FACULTY OF LAW
Lund University

Mónika Trombitás Andersson

Rubberducking EU copyright law
Generative AI, machine learning and infringement

LAGF03 Essay in Legal Science

Bachelor Thesis, Master of Laws program
15 higher education credits

Supervisor: Aurelija Lukoševičienė

Term: Autumn 2023

Read a thousand books and your words will flow like a river.

Virginia Woolf

Contents

Summary

Whether generative AI infringes copyright holders' economic rights is a normative question that, due to the requirement of legal certainty inherent to the principle of rule of law should have an answer. And yet, it is not a clear-cut case. This is partly because, due to the wide variety of generative AI tools available on the market or under development, an abstract analysis on the legality of all of these has its limitations. More importantly, the answer depends on a range of different legal criteria, the interpretation of which is unclear at the moment. Therefore, interpretative guidance from the European courts is desirable for the sake of clarity and legal predictability.

Sammanfattning

Huruvida generativ AI utgör intrång i upphovsrättsinnehavarens ekonomiska rättigheter är en normativ fråga som, på grund av rättssäkerhetsprincipen som är en viktig del av rättsstatsprincipen borde kunna besvaras. Ändå verkar det inte finnas något entydigt svar. Detta beror delvis på svårigheterna med att utföra analysen endast på en abstrakt nivå när det finns en betydligt stor variation av generativa AI-verktyg på marknaden. Ännu viktigare är att svaret beror på en rad olika kriterier som för närvarande verkar att vara oklara eller oprecisa. Ett avgörande från en europeisk domstol om artikel 4 DSM-direktivet är önskvärd för rättssäkerheten.

Abbreviations

AI   artificial intelligence

CJEU   Court of Justice of the European Union

EP   European Parliament

EU   European Union

JRC   Joint Research Centre

TDM   text and data mining

1 Introduction

Although the artificial intelligence phenomenon appears undeniably current – 2023 has been declared the year AI went mainstream[1] – humanity's hopes and fears for new technologies, including intelligent machines, are nothing new.[2] Ancient fears aside, scepticism against the rapid development of AI tools is now on the rise it may seem,[3] while many remain optimistic.[4] Due to the wide variety of possible applications, a legal analysis on modern AI is viable from several aspects such as data protection, privacy, criminal law, and migration law for instance. The EU has been moving towards regulation of AI technologies since 2020[5] with the aim to ensure that AI systems are safe, transparent, traceable, non-discriminatory and environmentally friendly.[6] The proposed legislation (Artificial Intelligence Act) was published this year.[7] In

---

[1] See for instance: Wilkins, A. (2023) '2023 was the year that artificial intelligence went mainstream', *New Scientist.* Available at: https://www.newscientist.com/article/mg26034693-900-2023-was-the-year-that-artificial-intelligence-went-mainstream/; Associated Press (2023) '2023 was the year AI went mainstream. It was also the year we started to panic about it', *euronews.next.* Available at: https://www.euronews.com/next/2023/12/27/2023-was-the-year-ai-went-mainstream-it-was-also-the-year-we-started-to-panic-about-it; Dib, D. (2023) 'How AI shook the world in 2023', *rest of the world.* Available at: https://restofworld.org/2023/the-year-of-ai-boom-2023/.

[2] Mayor, A. (2018) 'An AI Wake-Up Call From Ancient Greece', *Project Syndicate.* Available at: https://www.project-syndicate.org/commentary/artificial-intelligence-pandoras-box-by-adrienne-mayor-2018-10; Truitt, E.E. (2023). 'Surveillance, Companionship, and Entertainment: The Ancient History of Intelligent Machines', *THE MIT PRESS READER.* Available at: https://thereader.mitpress.mit.edu/the-ancient-history-of-intelligent-machines/.

[3] Metz, C. (2023) 'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead', *New York Times.* Available at: 'The Godfather of AI' Quits Google and Warns of Danger Ahead - The New York Times (nytimes.com). Yudowsky, E. (2022) 'AGI Ruin: A List of Lethalities', *AI Alignment Forum.* Available at: https://www.alignmentforum.org/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities.

[4] Langkilde, D. (2023) 'Why I am excited about AI' Available at: https://lang-kilde.se/blog/2023/6/8/why-i-am-excited-about-ai.

[5] See press release of the two legislative initiatives and one report on AI: https://www.europarl.europa.eu/news/en/press-room/20201016IPR89544/parliament-leads-the-way-on-first-set-of-eu-rules-for-artificial-intelligence.

[6] See Recital 1 of AI Act Proposal (text version of 14 June 2023).

[7] Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(1), available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html. On 9 December 2023, the Council and the Parliament reached a provisional agreement on the conceptual basis for a compromised text that will be then submitted to Coreper, the body comprised of the representatives of each Member States, in 2024. Hereinafter: AI Act Proposal.

addition, 2023 was also the year when companies made highly capable *generative* AI products – that is AI algorithms that can generate new content, such as images, videos, music, speech, text, software code, and product design – available to the general public,[8] followed by strong public disapproval from professionals within the creative industry, i.e. journalists, literary authors, graphic designers, and photographers. Opinion articles titled such as 'Is generative AI Stealing From Artists?',[9] 'The Future of Creativity is Threatened by AI – Artist in Danger',[10] and 'We must declare jihad against AI'[11] are just a few memorable examples. At the time this paper is submitted, there are several lawsuits pending against developers of generative AI systems, particularly in the US[12] and the UK,[13] with copyright infringement among the claims. Similar – or at least similarly publicized –cases are yet to be filed in an EU Member State.

As far as generative AI's society-wide disruptive effects are concerned, time will tell if the anti-AI sentiment is only alarmism or concerns are indeed justified. Whether generative AI infringes copyright holders' economic rights (or as popularly expressed "is stealing from artists") is on the other hand a

---

[8] Meta's chatbot launched in September 2023, and then was followed by Open AI's ChatGPT 3 in November 2023. Solutions currently on the market include OpenAI's GPT-4 and DALL E 2, Microsoft Copilot, Kapwing AI Video Generator just to name a few.

[9] Marr, B. (2023). 'Is generative AI Stealing From Artists?', *Forbes*. Available at: https://www.forbes.com/sites/bernardmarr/2023/08/08/is-generative-ai-stealing-from-artists/.

[10] Torres, E. (2023). 'The Future of Creativity is Threatened by AI – Artist in Danger'. *Bootcamp*. Available at: https://bootcamp.uxdesign.cc/the-future-of-creativity-2dfc568ccaa2.

[11] Cuenco, M. (2023). 'We must declare jihad against AI', *Compact*. Available at: https://compactmag.com/article/we-must-declare-jihad-against-a-i.

[12] See for instance: "*Getty images v. Stability AI,* challenging the legality of products such as Stable Diffusion and DreamStudio; Class action *Doe v. Github, Inc.*, challenging the legality of GitHub Copilot (and a related product, OpenAI Codex, which powers Copilot), filed against GitHub, Microsoft and OpenAI; Class action Authors Guild and Others as Plaintiffs v. OpenAI and Microsoft as Defendants, complaint filed 4 December 2023, available at: https://authorsguild.org/app/uploads/2023/12/Authors-Guild-OpenAI-Microsoft-Class-Action-Complaint-Dec-2023.pdf; The New York Times Company as Plaintiff v. Microsoft and OpenAi (various entities) as Defendants, complaint filed on 27 December 2023 available at: https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf.

[13] See for instance: Case No: IL-2023-000007 between Getty Images (various companies) as Claimants and Stability AI Ltd. as Defendant. Preliminary ruling of 1 December 2023 available at: https://s3.documentcloud.org/documents/24183636/getty-images-v-stability-ai-uk-ruling.pdf. The case was deemed to have sufficient merit for trial.

normative question that, due to the requirement of legal predictability and certainty inherent to the principle of rule of law,[14] urgently requires an answer.

Ever since copyright law has existed in the Western world,[15] it has been intertwined with technology and innovation. The very idea of protecting intellectual creations (a poem or a novel) – as opposed to the physical manifestations thereof (a codex or a book) – by means of property rights became first relevant after a series of ground-breaking inventions. Gutenberg's printing press and subsequent technological advancements made copying and distributing written text more effective and affordable. The previous, feudal system of privileges proved to be inefficient for renumerating authors and an exclusive property right emerged instead.[16] Technological advancement was not only the spark that set off codification of authors' exclusive rights (copyright meaning what is worth copying, is worth protecting) but it remained one of the most important factors shaping copyright law's development. Inventions such as the cinema, cable television, photocopying machines, various sound and image carriers (cassette tapes, VCRs, CDs, etc.) posed challenges to this field of law and thereby contributed to its development, refined its set of rules. Most recently, it has been the Internet's emergence that shook the foundations of copyright law.[17] E-book readers, virtual book publishing, digital rights management software all tested

---

[14] CDL-AD(2011)003rev-e - Report on the rule of law - Adopted by the Venice Commission at its 86th plenary session (Venice, 25-26 March 2011), para 41.

[15] It is customary to refer to the British Statute of Ann 1710 as the first codified copyright law, which was followed by several revolutionary codifications in the US (Copyright Act of 1790) and France (Chénier Act 1793). Laws akin to modern copyright law were however first codified in Europe and in the US first during the 19th century.

[16] Joyce, C. (ed.) (2013). Copyright Law. 9th ed. New Providence, LexisNexis, pp.17-19; Lontai, E., Faludi G., Vékás, G. (ed.) (2012). *Szerzői jog és iparjogvédelem: [oktatási segédanyag]*. Budapest : Eötvös J. Kvk., pp. 13-14.; Bettig, R.V. (1996) *Copyrighting culture : the political economy of intellectual property.* Westview Press (Critical studies in communication and in the cultural industries), p. 11.

[17] Marshall, E. 'Copyright Obsolete in An Electronic Age, OTA Finds'. (1986) *Science*, New Series, Vol. 232, No. 4750 (May 2, 1986), p. 572 (1 page); Mills, M. L. (1989) 'New Technology and the Limitations of Copyright Law: An Argument for Finding Alternatives to Copyright Legislation in an Era of Rapid Technological Change', *Chicago-Kent Law Review*. 65(1). Available at: https://scholarship.kentlaw.iit.edu/cklawreview/vol65/iss1/14; European Commission's Green Paper on 'Copyright and the Challenge of Technology'. COM (88) 172 Final, 7.6.1988. Available at: https://op.europa.eu/en/publication-detail/-/publication/f075fcc5-0c3d-11e4-a7d0-01aa75ed71a1.

some of the traditional notions of copyright law,[18] not to mention the issues peer-to-peer file sharing brought.[19] Internet-related copyright problems are far from resolved[20]. In this digital environment, enforcement against unauthorized mass production and distribution, and thus mass infringement is also challenging because the infringements online typically have a cross-border nature.[21] It is undoubted that generative AI is yet another "stress test" for copyright law.[22] Some aspects are familiar: potential infringement on a large scale, geographically unlimited impact, and eventual difficulties in enforcement, whilst others are completely new. This paper aims to join the debate on *whether generative AI potentially infringes copyright, more specifically economic rights, under EU law*. And while it is true that there is no genuine "European Copyright Law" to speak of,[23] and the particularities of national copyright and private procedural laws play a pivotal rule in copyright infringement cases, there is sufficient convergence between national copyright laws with regard to economic rights so that a legal analysis with an EU focus may provide some relevant findings, even in a national context, such as a court case in any Member State.

---

[18] Mancini, A. (2006). *Copyright is Obsolete*. New York, Buenos Books.

[19] Coincidentally, the torrent website Pirate Bay started exactly 20 years ago, in 2003. The Pirate Bay-case was one of the most publicised copyright infringement cases at the time. For a legal commentary see: Asp, P Rosén, J. 'The Pirate Bay – en kommentar'. (2011) SvJT. Available (in Swedish) at: https://svjt.se/svjt/2011/103.

[20] Longan, M. (2021) 'Rethinking copyright and the Internet : a new model for users' rights'. Available at: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=edsble&AN=edsble.833961&site=eds-live&scope=site.

[21] Joyce, C. p. 16; (2013); Mills (1989).

[22] Lim, D. 'Generative AI and copyright: principles, priorities and practicalities', *Journal of Intellectual Property Law & Practice*, Volume 18, Issue 12, December 2023, Pages 841–842, https://doi.org/10.1093/jiplp/jpad081.

[23] Lucas-Schloetter, A. 'Is there a concept of European copyright law? History, evolution, policies and politics and the *acquis communautaire'* in *EU copyright law : a commentary*. Second edition (2021). Edward Elgar Publishing (Elgar commentaries), p. 12.

2   On the technology behind AI

2.1   Methodological challenges

In a field of law, which is so heavily intertwined with technology, it is paramount to "translate" data science terminology to useful and, for the uninitiated, comprehensible definitions and then compare these against the normative terminology of law. An additional challenge in the present case is that data science concepts in themselves – such as data mining for instance – are rather vague terms that may cover a great variety of actual computing activities, that is one purpose or goal can be reached in various technical manners. Because the relevant segment of data science is relatively new,[24] and because computing technologies evolve at a very rapid pace, the meaning of these technical concepts is currently changing.

Nonetheless, recognizing the necessity of policy making but aware of the challenges around definitions, the EU commissioned an inquiry for establishing "an operational definition and taxonomy of artificial intelligence" in 2020. Two JRC Technical Reports have been published, based on a collection of definitions published between 1955 and 2021.[25] For defining the relevant *technical concepts*, the findings of these reports, as well as the recitals of the AI Act Proposal, are indispensable, although not always sufficient. Therefore, in this chapter, this paper also relies on academic definitions, bearing in mind that none of these may be carved into stone just yet.

---

[24] The famous expression "data is the new oil" was coined by Clive Humby in 2006.

[25] Samoili, S., Lopez Cobo, M., Gomez Gutierrez, E., De Prato, G., Martinez-Plumed, F. and Delipetrev, B., (2020). 'AI WATCH. Defining Artificial Intelligence', EUR 30117 EN, Publications Office of the European Union, Luxembourg. Available at: https://publications.jrc.ec.europa.eu/repository/handle/JRC118163 and Samoili, S., Lopez Cobo, M., Delipetrev, B., Martinez-Plumed, F., Gomez Gutierrez, E. and De Prato, G., (2021). 'AI Watch. Defining Artificial Intelligence 2.0', EUR 30873 EN, Publications Office of the European Union, Luxembourg. Available at: https://publications.jrc.ec.europa.eu/repository/handle/JRC126426.

2.2   Technical concepts and definitions

Data mining is, in a technical sense, the process of extracting and discovering patterns in large data sets. The term *data mining* is a misnomer because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself. As Han, Kamber, and Pei (2012) wittily explain, "data mining should have been more appropriately named 'knowledge mining from data,' which is unfortunately somewhat long".[26]

Machine learning defined by the report 'AI Watch - Defining Artificial Intelligence 2.0' is the scientific study of computer algorithms that improve automatically through experience, and that can be considered as the practical implementation of artificial intelligence linked to theoretical models, such as generative models. Machine learning algorithms build a model based on training data in order to make predictions or decisions without being explicitly programmed to do so.[27] Machine learning approaches include, for instance, supervised, unsupervised and reinforcement learning, using a variety of methods including deep learning with neural networks.[28]

Although both *data mining* and *machine learning* are automated processes, they are distinctly separate technical concepts and are understood as such by practitioners.[29] The data sets required for data mining and machine learning respectively can be obtained in various ways, including gathering data by data crawling or more typically by data scraping, and purchasing data sets from Data-as-Service companies. Machine learning models do not rely on data mining in a technical sense. Data used in both types of processes may include copyrighted material.

---

[26] Han, J., Kamber, M., Pei, J. (2012). 'Data mining: concepts and techniques.' (3rd ed), para. 1.2. Available at: https://www-sciencedirect-com.ludwig.lub.lu.se/science/article/pii/B9780123814791000010?via%3Dihub.
[27] See page 13.
[28] AI Act Proposal, Recital 6 a.
[29] 'Difference Between Data Mining and Machine Learning. (2023). Available at: https://differencebetween.io/data-mining-and-machine-learning/; 'Difference Between Data mining and Machine learning'. (2023). Available at: https://www.geeksforgeeks.org/difference-between-data-mining-and-machine-learning/.

Generative artificial intelligence is a broad term referring to the technology as well as the product that can generate new content, such as images, videos, music, speech, text, software code, and product design. These are based on so-called *generative models* (a type of machine learning model, i.e. built by machine learning algorithms) that learn patterns and structure of input training data and then generate new data that has similar characteristics.[30]

Many generative AI tools are based on *foundation models,* pretrained models that serve as a basis for other AI models. These are general-purpose technologies that can support a wide variety of use cases,[31] including generative AI. Since content generation is only a segment of these use cases, scholars and legislators tend to focus on foundation models.[32] As the AI Act Proposal states in Recital 60 e, these models

> "are often trained on a broad range of data sources and large amounts of data to accomplish a wide range of downstream tasks, including some for which they were not specifically developed and trained. (…) AI systems with specific intended purpose or general purpose AI systems can be an implementation of a foundation model, which means that each foundation model can be reused in countless downstream AI or general purpose AI systems. These models hold growing importance to many downstream applications and systems."

Consequently, data mining in the technical, or narrow, sense is not necessary for the training of generative models, nor for creating the machine learning algorithms training the models.

---

[30] Pasick, A. (2023). 'Artificial Intelligence Glossary: Neural Networks and Other Terms Explained'. *The New York Times*. Available at: https://www.nytimes.com/article/ai-artificial-intelligence-glossary.html.

[31] Competition and Markets Authority (2023). AI Foundation Models: Initial Report. Available at: https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full_report_.pdf.

[32] For scholarly work see for instance: Henderson, P. et al. (2023) 'Foundation Models and Fair Use'. Available at: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=edsarx&AN=edsarx.2303.15715&site=eds-live&scope=site. For legislation, see AI Act Proposal.

2.3   Machine learning as reproduction

There are some differences between Member States laws in terms of the subjects of copyright, namely which types of works are covered and how: protected by copyright or neighbouring rights.[33] Nonetheless, literary works, photographs, digital art, and computer software are protected under copyright laws in every Member States.

A potential copyright infringement by generative AI may occur in two situations. *Firstly*, during machine learning. When generative models are trained, the input data is highly likely to include such protected material, considering the amount of data necessary to create generative models. The relevant machine learning process involves creating and storing multiple copies of copyrighted material, which affects the rightsholder's economic rights, above all the exclusive right of reproduction within the meaning of Article 2 InfoSoc.[34] A relevant, and often overlooked aspect is the so-called "memorisation" in more advanced generative models.[35] Ideally, it is only "training observations" that are memorised but such training observations might in fact coincide with copyrighted material, which would mean reproduction of a copyrighted item and storage of the same for unlimited time.[36]

---

[33] See for instance Article 1:1 Section 1 Point 5 and paragraph 5:49a of the Swedish Copyright Act (1960:729). Photographs might be protected under copyright law or under neighboring rights depending on the level of originality the photograph possesses. In other Member States, such as Hungary for instance, photographs are only protected under copyright law if they are original (Article 1(2)(i) of the Hungarian Copyright Act LXXVI of 1999), there is no neighboring right of photographers. In case of film, both Sweden and Hungary recognizes original works protected by copyright as well as neighboring rights (see: Article 1:1 Section 1 Point 4 and paragraph 5:49a Swedish Copyright Act as well as Article 1(2)(g) and Article 64 (3) of the Hungarian Copyright Act). The similarity in this case is due to harmonization, i.e. the Rental and Lending Rights Directive (Directive 2006/115 - Rental right and lending right and on certain rights related to copyright in the field of intellectual property).

[34] 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.

[35] Van den Burg, G. J. J., Williams, Ch. K. I. (2021). 'On Memorization in Probabilistic Deep Generative Models'. Available at: https://open-review.net/pdf?id=PlGSgjFK2oJ#:~:text=Memorization%20in%20generative%20models%20is,more%20likely%20to%20be%20generated..

[36] For evidence see The New York Times complaint filed on 27 December 2023, see footnote 12, particularly page 29 (para 98.).

Secondly, the content generated by a generative AI tool might infringe on the exclusive right of reproduction or the right of alteration, i.e. the right of creation of derivative works. In this second case, deciding on whether the output content is infringing on the original copyrighted work needs to be assessed on a case-by-case basis following the long established principles of copyright law. Due to the limits of this legal essay, the analysis will focus on the first case of possible infringement, namely *whether generative AI infringes copyright, more specifically economic rights by training the generative (foundation) models.*

This regard, it is important to stress that reproduction of *parts* of a copyrighted work might be sufficient for finding copyright infringement. It is established case law of the CJEU that even an excerpt of 11 words may be protected, to the extent that it forms the expression of its author's own intellectual creation.[37]

Based on the above, unless there is a copyright exception to the right of reproduction applicable in case of machine learning, particularly the type of machine learning involved in training generative foundation models, it is likely that generative AI infringes the economic rights of rightsholders by training the generative (foundation) models.

---

[37] Judgment of the Court (Fourth Chamber) of 16 July 2009, in case C-5/08 'Infopaq'.

3 Copyright exceptions

Although copyright entails a set of exclusive rights, these are not unlimited. For the sake of flexibility and long-term development as well as for balancing the right to property with other fundamental rights – including the right to education, right to health, freedom of expression – some limitations on these exclusive rights have been put in place. In the common-law countries, the fair use doctrine is prevalent. In contrast, in continental, *droit d'auteur regimes*, an exhaustive list of codified exceptions, and limitations,[38] is typical,[39] with the UK copyright system being somewhat in between.[40] National catalogues of limitations are partially harmonized and supplemented with the so-called three-step test, which has its roots in international copyright law.[41] A European three-step test is codified under Article 5(5) of the InfoSoc Directive effectively limiting the Member State's discretion in implementing the limitations. All in all, there are two limitations enshrined in EU copyright law instruments that can be relevant in a machine learning context.

3.1 Temporary acts of reproduction

Firstly, there is the mandatory exception for temporary acts of reproduction provided for in Article 5(1) InfoSoc.

Permitted acts of temporary reproduction are *transient* or *incidental* reproductions, forming an integral and essential part of a technological process and carried out for the sole purpose of enabling either efficient transmission in a network between third parties by an intermediary, or a lawful use of a work or other subject-matter to be made. The acts of reproduction concerned should have no separate economic value on their

---

[38] In this paper the word exception is used for both copyright exceptions and limitations, despite the dogmatic differences.

[39] Toth, A.K. (2019) 'Algorithmic Copyright Enforcement and AI: Issues and Potential Solutions, through the Lens of Text and Data Mining', Masaryk University Journal of Law and Technology, 13(2), pp. 361–388.

[40] Geiger Ch., Schönherr F., Stamatoudi, I., Torremans, P. 'The Information Society Directive', in *EU copyright law : a commentary*. Second edition (2021). Edward Elgar Publishing (Elgar commentaries), p. 314.

[41] Ibid., p. 316-317.

own. This exception was created primarily to enable browsing as well as acts of caching to take place, including those which enable transmission systems to function efficiently, provided that the intermediary does not modify the information and does not interfere with the lawful use of technology, widely recognised and used by industry, to obtain data on the use of the information.[42] This provision does not mean a certainly applicable exception for generative models, primarily because reproductions for creating "training data" is hardly *transient* or *incidental.*

Moreover, according to the Court of Justice in its decision Infopaq II case, the temporary copy must result from an automated process that deletes it automatically, without human intervention, once its function of enabling the completion of such a process has come to an end.[43] Therefore, the exception for temporary copies do not apply to copies resulting from a download, as these are considered permanent reproductions. By analogy, it is not likely to apply for reproductions in a machine learning context either.

Lastly, as part of the legislative efforts around the DSM Directive, EU commissioned several studies on whether a new limitation on copyright law for "text and data mining" was necessary to adopt at the time. The conclusion was that there was a great uncertainty to which extent Article 5(1) of the InfoSoc Directive covered data mining activities.[44] As described in section 2.2. above, machine learning results in more instances of reproduction than classical data mining activities. Indeed, the referenced report described text- and data mining (TDM) as a research technique that works by

---

[42] Recital 33 InfoSoc.

[43] Order of the Court (Third Chamber), 17 January 2012, Infopaq International A/S v Danske Dagblades Forening, Case C-302/10.

[44] Meeûs d'Argenteuil, J., Triaille, J., Francquen, A., (2014). 'Study on the legal framework of text and data mining (TDM)', European Commission, Directorate-General for the Internal Market and Services, Publications Office. Available at: https://op.europa.eu/en/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en.; Rosati, E. (2018). 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspects' Briefing to the JURI committee, available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/604942/IPOL_BRI(2018)604942_EN.pdf; Geiger, Ch., Frosio, G, Bulayenko, O. (2018). 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects' IN-DEPTH ANALYSIS For the JURI committee, available at: https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf.

14

"1. identifying input materials to be analysed, such as works, or data individually collected or organised in a pre-existing database;

2. copying substantial quantities of materials—which encompasses a. pre-processing materials by turning them into a machine-readable format compatible with the technology to be deployed for the TDM so that structured data can be extracted and b. possibly, but not necessarily, uploading the pre-processed materials on a platform, depending on the TDM technique to be deployed;

3. extracting the data; and

4. recombining it to identify patterns into the final output".[45]

In contrast, the reproduction machine learning entails is less transient or incidental. Moreover, it is also questionable whether those acts of reproduction, which become part of a foundation model, have any "separate economic value on their own". One could make the argument that it is the data the machine is trained on – with other words the reproduction – that create value. Therefore, it may be concluded that if data mining activities (as described above) are not entirely permitted under the exception of temporary acts of preproduction, then the relevant types of machine learning are even less likely to be legal.

3.2 Text and data mining (TDM)

The second exception is the text and data mining exceptions, in Article 3 and 4 of the DSM Directive,[46] that is the most referenced exception in context of generative AI and machine learning. The DSM Directive was adopted on 17 April 2019. Member States have of course implemented the rules in various times. Generally, the TDM exception will only be applicable from the entry

---

[45] Geiger, Ch., Frosio, G, Bulayenko, O. (2018). Page 4.
[46] 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

into force of the national rules implementing the directive.[47] This means that if reproduction took place prior to these dates, generative AI companies can only rely on the temporary acts of reproduction exception (section 3.1.).

3.2.1 Is training generative models a "TDM"?

As explained, *machine learning* and *data mining* are two different technological concepts. However, "text and data mining" is a legal term that is defined by Article 2 (2) DSM Directive as meaning "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". As many have pointed it out, this is a very broad definition that goes way beyond the technical concept of data mining described in section 2.2. above and does therefore likely encompass all machine learning activities conducted when developing generative foundation models or concrete generative AI tools.[48]

Much indicates, however, that this was not necessarily the intention of the law maker, and that the formulation "in order to generate information" was not intended to cover generating content. *Firstly*, Article 3 DSM Directive is only applicable for research organisations and cultural heritage institutions, when the purpose of the TDM is scientific research. The preparatory works emphasise the necessity to create the exception for research and data analysis as opposed to AI development, TDM is often referenced as a "basic means of research".[49] Indeed, the common proposal of the Parliament and the Council of 2016 only contained this exception. [50] The DSM Directive was reportedly targeted by unprecedented lobbying campaigns addressed to the European

---

[47] In Sweden for instance, the exception is applicable from 1 January 2023, in Hungary, from 1 June 2021. In Germany, a similar exception was adopted in 2018 for research purposes only. The amendment implementing the DSM directive-conform TDM exception entered into force 1 August 2021.

[48] Called "excessively broad" by Margoni, T., Kretschmer, M., A. 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology', *GRUR International*, Volume 71, Issue 8, August 2022, Pages 685–701.

[49] This wording is even adopted by the commentary, see: Stamatoudi, S., Torremans, P. (2021) 'The Digital Single Market Directive' in EU copyright law : a commentary. Second edition (2021). Edward Elgar Publishing (Elgar commentaries), p. 682.

[50] COM/2016/0593 final - 2016/0280 (COD).

Parliament,[51] und perhaps not entirely independently from this factor, the exception allowing for a for-profit, business use in Article 4 was introduced quite late in the legislative process. *Secondly*, it is true that innovation was one of the key factors accepting Article 4. Recital 18 states that, "in addition to their significance in the context of scientific research, text and data mining techniques are widely used both by private and public entities to analyse large amounts of data in different areas of life and for various purposes, including for government services, complex business decisions and the development of new applications or technologies." However, the legislators could not possibly have foreseen the upcoming breakthrough with accessible generative AI tools at the time. Nonetheless, arguing that Article 4 should be interpreted so that it does not encompass machine learning technologies would mean advocating for an interpretation contrary to the original meaning of the otherwise not very ambiguous law text. Therefore, the conclusion must be that "text- and data mining" within the meaning of Article 2(2) DSM Directive covers machine learning technologies.

3.2.2   Article 3 or Article 4?

It may prove to be hard to establish when a technique, which results in a reproduction, is deployed for research purposes. If a non-profit, research company develops a generative foundational model, that is then disclosed and becomes a basis of for-profit AI tools, can the non-profit rely on Article 3? This question is particularly relevant due to the inherent differences between Article 3 and Article 4 but might prove to be particularly difficult taking into account the long supply chain in the AI industry, as well as the blurred line between research facilities and for-profit companies.

Nonetheless, it is most likely that for the vast majority of tools, the research exception will not be applicable.

---

[51] Ferri, F. (2021) 'The dark side(s) of the EU Directive on copyright and related rights in the Digital Single Market.' *China-EU Law J* 7, 21–38.

As for Article 4, that is most likely relevant in our case, there are three cumulative criteria that must be met for the exception to be applicable, and accordingly, find no infringement, namely:

1) reproductions have not been retained for longer than what is necessary,

2) the original work has been lawfully accessed, and

3) the rightsholder has not reserved the right to use the work for text and data mining.

Below, we will analyse these three criteria separately.

3.2.3   Retention/storing

If reproductions are retained "for longer than what is necessary" for text and data mining purposes, creating unauthorised reproductions will result in an infringement. Since the purpose of TDM according to Article 2(2) is to "generate information", which is, again, a very broad formulation, this rule has very little restraining effect.

A relevant aspect in this regard is the previously mentioned "memorisation". Although this criterion appears to be easy to comply with, one can argue that the formulation must mean at least a theoretical possibility for deleting the retained reproductions so that once these are no longer necessary to be kept, they can be removed. However, trained model's memorisation means that deletion is only possible by deleting the model itself. Therefore, using generative models, especially those trained with deep learning (deep generative models) might be a conduct that cannot by its nature fulfil this requirement. It is worth mentioning that the most advanced text generating AI tools are based on such deep generative models.

3.2.4   Lawful access

This criterion is the least likely to be problematic, as CJEU has interpreted the expression "lawful use" in an intellectual property context in several of

its decisions as use authorised by the rightsholder or not restricted by the applicable legislation.[52]

In addition, Recital 14 states:

> "Lawful access should be understood as covering access to content based on an open access policy or through contractual arrangements between rightsholders and research organisations or cultural heritage institutions, such as subscriptions, or through other lawful means. For instance, in the case of subscriptions taken by research organisations or cultural heritage institutions, the persons attached thereto and covered by those subscriptions should be deemed to have lawful access. Lawful access should also cover access to content that is freely available online."

Although the above refers to research organisations or cultural heritage institutions only, the underlying rule of interpretation ought to be applicable within the context of Article 4.

3.2.5   Opt-out

Arguably, this is the most problematic criterion of Article 4. According to Article 4 (3), the TDM exception in Article 4 (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 rightsholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online. The term "machine readable" has been implemented as legal requirement in some countries (Germany, Hungary for example), whilst in other Member States, the term "appropriate manner" is featured in the amended Copyright Act and the requirement for "machine readability" remained a clarification only in the preparatory works (Sweden for instance).

---

[52] Judgment of the Court (Second Chamber) of 26 April 2017 in case C-527/15 'Filmspeler', para 65.; Judgment of the Court (Grand Chamber) of 4 October 2011 in joined cases C-403/08 and C-429/08 'Football Association Premier League' para. 168.; Order of the Court (Third Chamber), 17 January 2012 in case C-302/10 'Infopaq', para. 42.

This opt-out regime is completely absent from Article 3 and is the most crucial criterion of Article 4. So much so, that the latest working version of the AI Act Proposal[53] contains a direct reference to it, in a legislation that is not focusing on intellectual property issues otherwise. According to the draft, providers of foundation models would need to

1) "prepare and make publicly available a sufficiently detailed summary of the content used to train the model or system and information on the provider's internal policy for managing copyright-related aspects" and

2) demonstrate "that adequate measures have been taken to ensure the training of the model or system is carried out in compliance with Union law on copyright and related rights, in particular with regards to Article 4(3) of Directive (EU) 2019/790."[54]

Its importance also stems from the fact that, contrary to Article 3, the limitation in Article 4 does not have a strong public interest supporting it, such as supporting non-profit research, and as such it is a fairly strong intrusion to the exclusive rights of rightsholders. Yet, as Paul Keller highlights, it is unclear how opt-outs would work in practice.[55] There are examples where a copyright collective published a statement on its website opting out on behalf of all the represented rightsholders,[56] where editorial offices/newspapers declared that they opt out.[57] There are various technological measures emerging as well, where the uncertainty regarding

53 https://www.openfuture.eu/wp-content/uploads/2023/11/231107AIA_FM_compro-mises.pdf.

54 Summary from Keller, P. ' Generative AI and copyright: Convergence of opt-outs?'. Kluwer Copyright Blog. Available at: https://copy-rightblog.kluweriplaw.com/2023/11/23/generative-ai-and-copyright-convergence-of-opt-outs/.

55 Ibid, page 2/9.

56 TDM objection y ARTISJUS. Available at: https://www.artisjus.hu/felhasznalok-nak/mas-felhasznalas/ai-tdm-objection/.

57 Milmo, D. ,The Guardian blocks ChatGPT owner OpenAI from trawling its content'. The Guardian. Available at: https://www.theguardian.com/technology/2023/sep/01/the-guardian-blocks-chatgpt-owner-openai-from-trawling-its-content.

whether AI providers have taken notice of the opt-out becomes secondary.[58] However, at the moment it is rather unclear which strategies are compliant with the DSM Directive, in other words, which types of "opt out" would be accepted by the courts in an eventual infringement case. Paul Keller and Zuzanna Warso argue that an intervention of an actor with sufficient credibility is required to provide guidance on how to express machine-readable rights reservations. They state: "In the current constellation, the entity best placed to take on this role is the European Commission, which is responsible for ensuring the implementation of the provisions of the CDSM Directive."[59]

3.2.6   Fact vs. expression

Articles 3 and 4 were created to cope with the legal uncertainty as to whether reproductions and extractions made for the purposes of TDM can be carried out on lawfully accessed works. This is despite the common argument that there should be no need for a TDM exception "as the extraction of factual information from protected content is external to the remit of copyright".[60]

From a generative AI context, we can see the following. *Firstly*, generative (foundation) models are trained with complete works under copyright protection and due to the "memorisation" phenomenon, reproductions of these works are stored within the algorithms. In other words, not only the facts but the actual expression is affected. *Secondly*, although TDM as defined in Article 2(2) DSM Directive is so broad that it does cover even these fairly invasive type of uses, Articles 3 and particularly 4 does not provide the desired legal certainty. *Thirdly* and conclusively, the broad definition of TDM ap-

---

[58]Heikkilä, M. (2023) 'This new data poisoning tool lets artists fight back against generative AI'. *Technology Review*. Available at: https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/?fbclid=IwAR27wEL-wEnq05lPxkY0mmn82wuGQKQLmSxcIvcTbDmU_9bz5icau9f5A3nA.

[59] Keller, P., Warso, Z. (2023) 'Defining best practices for opting out of ML training". *Open Future.* Available at: https://openfuture.eu/wp-content/uploads/2023/09/Best-_practices_for_optout_ML_training.pdf.

[60] Rosati, E. (2018), Margoni, T., Kretschmer, M., A. (2022), and footnote 38 therein.

pears therefore unbalanced, and arguably jeopardises the fact/expression doctrine that is part of EU copyright law regime[61]: unless right holders are actively objecting to such use, AI providers may copy and store large volume of copyrighted work, without authorization or any compensation. This is particularly problematic considering that the trained models then are employed to create content, expression, that – even *if* it is not a direct infringement on the right holder's copyrighted works – is positioned to compete with those rightsholders and might therefore have a disincentivising effect on original creation.

---

[61] Indirect reference can be found in Recital 9 of the DSM Directive, see Margoni, T., Kretschmer, M., A. (2022), page 5.

4 Conclusion

All in all, whether generative AI models' training is infringing on the rightsholders' exclusive right of reproduction would depend on a range of different factors, the interpretation of most of which is somewhat unclear at the moment. Crucially, the interpretation of Article 4 DSM Directive by a European instance could be beneficial to create additional clarity and legal certainty.

The AI Act Proposal is a good example where disclosure requirements for AI providers can aid copyright enforcement efforts of rightsholders. Similar legislative efforts could be beneficial for a better balanced copyright law. This is because the current system seems to favour AI providers to a degree that might not have been intended by the law makers to begin with. Indeed, the legal definition of a text and data mining is extensively broad, and the criteria set for beneficiaries too unclear, to the extent that the current regime risks emptying the economic rights in our digital age. The idea behind a catalogue of precise exceptions and limitations is to promote legal certainty and predictability, however, the TDM exception does not seem to be fit for purpose.

A good indication of the above is that some very relevant stakeholders rather enter into licensing agreements in order to create certainty.[62] Not all rightsholders have, however, the weight to be able to negotiate an agreement, and taking into account the large content required for training a generative model, AI providers do not have the capacity to enter into agreements with each rightsholder individually either.

---

[62] Axel Springer and OpenAI partner to deepen beneficial use of AI in journalism. The parties have reportedly agreed on a licencing scheme: OpenAI would renumerate Axel-Springer and the publishing company will let OpenAI to train its generative models on their protected content. AxelSpringer statement available at: https://www.axelspringer.com/en/ax-press-release/axel-springer-and-openai-partner-to-deepen-beneficial-use-of-ai-in-journa-lism.

Bibliography

Case law, official documents

EU

Judgment of the Court (Fourth Chamber) of 16 July 2009, in case **C-5/08** 'Infopaq'.

Order of the Court (Third Chamber), 17 January 2012, in case **C-302/10** 'Infopaq II'.

Judgment of the Court (Second Chamber) of 26 April 2017 in case **C-527/15** 'Filmspeler',

Judgment of the Court (Grand Chamber) of 4 October 2011 in joined cases **C-403/08** and **C-429/08** 'Football Association Premier League' para. 168.;

Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(1),

Report on the rule of law - Adopted by the Venice Commission at its 86th plenary session (Venice, 25-26 March 2011), CDL-AD(2011)003rev-e -.

European Commission's Green Paper on 'Copyright and the Challenge of Technology'. COM (88) 172 Final, 7.6.1988.

Proposal for a Directive of The European Parliament and of The Council on copyright in the Digital Single Market, COM/2016/0593 final - 2016/0280 (COD).

Press release of the European Parliament 'Parliament leads the way on first set of EU rules for Artificial Intelligence'

TDM objection by copyright collective ARTISJUS.

USA

Competition and Markets Authority (2023). AI Foundation Models: Initial Report

Pending cases

USA

"*Getty images v. Stability AI,* challenging the legality of products such as Stable Diffusion and DreamStudio;

Class action *Doe v. Github, Inc.*, challenging the legality of GitHub Copilot (and a related product, OpenAI Codex, which powers Copilot), filed against GitHub, Microsoft and OpenAI;

Class action Authors Guild and Others as Plaintiffs v. OpenAI and Microsoft as Defendants, complaint filed 4 December 2023

The New York Times Company as Plaintiff v. Microsoft and OpenAi (various entities) as Defendants, complaint filed on 27 December 2023

UK

Case No: IL-2023-000007 between Getty Images (various companies) as Claimants and Stability AI Ltd. as Defendant. Preliminary ruling of 1 December 2023

Books, academic journals

Asp, P Rosén, J. 'The Pirate Bay – en kommentar'. (2011) SvJT.

Bettig, R.V. (1996) *Copyrighting culture : the political economy of intellectual property.* Westview Press (Critical studies in communication and in the cultural industries)

Ferri, F. (2021) 'The dark side(s) of the EU Directive on copyright and related rights in the Digital Single Market.' *China-EU Law J* 7, 21–38.

Geiger, Ch., Frosio, G, Bulayenko, O. (2018). 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects' IN-DEPTH ANALYSIS For the JURI committee.

Geiger Ch., Schönherr F., Stamatoudi, I., Torremans, P. 'The Information Society Directive', in *EU copyright law : a commentary*. Second edition (2021). Edward Elgar Publishing (Elgar commentaries).

Han, J., Kamber, M., Pei, J. (2012). 'Data mining: concepts and techniques.' (3rd ed), para. 1.2

Henderson, P. et al. (2023) 'Foundation Models and Fair Use'

Joyce, C. (ed.) (2013). Copyright Law. 9th ed. New Providence, LexisNexis

Keller, P. ' Generative AI and copyright: Convergence of opt-outs?'. Kluwer Copyright Blog.

Keller, P., Warso, Z. (2023) 'Defining best practices for opting out of ML training". *Open Future.*

Lim, D. 'Generative AI and copyright: principles, priorities and practicalities', *Journal of Intellectual Property Law & Practice*, Volume 18, Issue 12, December 2023.

Longan, M. (2021) 'Rethinking copyright and the Internet : a new model for users' rights'.

Lontai, E., Faludi G., Vékás, G. (ed.) (2012). *Szerzői jog és iparjogvédelem: [oktatási segédanyag]*. Budapest : Eötvös J. Kvk.

Lucas-Schloetter, A. 'Is there a concept of European copyright law? History, evolution, policies and politics and the *acquis communautaire'* in *EU copyright law : a commentary*. Second edition (2021). Edward Elgar Publishing (Elgar commentaries).

Mancini, A. (2006). *Copyright is Obsolete*. New York, Buenos Books.

Margoni, T., Kretschmer, M., A. (2022) 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology', *GRUR International*, Volume 71, Issue 8.

Marshall, E. 'Copyright Obsolete in An Electronic Age, OTA Finds'. (1986) *Science*, New Series, Vol. 232, No. 4750

Meeûs d'Argenteuil, J., Triaille, J., Francquen, A., (2014). 'Study on the legal framework of text and data mining (TDM)', European Commission, Directorate-General for the Internal Market and Services, Publications Office.

Mills, M. L. (1989) 'New Technology and the Limitations of Copyright Law: An Argument for Finding Alternatives to Copyright Legislation in an Era of Rapid Technological Change', *Chicago-Kent Law Review.* 65(1).

Pasick, A. (2023). 'Artificial Intelligence Glossary: Neural Networks and Other Terms Explained'. *The New York Times*.

Rosati, E. (2018). 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspects' Briefing to the JURI committee,

Samoili, S., Lopez Cobo, M., Gomez Gutierrez, E., De Prato, G., Martinez-Plumed, F. and Delipetrev, B., (2020). 'AI WATCH. Defining Artificial Intelligence', EUR 30117 EN, Publications Office of the European Union, Luxembourg.

Samoili, S., Lopez Cobo, M., Delipetrev, B., Martinez-Plumed, F., Gomez Gutierrez, E. and De Prato, G., (2021). 'AI Watch. Defining Artificial Intelligence 2.0', EUR 30873 EN, Publications Office of the European Union, Luxembourg.

Stamatoudi, S., Torremans, P. (2021) 'The Digital Single Market Directive' in EU copyright law : a commentary. Second edition (2021). Edward Elgar Publishing (Elgar commentaries), p. 682.

Toth, A.K. (2019) 'Algorithmic Copyright Enforcement and AI: Issues and Potential Solutions, through the Lens of Text and Data Mining', Masaryk University Journal of Law and Technology, 13(2), pp. 361–388.

Truitt, E.E. (2023). 'Surveillance, Companionship, and Entertainment: The Ancient History of Intelligent Machines', *THE MIT PRESS READER*.

Van den Burg, G. J. J., Williams, Ch. K. I. (2021). 'On Memorization in Probabilistic Deep Generative Models'.

News articles, magazines, other publications

Associated Press (2023) '2023 was the year AI went mainstream. It was also the year we started to panic about it', *euronews.next*.

Cuenco, M. (2023). 'We must declare jihad against AI', *Compact*.

Dib, D. (2023) 'How AI shook the world in 2023', *rest of the world*.

Heikkilä, M. (2023) 'This new data poisoning tool lets artists fight back against generative AI'. *Technology Review*.

Langkilde, D. (2023) 'Why I am excited about AI'

Marr, B. (2023). 'Is generative AI Stealing From Artists?', *Forbes*.

Mayor, A. (2018) 'An AI Wake-Up Call From Ancient Greece', *Project Syndicate*.

Metz, C. (2023) 'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead', *New York Times.*

Milmo, D. ,The Guardian blocks ChatGPT owner OpenAI from trawling its content'. *The Guardian..*

Torres, E. (2023). 'The Future of Creativity is Threatened by AI – Artist in Danger'. *Bootcamp*.

Wilkins, A. (2023) '2023 was the year that artificial intelligence went mainstream', *New Scientist.*

Yudowsky, E. (2022) 'AGI Ruin: A List of Lethalities', *AI Alignment Forum*.

Websites

Difference Between Data Mining and Machine Learning. (2023). Available at: https://differencebetween.io/data-mining-and-machine-learning/;

'Difference Between Data mining and Machine learning'. (2023). Available at: https://www.geeksforgeeks.org/difference-between-data-mining-and-machine-learning/.