These functions encompass content initiation, editing, and publication oversight.[45] Any demonstrable investment is

32 S. i.e. Case C-469/17 *Funke Medien NRW* ECLI:EU:C:2019:623 para 19; Case C-5/08 *Infopaq International* ECLI:EU:C:2009:465 para 49; Case C-145/10 *Painer* ECLI:EU:C:2011:798 para 89, 92.

33 Case C-469/17 *Funke Medien NRW* ECLI:EU:C:2019:623, Opinion AG Szpunar ECLI:EU:C:2018:870 para 19.

34 Case C-469/17 *Funke Medien NRW* ECLI:EU:C:2019:623, Opinion AG Szpunar ECLI:EU:C:2018:870 para 37 cited in C., Geiger & E. Izyumenko, "Freedom of Expression as an External Limitation to Copyright Law in the EU: The Advocate General of the CJEU Shows the Way", (2019), 41(3) E.I.P.R., 133.

35 U., Furgal, "Rights on News: expanding copyright on the internet", (2020), Florence: European University Institute, EUI, LAW, PhD Thesis, 150–152 <https://doi.org/10.2870/82845> last accessed 13.05.2025.

36 Intellectual Property rights are not protected as absolute rights, s. i.e. Case C-469/17 *Funke Medien NRW* ECLI:EU:C:2019:623 para 72.

37 Case C-476/17 *Pelham* ECLI:EU:C:2019:624.

38 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 266.

39 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 266 citing Case C-476/17 *Pelham* ECLI:EU:C:2019:624 para 33, 34.

40 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 266.

41 Case C-476/17 *Pelham* ECLI:EU:C:2019:624 para 28., Case C-201/13 *Deckmyn and Vrijheidsfonds* ECLI:EU:C:2014:2132 para 19 and the case-law cited.

42 "investment" <https://dictionary.cambridge.org/dictionary/english/investment> last accessed 13.05.2025.

43 A., Hayes, "Investment: How and Where to Invest" <https://www.investopedia.com/terms/i/investment.asp> last accessed 13.05.2025.

44 WIPO, Understanding Copyright and Related Rights, (2016), 27 <https://doi.org/10.34667/tind.28946> last accessed 13.05.2025.

45 M. C., Caron, "Legal Analysis with focus on Article 11 of the proposed Directive on copyright in the Digital Market", (European Parliament, Policy Department for Citizens' Rights and Constitutional

sufficient to trigger protection. However, not every minor or insubstantial use will interfere with the opportunity to recoup such investment.

The *Pelham* decision recognised that phonograms are protected as indivisible wholes due to the fixation requirement.[46] By contrast, press publications are not defined by fixation, and may consist of both protected and unprotected elements. Thus, a recognisability test alone is inadequate for the PPR.

Examining the explicit exclusions in Article 15 CDSMD could help clarify where investment is typically absent and, by contrast, where it may be inferred. However, these exclusions do not imply an absence of investment per se; rather, each use must be assessed individually. If the reused material reflects investment, it may still fall within the right's scope, subject to applicable exceptions.[47] Thus, the exclusions inform – but do not fix – the boundaries of protection, underscoring the need for a flexible, context-sensitive standard.

To this end, a three-step functional test is proposed:

I. **Recognisability of editorial elements**: Recognisability, though not a standalone test under the PPR, serves as a meaningful entry point for assessing infringement due to the right's inherently vague and non-fixed subject matter. Unlike the phonogram producers' right, where the object of protection is concretely fixed,[48] the PPR protects investment without a fixation requirement. It can therefore be subtle and difficult to isolate. This makes the presence of recognisable elements – such as distinct editorial structure, wording, or formatting – especially significant. If reused material is identifiable despite the lack of fixation and the diffuse nature of the subject matter, this strongly suggests that protected investment has been appropriated.

II. **Value contribution**: The part used must contribute to the economic value of the original publication. The idea for added value as a tool for assessing infringement stems from the concept of financial investment, which implies an expectation of return and value enhancement.[49] Since added value is more tangible and measurable – i.e. through user engagement or licensing demand – it serves as a practical proxy for determining whether a use interferes with the publisher's ability to recoup that investment. This is easier than assessing the precise location of editorial investment, which is often diffuse and intangible.

III. **Substitution potential**: The idea of substitution potential as the last indicator arises from the original rationale behind the right – namely, to counteract losses caused by news aggregators diverting users away from original sources.[50] While actual substitution is rare, the potential to fulfil the same user need as the original can interfere with investment recoupment. This step introduces a subjective but necessary inquiry into market dynamics and content function.

These steps should be cumulatively applied to establish infringement. However, each may also serve as an indicator on its own. Most importantly, the exclusion of "mere facts" remains a mandatory limiting principle and must be considered throughout.

In conclusion, Article 15 CDSMD creates a low-threshold, investment-based related right aimed at press sector sustainability. The proposed test offers legal clarity in assessing infringement without undermining fundamental rights. Ultimately, judicial interpretation – particularly by the CJEU – will be necessary to define its boundaries and ensure a fair balance between rightholders and users in the digital environment.

3. IS AI TRAINING INFRINGING ARTICLE 15 CDSMD?

While AI lacks a universally accepted definition,[51] the EU AI-Act[52] describes it as a system capable of inferring outputs from inputs.[53] This article will focus on Generative AI, a special branch of AI dedicated to drafting new content,[54] and specifically on Large Language Models (LLMs), as a particular form of Generative AI.[55] These produce new textual content by recognizing patterns

Affairs 2017), 2 <https://www.europarl.europa.eu/RegData/etudes/BRIE/2017/596834/IPOL_BRI(2017)596834_EN.pdf> last accessed 13.05.2025.

46 Case C-476/17 *Pelham* ECLI:EU:C:2019:624, Opinion AG Szpunar ECLI:EU:C:2018:1002 para 30.

47 E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 277, 278.

48 Case C-476/17 *Pelham* ECLI:EU:C:2019:624, Opinion AG Szpunar ECLI:EU:C:2018:1002 para 30.

49 A., Hayes, "Investment: How and Where to Invest", (08 May 2025) <https://www.investopedia.com/terms/i/investment.asp> last accessed 13.05.2025.

50 Recitals 54 & 55 CDSMD; E., Treppoz., "The Past and Present of Press Publishers' Rights in the EU", (2023), 46 (3) Colum. J.L. & Arts, 276.

51 M. U., Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies, (2016), 29(2) JOLT, 359, for a detailed discussions <https://jolt.law.harvard.edu/Articles/pdf/v29/29HarvJLTech353.pdf> last accessed 14.05.2025.

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

53 S. Art. 3 (1) AI-Act.

54 J. L., Gillotte, "Copyright Infringement in AI-generated Artworks", (2020), 53(5) U.C.Davis L. Rev., 2661 <https://lawreview.law.ucdavis.edu/archives/53/5/copyright-infringement-ai-generated-artworks> last accessed 14.05.2025.

55 S., Warudkar & R., Jalit, "Unlocking the Potential of Generative AI in Large Language Models" in proceedings of the 2024 Parul International Conference on Engineering and Technology (PICET), 2 <https://doi.org/10.1109/PICET60765.2024.10716156> last accessed 14.05.2025.

in massive text datasets.[56] The technological disruption these models pose was not anticipated by the PPR. Unlike traditional aggregators, LLMs can ingest vast quantities of press content, distil its substance, and return user-specific outputs – thus eliminating referral traffic and undermining the economic sustainability of quality journalism.[57] Given the Directive's objective to safeguard the sustainability of quality journalism,[58] it is imperative that the PPR be interpreted dynamically to accommodate technological developments. Article 2 of the InfoSoc-Directive, incorporated into Article 15 CDSMD, adopts a technologically neutral definition of reproduction that includes reproduction "by any means and in any form".[59] This formulation supports the adaptability of reproduction rights to new processes such as AI training.

3.1 Understanding AI Systems and Their Training Processes

Generative AI, particularly LLMs, function through natural language processing to predict textual sequences based on previously observed patterns.[60] These models are trained on vast corpora[61] using machine learning architectures – especially transformers – that convert text into numerical representations (tokens) and encode semantic relationships through layers of weighted nodes known as neural networks.[62] The core stages involve data collection, pre-processing, which relates to preparing inputs by removing irrelevant data and segmenting text into tokens,[63] followed by the model training itself.[64] A useful pedagogical analogy likens this process to the education of a law student who, by analysing diverse case law, internalises legal principles to apply them to new factual scenarios.[65] Similarly, LLMs iteratively adjust internal parameters to better predict textual outcomes, based on exposure to large volumes of structured training data.

3.2 Is AI Reproducing?

To determine whether Generative AI Training infringes the reproduction right under Article 15 CDSMD, it is essential to assess the discrete stages of the training process where reproduction may occur. Scholarly analyses increasingly converge on the conclusion that reproduction in the light of copyright takes place at several levels, particularly during the initial acquisition.[66] Whether this can be transferred to the related right of press publishers will be analysed in the following.

3.2.1 Dataset Compilation

The first stage – dataset compilation – typically involves the use of automated web scraping tools to extract content, often in HTML format,[67] from online sources.[68] Although HTML structures text using technical tags, it still captures and reproduces the original editorial content, including headlines and introductory paragraphs[69] – elements that exemplary embody the publisher's investment through phrasing and structure. Applying the tripartite test for infringement under Article 15 CDSMD – recognisability, contribution to value, and substitution potential – the web scraping of news websites readily satisfies all three criteria. The editorial structure and substantive content remain recognisable in HTML, as the underlying text is typically reproduced verbatim and the fundamental structural elements are preserved through

56 i.e. N., Lucchi, "ChatGPT: A Case Study on Copyright Challenges for Generative Articifical Intelligence Systems", (2024),15(3) EURJRR, 603 <https://doi.org/10.1017/err.2023.59> last accessed 13.05.2025.

57 Gartner Inc, Gartner Predicts Search Engine Volume Will Drop 25% by 2026, Due to AI Chatbots and Other Virtual Agents, (Press Release, 2024) < https://www.gartner.com/en/newsroom/press-releases/2024-02-19-gartner-predicts-search-engine-volume-will-drop-25-percent-by-2026-due-to-ai-chatbots-and-other-virtual-agents> last accessed 14.05.2025; A., Schiffrin & H. Mateen, "Startup Aims To Help Publishers Collect Fees from AI Companies", (2024) <https://www.techpolicy.press/startup-aims-to-help-publishers-collect-fees-from-ai-companies/> last accessed 13.05.2025 & G., De Vynck & C., Zakrzewski, "Web publishers brace for carnage as Google adds AI answers", (2024) <https://www.washingtonpost.com/technology/2024/05/13/google-ai-search-io-sge/> last accessed 14.05.2025.

58 Recital 53 CDSMD.

59 R. Ducato & A., Strowel, "Ensuring text and data mining: remaining issues with the EU copyright exceptions and possible ways out", (2021), 43(5) E.I.P.R., 338 footnotes 79, 80 mentioning that there are other ways of defining technological neutrality.

60 M., Senftleben, "Remuneration for AI Training – A New Source of Income for Journalists?", (2024), 4 forthcoming in M., Senftleben and Others (eds), The Cambridge Handbook of Media Law and Policy in Europe, Cambridge University Press; N., Lucchi, "ChatGPT: A Case Study on Copyright Challenges for Generative Articifical Intelligence Systems", (2024),15(3) EURJRR, 603.

61 M., Iglesias Portela, S., Shamuilia & A., Anderberg, "Intellectual Property And Artificial Intelligence. A literature review", (Publications Office of the European Union 2019), 10 <https://op.europa.eu/sv/publication-detail/-/publication/912bc3f8-7d67-11eb-9ac9-01aa75ed71a1/language-en> last accessed 13.05.2025.

62 M., Senftleben, "Remuneration for AI Training – A New Source of Income for Journalists?", (2024), 4 forthcoming in M., Senftleben and Others (eds), The Cambridge Handbook of Media Law and Policy in Europe, Cambridge University Press; A., Zewe, "Explained: Generative AI", (2023); J. L., Gillotte, "Copyright Infringement in AI-generated Artworks", (2020), 53(5) U.C.Davis L. Rev., 2661; EUIPO, "The Development of Generative Artificial Intelligence from a Copyright Perspective", (2025), 26 <https://www.euipo.europa.eu/sv/publications/genai-from-a-copyright-perspective-2025> last accessed 13.05.2025.

63 EUIPO, "The Development of Generative Artificial Intelligence from a Copyright Perspective", (2025), 30.

64 See all stages W., Huang & X., Chen, "Does Generative AI copy? Rethinking the right to copy under copyright law", (2025), 56 CLSR, 2 <https://doi.org/10.1016/j.clsr.2024.106100> last accessed 14.05.2025 confirmed by the EUIPO, "The Development of Generative Artificial Intelligence from a Copyright Perspective", (2025), 30, 128.

65 Analogy derived from V., Lindberg, "Building and Using Generative Models under US Copyright Law", (2023), 18(2) Rutgers Bus. L.R., 6,7 <https://ssrn.com/abstract=4464001> last accessed 14.05.2025.

66 W., Huang & X., Chen, "Does Generative AI copy? Rethinking the right to copy under copyright law", (2025), 56 CLSR, 2.

67 I., Vistorskyte, "News Scraping: Everything You Need to Know", (2021) <https://oxylabs.io/blog/news-scraping> last accessed 14.05.2025.

68 I., Cohen, "From Headlines to AI: Narrowing the Bargaining Gap between News and AI companies", (2024), 1, 6, 7 < https://dx.doi.org/10.2139/ssrn.4878254> last accessed 14.05.2025.

69 A., Sellers, "Twenty Years of Web Scraping and the Computer Fraud and Abuse Act", (2018), 24 Boston Journal of Science & Technology Law, 384, 386 <https://scholarship.law.bu.edu/faculty_scholarship/465/?utm_source=scholarship.law.bu.edu%2Ffaculty_scholarship%2F465&utm_medium=PDF&utm_campaign=PDFCoverPages> last accessed 14.05.2025; A., Sharma, "Introduction to HTML (Hyper Text Markup Language) – A Review Paper", (2018), 7(5) IJSR, 1337 < https://www.ijsr.net/getabstract.php?paperid=ART20182355> last accessed 14.05.2025.

HTML mark-up.[70] This technical representation maintains the investment inherent in both the linguistic formulation and the organisational layout of the original publication. The components extracted – most notably headlines, lead paragraphs, and introductory summaries – are of particular economic relevance, given their role in capturing user attention, enhancing search engine visibility, and driving traffic. Increased user engagement directly correlates with advertising revenue, thereby evidencing a clear contribution to the publication's economic value. Lastly, the systematic aggregation and ingestion of such content by Generative AI systems facilitates the generation of outputs that may serve as functional substitutes for original press content. While complete market substitution has not yet materialised, the legal criterion of substitution under the developed test does not require actual displacement, but merely the potential for such an effect. Accordingly, the indirect but substantial substitution potential affirms the legal relevance of this early-stage act of reproduction.

3.2.2 Pre-Processing Stage

Following dataset compilation, raw text undergoes preprocessing, so data cleaning and tokenisation. During tokenisation, the text is fragmented into units, singular words, word parts, numbers and punctations, that get assigned a numerical value. These so-called tokens can be algorithmically analysed.[71] Whether this process constitutes reproduction under Article 15 CDSMD is less clear. At this point the meaning of "recognisability"[72] would be challenged. While traditional interpretations of "recognisability" would rely on perceptibility to human users, a broader, technologically informed view might encompass algorithmic recognisability, particularly if tokens retain structural or semantic traces of the original content. Nonetheless, the fragmented and abstracted nature of tokens challenges their economic and communicative value. Furthermore, the exclusion of "individual words", while not judicially defined yet, strengthens the implication that tokens – often smaller than words – are unlikely to meet the threshold for reproduction. Therefore, although arguable under a non-exhaustive test, tokenisation alone appears to be insufficient to establish infringement in most cases.

3.2.3 The Model Itself

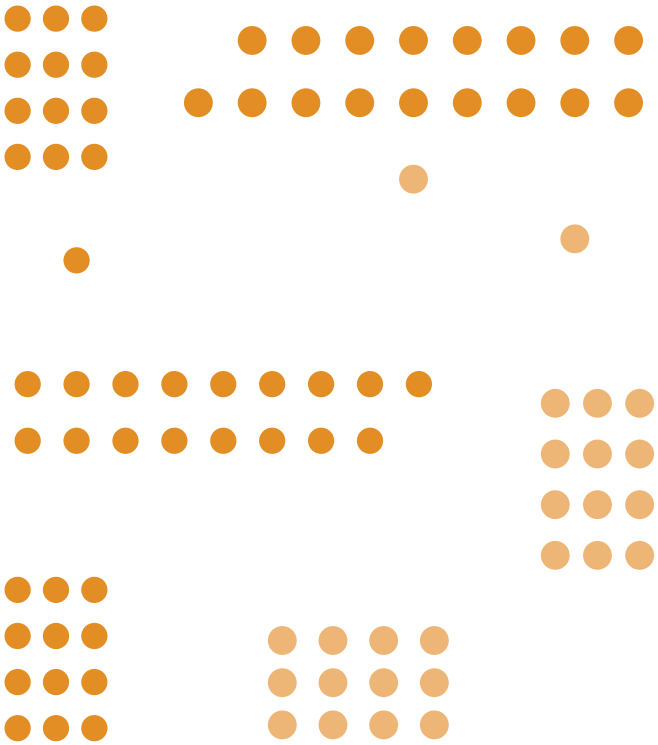
The final consideration is whether reproduction occurs within the trained model itself. LLMs encode knowledge through adjustments in neural weights and statistical correlations rather than by storing literal content.[73] These

70 A., Sellers, "Twenty Years of Web Scraping and the Computer Fraud and Abuse Act", (2018), 24 Boston Journal of Science & Technology Law, 384, A., Sharma, "Introduction to HTML (Hyper Text Markup Language) – A Review Paper", (2018), 7(5) IJSR, 1337.

71 EUIPO, "The Development of Generative artificial Intelligence from a Copyright Perspective", (2025), 145–149, inter alia with the example of ChatGPT.

72 For recognisability in *Pelham* see E., Rosati, "Of tables and other furniture: AG Szpunar advises CJEU on originality (but also proposes adoption of recognisability test for infringement), (2025) < https://ipkitten.blogspot.com/2025/05/of-tables-and-other-furniture-ag.html> last accessed 14.05.2025; J., Kiiski, "Recognising music samples – whose ear to trust in IP?", (2024), 46(10) E.I.P.R., 676–683.

73 EUIPO, "The Development of Generative Artificial Intelligence from a Copyright Perspective", (2025), 151.



distributed representations lack perceptibility and do not enable direct retrieval of protected material. Accordingly, recognisability and value contribution are virtually non-existent at this stage. Furthermore, the CDSMD's recitals suggest that relevant acts of copying occur during data preparation, not within the internal structure of the trained model. Thus, reproduction in the legal sense is not sustained at this level.

3.2.4 Interim Conclusion

Generative AI Training implicates the reproduction right under Article 15 CDSMD primarily during the data acquisition phase, where web scraping results in the capture and storage of protected content. While later stages such as tokenisation and model training involve substantial transformation, they present a weaker case for infringement due to diminished recognisability and commercial relevance of the singular parts. Accordingly, legal enforcement of the PPR in the context of AI training should focus on the early- stage act of web scraping, which most directly interferes with the press publishers' ability to recoup their investment.

4. EXCEPTIONS, LICENSING AND FUTURE IMPLICATIONS

Assuming, as this article has argued, that the training of Generative AI models constitutes acts of infringing reproduction under Article 15 CDSMD, it becomes necessary to examine the potential applicability of relevant exceptions. In the absence of such exceptions, licensing remains the necessary legal mechanism to authorise such use.

4.1 Exceptions

While the applicability of the Text and Data Mining (TDM) exceptions under Articles 3 and 4 CDSMD is acknowledged, their analysis is excluded due to the commercial nature of most AI training, thereby rendering Article 3 CDSMD inapplicable, and the unresolved legal uncertainty surrounding the opt-out mechanism under Article 4 (3) CDSMD.[74] The only remaining potentially applicable provision is the exception for temporary acts of reproduction under Article 5(1) of the InfoSoc-Directive. The temporary reproduction exception requires five cumulative conditions to be fulfilled: the act must be (1) temporary; (2) transient or incidental; (3) an integral part of a technological process; (4) serve either lawful use or transmission between third parties; and (5) lack independent economic relevance.[75] These criteria were originally designed to ensure the technical operability of the internet, balancing broad reproduction rights with the need for technological innovation. Whether these conditions apply to AI training processes remains contested.[76] In LAION, the Hamburg Regional Court held that the reproduction of photographs for an AI training dataset did not meet the necessary requirements, particularly because the copies were not deleted automatically and because their function was preparatory rather than incidental.[77] While this national ruling is instructive, it is not biding at the EU level, and the Court of Justice of the European Union (CJEU) has not yet addressed the issue. In the absence of authoritative clarification, licensing emerges as the more secure legal avenue for both rightholders and AI developers.

4.2 Licensing

Where no exception applies and the PPR is infringed, licensing becomes essential. Furthermore, licensing offers not only greater legal certainty – particularly in contrast to the unresolved requirements of exceptions such as the opt-out mechanism under Article 4 (3) CDSMD – but also serves to address broader ethical and societal considerations. Generative AI systems depend on human-created journalistic content – often accessed without authorisation or compensation.[78] Licensing ensures fairness, supports new revenue streams for press publishers,[79] and helps sustain professional journalism,

74 i.e. N., Lucchi, "ChatGPT: A Case Study on Copyright Challenges for Generative Articifical Intelligence Systems", (2024), 15(3) EURJRR, 616.

75 EUIPO, "The Development of Generative Artificial Intelligence from a Copyright Perspective", (2025), 50,51.

76 EUIPO, "The Development of Generative Artificial Intelligence from a Copyright Perspective", (2025), 51.

77 *Kneschke v LAION*, LG Hamburg, Judgement of 27 September 2024 para 62, 66.

78 Initiative Urheberrecht, "Authors and Performers Call for Safeguards Around Generative AI in the European AI Act", (2023), 2 <https://urheber.info/diskurs/call-for-safeguards-around-Generative-ai> last accessed 14.05.2025.

79 Maverick Publishing Specialists, "Licensing content to Generative AI platfroms: a pubisher's perspective", (2025) <https://www.maverick-os.com/news-events/news/licensing-content-to-Generative-ai-platforms-

which plays a critical role in democratic discourse.[80] These concerns are reflected in the legislative history of the AI-Act. Recital 105 affirms that the use of protected content requires prior authorisation, unless a statutory exception applies.[81] Author and performer organisations have repeatedly stressed the need for consent, remuneration, and human-centric AI development.[82] Such advocacy has shaped industry practices: some rightholders have turned to litigation while other have signed licensing deals with AI developers.[83] Although these agreements are often confidential,[84] a *Reuters institutes survey* found that a majority of publishers favour collective licensing frameworks benefiting the sector as a whole over individual negotiations.[85]

However, the appropriate structure of such licensing frameworks remains debated. Individual licensing offers flexibility[86] but is often impractical due to the volume of content and number of rightholders involved.[87] In the press publishing sector, this situation is somewhat simplified by the fact that publishers frequently control bundled rights, having acquired author rights contractually.[88] Still, high transaction costs and imbalanced negotiating power make one-to-one licensing unsustainable – particularly for smaller or regional publishers with limited market leverage.[89]

Collective licensing, administered by Collective Management Organisations (CMOs),[90] provides a more viable solution.[91] It allows for the aggregation of rights, simplifies negotiating processes, and can be tailored to the need of specific sectors.[92] Recent developments, such as the Copyright Clearance Centre's (CCC) introduction of AI-specific licensing tools,[93] indicate the growing feasibility of such schemes.[94] Nonetheless, collective licenses face challenges, including limited representativeness of CMOs and difficulties allocating revenue – especially given the opacity of AI training processes.[95]

More far-reaching is the proposal for statutory[96] or extended collective licensing (ECL), such as that in Spain's draft Royal Decree.[97] ECL allows licenses granted by CMOs to apply to non-members, provided opt-out options are available.[98] While this addresses the scale issue, it risks overriding rightholder autonomy and raises practical difficulties, such as the effectiveness of post-training opt-outs.[99] Although Article 12 CDSMD allows for ECL in situations where individual licensing is impractical,[100] its use remains controversial. It may offer legal coverage, but its automatic inclusion of non-con-

a-publishers-perspective/> last accessed 14.05.2025; M., Senftleben, "Remuneration for AI Training – A New Source of Income for Journalists?", (2024), 4 forthcoming in M., Senftleben and Others (eds), The Cambridge Handbook of Media Law and Policy in Europe, Cambridge University Press; N., Newman & Cherubini, F., "Journalism, media, and technology trends and predictions 2025", (Reuters Institute 2025) <https://doi.org/10.60625/risj-vte1-x706> last accessed 14.05.2025.

80 M., Senftleben, "Remuneration for AI Training – A New Source of Income for Journalists?", (2024), 4 forthcoming in M., Senftleben and Others (eds), The Cambridge Handbook of Media Law and Policy in Europe, Cambridge University Press.

81 Recital 105, AI-Act.

82 Authors', Performers' and Other Creative Workers' Organisations, "Joint Statement on Artificial Intelligence and the Draft AI Act", (2023), 1 <https://screendirectors.eu/joint-statement-on-artificial-intelligence-and-the-draft-eu-ai-act/> last accessed 14.05.2025.

83 C., Tobitt, "Who's suing AI and who's signing: Ziff Davis sues OpenAI after Washington Post signs deal. 14 major publishers sue AI start-up Cohere Inc.", (2025).

84 G., Kahn, "How AI is reshaping copyright law and what it means for the news industry", (Reuters Institute 2025) <https://reutersinstitute.politics.ox.ac.uk/news/how-ai-reshaping-copyright-law-and-what-it-means-news-industry> last accessed 14.05.2025.

85 N., Newman & Cherubini, F., "Journalism, media, and technology trends and predictions 2025", (Reuters Institute 2025).

86 D., Gervais and Others, "The Heart of the Matter: Copyright, AI Training, And LLM's" (2024), 71 Journal of the Copyright Society, 27 <https://copyrightsociety.org/wp-content/uploads/2025/04/713_The-Heart-of-the-Matter.pdf> last accessed 14.05.2025.

87 R., Matulionyte, "Generative AI and Copyright: Exception, Compensation or Both?", (2023), 134 IPF, 5 <https://dx.doi.org/10.2139/ssrn.4652314> last accessed 14.05.2025.

88 S., Karapapa, "The Press Publishers Right under EU Law – Rewarding Investment through Intellectual Property" in E., Bonadio & P., Goold (eds), The Cambridge Handbook of Investment-Driven Intellectual Property, (CUP 2023), 164; M., Stratton, "Market-Based Licensing for Publishers' Works is Feasible. Big Tech Agrees.", (forthcoming 2025), 48 Colum. J.L. & Arts, 7 <https://dx.doi.org/10.2139/ssrn.5072814> last accessed 14.05.2025.

89 C., Geiger & V., Iaia, "The forgotten creator: Towards a statutory remuneration right for machine learning of Generative AI", (2024), 52 CLSR, 12 <https://doi.org/10.1016/j.clsr.2023.105925> last accessed 14.05.2025; M., Stratton, "Market-Based Licensing for Publishers' Works is Feasible. Big Tech Agrees.", (forthcoming 2025), 48 Colum. J.L. & Arts, 7 citing Andreessen Horowitz, Comments on the US Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright (2023), 8 <https://www.regulations.gov/comment/COLC-2023-0006-9057> last accessed 14.05.2025; I., Cohen, "From Headlines to AI: Narrowing the Bargaining Gap between News and AI companies", (2024), 13.

90 R., Matulionyte, "Generative AI and Copyright: Exception, Compensation or Both?", (2023), 134 IPF, 5.

91 D., Gervais and Others, "The Heart of the Matter: Copyright, AI Training, And LLM's" (2024), 71 Journal of the Copyright Society, 27.

92 D., Gervais and Others, "The Heart of the Matter: Copyright, AI Training, And LLM's" (2024), 71 Journal of the Copyright Society, 27.

93 Copyright Clearance Centre, "CCC announces AI Systems Training License for the External Use of Copyrighted Works coming soon", (2025) <https://www.copyright.com/media-press-releases/ccc-announces-ai-systems-training-license-for-the-external-use-of-copyrighted-works-coming-soon/> last accessed 14.05.2025; for more information about this new type of licence s. Copyright Clearance Centre, "Responsible AI Starts with Licensing" <https://www.copyright.com/solutions-annual-copyright-license/business/> last accessed 14.05.2025.

94 R., Levy, "Navigating Copyright in the Age of Generative AI: Responsible AI Starts with Licensing", (2024); now also in Japan according to Copyright Clearance Centre, "Japan Academic Association for Copyright Clearance and RightsDirect Japan Announce the Availability of AI Re-Use Rights for Digital Copyright License" (2025) <https://www.copyright.com/media-press-releases/japan-academic-association-for-copyright-clearance-and-rightsdirect-japan-announce-the-availability-of-ai-re-use-rights-for-digital-copyright-license/> last accessed 14.05.2025.

95 R., Matulionyte, "Generative AI and Copyright: Exception, Compensation or Both?", (2023), 134 IPF, 5,6.

96 S. inter alia C., Geiger & V., Iaia, "The forgotten creator: Towards a statutory remuneration right for machine learning of Generative AI", (2024), 52 CLSR, 12f.

97 T., Nobre, "A first look at the Spanish proposal to introduce ECL for AI training", (2024), <https://copyrightblog.kluweriplaw.com/2024/12/11/a-first-look-at-the-spanish-proposal-to-introduce-ecl-for-ai-training/> last accessed 14.05.2025.

98 Article 12 (1) CDSMD.

99 US Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-Publication Version), (2025),101 <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf> last accessed 14.05.2025.

100 Article 12 CDSMD as described and cited in T., Nobre, "A first look at the Spanish proposal to introduce ECL for AI training", (2024).

senting rightholders raises concerns about the erosion of exclusive rights.[101] This concern remains even more pronounced with statutory licensing. If we legally require a high level of protection for right holders[102], then forcing creators and publishers/other related right holders into statutory licensing without even the option to opt-out undermines that principle. It treats their works as public infrastructure – not protected expressions.

4.3 Interim Conclusion

Generative AI is reshaping how society produces and consumes information. While many remain sceptical of AI-generated news,[103] especially in politically sensitive contexts,[104] younger demographics show more openness.[105] As trust becomes a core concern,[106] access to high-quality, verifiable training data is essential – precisely what licensing enables. The relationship between AI developers and press publishers is interdependent: the former require high quality journalistic content, while the latter depend on fair compensation to continue producing it. Licensing is thus not merely a legal formality but a structural necessity. While ECL may offer broad coverage, it risks overreach. Individual licensing, though principled, lacks the scale of an industry solution. Collective licensing via CMOs offers the most balanced solution: it preserves rightholder autonomy, allows for coordinated rights management, and facilitates lawful AI training practices without compromising democratic values.

5. CONCLUSION

This article has analysed whether the training of Generative AI systems infringes the reproduction right granted under Article 15 CDSMD and, if so, what form of licensing is most appropriate in response. Applying a functional three-part test – assessing recognisability, value contribution, and substitution potential – it was shown that the most legally relevant act of reproduction occurs during dataset compilation via web scraping. Later stages, such as tokenisation and what is represented within the model itself, are less clearly infringing on their own. Given the absence of applicable (or practical) exceptions, licensing emerges as the necessary legal response. While individual licensing is burdensome and ECL potentially overreaching, collective licensing through CMOs offers a proportionate and workable middle ground. For press publishers who often control a coherent bundle of rights, CMOs are structurally well-positioned to facilitate such licensing efficiently. Ultimately, the viability of both Generative AI and the independent press sector depends on creating a legal and economic framework in which both can coexist. Licensing is not a barrier to innovation but a foundation for a sustainable digital ecosystem – one in which rights, quality journalism, and democratic values are respected and preserved.

Klara Schinzler

Klara is a recent graduate of the LL.M. program in European Intellectual Property Law at Stockholm University. Her thesis focused on the Press Publishers' Right under Article 15 of the Copyright in the Digital Single Market Directive (CDSMD) and its relevance to the training of generative AI models. Prior to her studies in Stockholm, she earned her first State Examination in Law at the University of Leipzig (Germany) in May 2024. Klara is now looking to build on her experience in Stockholm through an internship before returning to Germany to complete her legal clerkship and continue on the path to becoming a qualified lawyer.

101 US Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-Publication Version), (2025), 100.

102 i.e. Case C-5/08 *Infopaq International* ECLI:EU:C:2009:465 para 42–42 & Case C-476/17 *Pelham* ECLI:EU:C:2019:624 para 30 cited in E., Rosati, "Copyright in the Digital Single Market: Article-by-Article Commentary on the Provisions of Directive 2019/790", (OUP 2021), 264.

103 F., Simon, "Neither humans-in-the-loop nor transparency labels will save the news media when it comes to AI", Figure 17, (Reuters Institute 2024) <https://reutersinstitute.politics.ox.ac.uk/news/neither-humans-loop-nor-transparency-labels-will-save-news-media-when-it-comes-ai> last accessed 14.05.2025.

104 F., Simon, "Neither humans-in-the-loop nor transparency labels will save the news media when it comes to AI", Figure 18, (Reuters Institute 2024).

105 F., Simon, "Neither humans-in-the-loop nor transparency labels will save the news media when it comes to AI", Figure 17, (Reuters Institute 2024).

106 F., Simon, "Neither humans-in-the-loop nor transparency labels will save the news media when it comes to AI", Figure 19, (Reuters Institute 2024).