Technical Challenges of Rightsholders' Opt-out From Gen AI Training after Robert Kneschke v. LAION [1]

by **Stepanka Havlikova** *

Abstract: This paper explores the evolving legal landscape surrounding generative AI model training on publicly available - often copyrighted - data, spotlighting the challenges in the wake of recent decision of German Court in Robert Kneschke v. LAION. On top of already explored implementation of copyright reservations by machine-to-machine and human-to-machine communication, this paper explores potential gaps and technical challenges stemming from the text and data mining exception including technical issues surrounding Robots.txt as well as data memorisation and regurgitation of verbatim snippets in AI outputs.

The Robert Kneschke v. LAION case exemplifies how non-profit organizations may leverage the TDM exceptions and offers insights that could influence commercial development of Gen AI. While the TDM exceptions may seem workable in theory, implementing them in practice presents a variety of practical challenges. Practical implications, such as requirements for "machine-readable" opt-out options for rightsholders considering current technological landscape, may ultimately reduce the practical benefits of these exceptions. Dataset creation and AI model training in practices occurs via chain of parties from copyright holders, licensors or publishers, non-profit organisations populating datasets to commercial AI developers which may bring additional interpretational issues and gaps when applying exception for research purposes or searching for validly applied opt-out. This paper discusses legal requirements and interpretation introduced by Robert Kneschke v. LAION and presents practical and technical implications stemming from the TDM exceptions and suggests possible outcomes thereof.

Keywords: Artificial Intelligence, Web Scraping, Text and Data Mining, Machine-readable, Copyright

Recommended citation: Stepanka Havlikova, Technical Challenges of Rightsholders' Opt-out From Gen AI Training after Robert Kneschke v. LAION, 16 (2025) JIPITEC XX para 1

1 LG Hamburg, Urteil vom 27. September 2024 – 310 O 227/23 (Robert Kneschke v. LAION).

* PhD Candidate at the Institute of Law and Technology at Masaryk University and a Senior Associate at Dentons Law Firm. I thank Pavel Koukal, Jacopo Ciani Sciolla, Massimo Durante, Alessandro Cogo and Péter Mezei, for their feedback and helpful suggestions either on various drafts of this paper or ideas presented therein. This article is the result of the project of the Grant Agency of the Czech Republic [Copyrighted Works and the Requirement of Sufficient Precision and Objectivity (GA22-22517S)].

A. Introduction

1 During the preceding months we can see a significant rise of lawsuits in the United States based on copyright infringement[3] in connection with generative artificial intelligence[4] and scraping of large amounts of publicly available information to train artificial intelligence.[5] As the Economist recently pointed out in its article addressing copyright and artificial intelligence, "*it is the oceans of copyrighted data the bots have siphoned up while being trained to create humanlike content*" while "*often, it is alleged, AI models plunder the databases without permissions*".[6] Lemley and Casey noted that this may well be one of the most important legal questions of the coming century: *Will copyright law allow robots to learn?*[7] It may be only question of time whether and when similar cases are initiated in the EU, especially in connection with the Representative Action Directive[8] currently being implemented across the EU[9] while at the same time heavily supporting AI development and launch across the EU.[10] Considering the broad interpretation of the concept of reproduction[11] (for copyright) and extraction[12] (for database rights) under EU law, scraping publicly available copyright (or database) protected content may indeed constitute copyright or database right infringements,[13] unless rightsholders grant their authorisation or statutory exception applies.[14]

2 When considering potential development of similar cases under EU law, recently adopted set of two

3 For example the Author's Guild claims that OpenAI's and Microsoft's AI models were "*trained," .. by reproducing a massive corpus of copyrighted material, including, upon information and belief, tens or hundreds of thousands of fiction and nonfiction books*" and that "*the only way that Defendants' models could be trained to generate text output that resembles human expression is to copy and analyze a large, diverse corpus of text written by humans*". With this argumentation the plaintiffs are requesting the defendants namely to cease using the infringing content and to provide financial compensation for past infringements. Brown, T.T., et al., Language Models are Few-Shot Learners. Available at: https://arxiv.org/pdf/2005.14165 [Accessed on 31.12.2024].

4 Hereinafter also abbreviated to Gen AI.

5 Cases filed before U.S. District Courts in 2023 against various global AI tools suppliers: Getty Images, Inc. v. Stability AI Ltd, U.S. District Court for the District of Delaware; Sarah Andersen v. Stability AI Ltd, U.S. District Court for the Northern District of California; Authors Guild v. Open AI, U.S. District Court for the Southern District of New York; Chabon v. OpenAI Inc., U.S. District Court for the Northern District of California; Richard Kadrey v. Meta Platforms, Inc., in the U.S. District Court for the Northern District of California; Sarah Silverman v. OpenAI, Inc., U.S. District Court for the Northern District of California.

6 'A battle royal is brewing over copyright and AI', The Economist [online], 2023. Available at: https://www.economist.com/business/2023/03/15/a-battle-royal-is-brewing-over-copyright-and-ai [Accessed on 31.12.2024].

7 Lemley, M.A. and Casey, B., 2020. Fair Learning. Available at SSRN: https://ssrn.com/abstract=3528447 or http://dx.doi.org/10.2139/ssrn.3528447 [Accessed on 31.12.2024].

8 Directive (EU) 2020/1828 of the European Parliament and of the Council of 25 November 2020 on representative actions for the protection of the collective interests of consumers and repealing Directive 2009/22/EC (the "Representative Actions Directive").

9 In accordance with deadline for implementation by 25 June 2023.

10 EU's long-term digital strategies identify the uptake of artificial intelligence as one of the objectives of the Digital Decade Policy Programme 2030. Artificial intelligence was named as one of the technologies (along with cloud computing and big data) which at least 75 % of Union enterprises should take up by 2030 (as part of the digital transformation of businesses which forms one of the digital targets in the Union); See Art. 4 (1) (3) Decision (EU) 2022/2481 of the European Parliament and of the Council of 14 December 2022 establishing the Digital Decade Policy Programme 2030, Available at: https://eur-lex.europa.eu/eli/dec/2022/2481/oj [Accessed on 31.12.2024]. The 2021 Coordinated Plan on Artificial Intelligence explicitly highlighted that "*availability of high-quality data, among other things, in respect of diversity, nondiscrimination, and the possibility to use, combine and re-use data from various sources in a GDPR compliant way are essential prerequisites and a precondition for the development and deployment of certain AI systems*". See the 2021 Coordinated Plan on Artificial Intelligence; Available at: https://digital-strategy.ec.europa.eu/en/policies/plan-ai

11 *Infopaq International A/S v. Danske Dagblades Forening*, Judgment of the Court of Justice dated 16.07.2009 in case C-5/08.

12 *Innoweb BV v. Wegener ICT Media BV*, Wegener Mediaventions BV. Judgment of the Court of Justice dated 19.12.2013 in case C-202/12. *CV-Online Latvia SIA v Melons SIA*. Judgment of the Court of Justice dated 3.6.2021 in case C-762/19.

13 Canellopoulou-Bottis, M., Papadopoulos, M., Zampakolas, C., and Ganatsiou, P., 2019. 'Text and Data Mining in Directive 2019/790/EU Enhancing Web-Harvesting and Web-Archiving in Libraries and Archives', Open Journal of Philosophy, p. 378.

14 R. Ducato and A. Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility"', CRIDES Working Paper Series (2018) 10.13140/RG.2.2.15392.84482. Available at SSRN: https://ssrn.com/abstract=3278901 or http://dx.doi.org/10.2139/ssrn.3278901 [Accessed on 31.12.2024]. Okediji, R., 2017. Copyright Law in an Age of Limitations and Exceptions. Cambridge: Cambridge University Press. ISBN 978131645090.

exceptions from copyright and database protection[15] for purposes of so-called "*text and data mining*"[16] introduced by the CDSM Directive[17] could emerge as pivotal when aiming to justify use of publicly available data to train artificial intelligence.[18] Existing case law addressing web scraping from various perspectives could also play significant role highlighting that scraping may lead to additional legal consequences such as unfair competition or free riding.[19]

3 Both TDM Exceptions are associated with legal uncertainties whereas some questions have been addressed by the recent decision of the German court in *Robert Kneschke v. LAION.*[20] In Robert Kneschke v. LAION German Hamburg Regional Court recently ruled on a lawsuit filed by German Photographer Robert Kneschke against the nonprofit organisation LAION which created a dataset consisting of image-text pairs subsequently used to train AI which included Kneschke's photos. The case against LAION was dismissed on the grounds of the scientific research TDM exception. Surprisingly, despite the fact the case was in fact dismissed based on TDM exception under Art. 3 CDSM Directive, significant part of the *obiter dictum* was dedicated to the court's view on TDM exception under Art. 4 CDSM Directive.

B. Applying TDM Exception on Gen AI Training

4 TDM Exceptions introduced by the CDSM Directive allow reproductions and extractions of protected content to carry out text and data mining defined as an "*automated analytical technique aimed at analysing text and data in digital form in order to generate information*".[21] Although scholars tend to agree TDM exceptions may serve as a suitable legal basis to justify use of data for generative AI training,[22] there are debates[23] to which extent did the development of artificial intelligence form a ratio behind enacting the TDM exceptions.[24]

15 And press publisher rights.

16 Text and data mining (further referred to as "TDM" and Text and Data Mining Exception under Art. 4 of the CDSM Directive also referred to as "TDM Exception").

17 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 (hereinafter referred to as the "CDSM Directive").

18 CDSM Directive introduces two exceptions or limitations allowing (i) text and data mining for the purpose of scientific research under Art. 3 CDSM Directive and (ii) text and data mining for other purposes unless reserved by rightsholders under Art. 4 of the CDSM Directive. Art. 3 of the CDSM Directive introduces an exception from reproduction rights under copyright protections, extraction rights under sui generis database protections and press publisher rights for reproductions and extractions of lawfully accessible works and other subject matters for the purposes of text and data mining for research purposes. Art. 4 of the CDSM Directive introduces an exception from reproduction rights under copyright protections, extraction rights under sui generis database protections and press publisher rights for reproductions and extractions of lawfully accessible works and other subject matters for the purposes of text and data mining, if such rights have 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.

19 See for example Pagallo U., Ciani Sciolla J., Anatomy of web data scraping: ethics, standards, and the troubles of the law. European Journal of Privacy Law & Technologies, (2023) 2 p. 1 - 19, available at: https://doi.org/10.57230/EJPLT232PS. [Accessed on 31.12.2024]. Due its limited extent, these consequences are excluded from the scope of this paper.

20 LG Hamburg, Urteil vom 27. September 2024 – 310 O 227/23 (Robert Kneschke v. LAION).

21 See footnote 18

22 Mezei, Péter, A saviour or a dead end? Reservation of rights in the age of generative AI (January 15, 2024). European Intellectual Property Review, 2024, 46(7), p. 461-469. Available at SSRN: https://ssrn.com/abstract=4695119 or http://dx.doi.org/10.2139/ssrn.469511. [Accessed on 31.12.2024]. Novelli, Claudio and Casolari, Federico and Hacker, Philipp and Spedicato, Giorgio and Floridi, Luciano, Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity (January 14, 2024). Available at SSRN: https://ssrn.com/abstract=4694565 or http://dx.doi.org/10.2139/ssrn.4694565. [Accessed on 31.12.2024].Rosati, E., Copyright in the Digital Single Market: Article-by-Article Commentary to the Provisions of Directive 2019/790, Oxford University Press, Oxford 2021. ISBN: 9780198858591. P. 72. Dusollier, Séverine, 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition' (2020) 57 Common Market Law Review 979, 984. Hanjo Hamann, Artificial Intelligence and the Law of Machine-Readability: A Review of Human-to-Machine Communication Protocols and their (In)Compatibility with Art. 4(3) of the Copyright DSM Directive, 15 (2024) JIPITEC 102 para 1.

23 EU accused of leaving 'devastating' copyright loophole in AI Act', The Guardian [online], 2025. Available at: https://www.theguardian.com/technology/2025/feb/19/eu-accused-of-leaving-devastating-copyright-loophole-in-ai-act [Accessed on 20 March 2025].

24 TDM exception introduced under Art. 4 CDSM Directive was not part of the Commission Proposal of the CDSM Directive which aimed to introduce solely exception for text and data mining for purposes of scientific research with no text and data mining exception for other purposes. TDM Exception - currently under Art. 4 – was subsequently proposed during the legislative procedure by the Committee on

5 Interestingly, the Commission Proposal of the CDSM Directive aimed to introduce solely the TDM Exception for purposes of scientific research.[25] Non-research TDM exception[26] was subsequently proposed during the legislative procedure by the Committee on Legal Affairs (JURI) and supported by the Parliament and the Council. For example, with the argumentation that "*this type of permitted use was not conceived for artificial intelligence*" the initial Polish legislative proposal for implementing the CDSM Directive included a controversial provision explicitly excluding the creation of generative AI models from the scope of the exceptions – which however did not stand and the final adopted law departed from this proposal and instead closely aligned with the original text of the CDSM Directive.[27] Although sometimes used as an argument against the applicability of the TDM Exception on AI training, such an interpretation was rejected by many scholars[28] as well as German court in *Robert Kneschke v. LAION*.[29] Lastly, the AI Act explicitly references the TDM exception in the context of training general-purpose AI models, underscoring that the exception might indeed be applicable when using protected content for AI training.[30]

6 The TDM Exception under Art. 3 CDSM Directive is limited to research organisations and cultural heritage institutions to carry out text and data mining for the purposes of scientific research and thus cannot be relied on by commercial companies scraping data to develop Gen AI (the interplay between Art. 3 and Art. 4 CDSM Directive will be further debated below). On the contrary, TDM exception under Art. 4 CDSM Directive is not limited by research purposes by research organisations – however applies only insofar such rights have not been "*expressly reserved by their rightsholders in an appropriate manner, such as machine-readable means*

Legal Affairs (JURI) and supported by the Parliament and the Council. See 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). Rosati, E., Copyright in the Digital Single Market: Article-by-Article Commentary to the Provisions of Directive 2019/790, Oxford University Press, Oxford 2021. ISBN: 9780198858591. P. 65. Mezei, Péter, A saviour or a dead end? Reservation of rights in the age of generative AI (January 15, 2024). European Intellectual Property Review, 2024, 46(7), p. 461-469. Available at SSRN: https://ssrn.com/abstract=4695119 or http://dx.doi.org/10.2139/ssrn.469511. [Accessed on 31.12.2024]. Report on the Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market (COM (2016)0593 – C8-0383/2016 – 2016/0280(COD)) (Rapporteur: MEP Axel Voss), Amendment 65. Dusollier, S., 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition' (2020) 57 Common Market Law Review 979, 984. Jan Bernd Nordemann and Jonathan Pukas, 'Copyright Exceptions for AI Training Data – Will There Be an International Level Playing Field?' (2022) 17 Journal of Intellectual Property Law & Practice 973, 974. Hajo Hamann, 'Artificial Intelligence and the Law of Machine-Readability: A Review of Human-to-Machine Communication Protocols and Their (In)compatibility with Art. 4(3) of the Copyright DSM Directive' (2024) 15(2) JIPITEC 102, 105–106.

25 Currently Art. 3 CDSM Directive.

26 Currently Art. 4 CDSM Directive.

27 Draft implementation law published by polish Government for consultation. Available at: https://legislacja.rcl.gov.pl/projekt/12382002. [Accessed on 31.12.2024].

28 Mezei, Péter, A saviour or a dead end? Reservation of rights in the age of generative AI (January 15, 2024). European Intellectual Property Review, 2024, 46(7), p. 461-469. Available at SSRN: https://ssrn.com/abstract=4695119 or http://dx.doi.org/10.2139/ssrn.469511. [Accessed on 31.12.2024]. Novelli, Claudio and Casolari, Federico and Hacker, Philipp and Spedicato, Giorgio and Floridi, Luciano, Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity (January 14, 2024). Available at SSRN: https://ssrn.com/abstract=4694565 or http://dx.doi.org/10.2139/ssrn.4694565. [Accessed on 31.12.2024]. Rosati, E., Copyright in the Digital Single Market: Article-by-Article Commentary to the Provisions of Directive 2019/790, Oxford University Press, Oxford 2021. ISBN: 9780198858591. P. 72. Dusollier, Séverine, 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition' (2020) 57 Common Market Law Review 979, 984. Hanjo Hamann, Artificial Intelligence and the Law of Machine-Readability: A Review of Human-to-Machine Communication Protocols and their (In)Compatibility with Art. 4(3) of the Copyright DSM Directive, 15 (2024) JIPITEC 102 para 1.

29 LG Hamburg, Urteil vom 27. September 2024 – 310 O 227/23 (Robert Kneschke v. LAION).

30 Recital 105 of the AI Act confirms that the use of literary and artistic works for AI training purposes has copyright relevance and involves acts of text and data mining that require the authorisation of rightholders: "[a]*ny use of copyright protected content requires the authorisation of the rightholder concerned unless relevant copyright exceptions and limitations apply*" and subsequently refers to the TDM exception and notes that "*Where the rights to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorisation from rightsholders if they want to carry out text and data mining over such works*". Also Mezei, Péter, The Multi-layered Regulation of Rights Reservation (Opt-out) Under EU Copyright Law and the AI Act -For the Benefit of Whom? (v1.0) (December 19, 2024). Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5064018 [Accessed on 30.12.2024].

in the case of content made publicly available online".[31]

C. Practical Challenges Associated with Machine-Readable Opt-Out

7 The TDM exception under Art. 4 CDSM Directive faced criticism for its impracticality, particularly due to the rightsholders' opt-out mechanism. As Hugenholtz aptly observed, *the TDM provisions of the CDSM Directive secure considerably less freedom to text and data mine than they initially appear to do. The opt-out clause of Art. 4, in particular, leaves for-profit miners in the EU at the mercy of the content owners."*[32] However, the lack of standardization, ambiguity in how to properly implement the reservation, and technical challenges in decoding these measures introduce further complications including the question who sets the standards and what the level of "*machine-readability*" is expected from reservations. A critical question remains: who will bear the burden: rightsholders, AI companies, or end users?

I. Is "Machine-Readability" a Strict Requirement to Validly Opt-Out?

8 First question arises in connection with interpretation of the "*machine-readable*" requirement which is cited in connection with content made publicly available online. It is worth noting that some scholars are of the view that the machine-readability is not a strict requirement on how the reservation must be made but rather an example of how the reservation could be made – meaning that even non-machine-readable reservation could have legal effect if expressed by appropriate means.[33] This extensive interpretation could lead to the conclusion that *any* reservation expressed by rightsholders is valid if "*appropriate*". However, the absence of "*machine-readable*" form could undermine the sole purpose of the TDM exception of *allowing the automated computational analysis of information*[34] and text and data mining as an "*automated analytical technique aimed at analysing text and data in digital form*".[35] Some countries have not expressly implemented the machine-readability requirement in their national legislation and implemented solely "*appropriate means*" requirement – such as in Italy.[36] On the other hand, countries such as Germany, Austria, Slovakia or the Czech Republic make it clear that machine-readability forms a requirement making the opt-out ineffective if these conditions are not met.[37]

9 In the author's view, machine-readability should in fact be considered as a mandatory legal requirement to form a legally effective reservation from the TDM Exception.[38] This follows also from recitals of the CDSM Directive which states that "*In the case of content that has been made publicly available online, it should only be considered appropriate to reserve those rights by the use of machine-readable means, [...]*" (emphasis added).[39] As a result, even the absence of explicit machine-readability requirement can be overcome by interpretation of the "*appropriate means*" requirement in light with the CDSM Directive.[40]

II. Interpretation of *"Expressly"* Reserved in *"Machine-Readable"* Form

10 The question remains how such "*machine-readable*" means shall be interpreted as CDSM Directive does not provide any legal definition thereof. According to Recital 18 of the CDSM Directive, such machine-readable means may include "*metadata and terms and conditions of a website or a service*".[41] Accordingly, machine-readable means could include for example technical restrictions and disallow commands[42] but

31 Defined as „*automated analytical technique aimed at analysing text and data in digital form in order to generate information*"

32 Hugenholtz, B. The New Copyright Directive: Text and Data Mining (Articles 3 and 4) [online]. Kluwer Copyright Blog. 2019. Available at: http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ [Accessed on 31.12.2024].

33 Discussion held during International Conference Techno-legal challenges of data Scraping hosted at the the University of Turin, Department of Law in November 2023.

34 Recitals 8 – 11 of the CDSM Directive.

35 Art. 2 (2) CDSM Directive.

36 Such as Italy. Section 70 of the Italian Copyright Act.

37 Löbling, L., Handschigl, Ch. Hofman, K., Schwedhelm, J. Navigating the Legal Landscape: Technical Implementation of Copyright Reservations for Text and Data Mining in the Era of AI Language Models. 14 (2023) JIPITEC 499 para 14.

38 The arguments for such interpretation are as follows. The beginning of the sentence starting with „such as" relates rather to the designation of „content made publicly available online" which requires as „appropriate means" the „machine-readable means". There may be other types of content not made publicly available online where the „appropriate means" are not specified by the CDSM Directive.

39 Recital 18 CDSM Directive.

40 *Costa v. ENEL*, Judgment of the Court of Justice in case 6/64.

41 As the Recital 18 of the CDSM Directive states: For that has been made publicly available online, it should only be considered appropriate to reserve those rights by the use of machine-readable means, including metadata and the terms and conditions of a website or a service.

42 Strowel, A., Ducato, R. Artificial Intelligence and Text

also reservations made via a website's terms of use provided they are in a machine-readable format.

11 By analogy, the Open-Data Directive defines machine-readable format of documents as "*a file format structured so that software applications can easily identify, recognise and extract specific data, including individual statements of fact, and their internal structure*". Nevertheless, the ratio behind Open-Data Directive significantly differs from the ration of Art. 4 CDSM Directive and thus it may not be suitable as *analogia legis*. As follows from Recital 35 of the Open Data Directive, "*A document should be considered to be in a machine-readable format if it is in a file format that is structured in such a way that software applications can easily identify, recognise and extract specific data from it. Data encoded in files that are structured in a machine-readable format should be considered to be machine-readable data."* While the Open-Data Directive aims to ensure access and reuse of public-sector information, the CDSM Directive aims to strike a balance between the interests of users of text and data mining (to be able to conduct automated analysis of data) and the interests of rights holders (to protect their rights). As a result, the requirement on machine-readability set forth by the Open-Data Directive is set as low as possible to ensure the easiest possible access of the public to the relevant information. However, setting the same benchmark for "*machine-readability*" under the CDSM Directive would mean shifting the balance significantly to the benefit of the users utilizing text and data mining. As a result, the definition of "*machine-readability*" under the Open-Data Directive cannot be relied on when interpreting the CDSM Directive.

12 Aim of the CDSM Directive is to allow the text and data mining which is defined as *"automated analytical technique [...]"* with the intention of making possible "*the processing of large amounts of information with a view to gaining new knowledge and discovering new trends*" and to "*analyse large amounts of data*".[43] German explanatory memorandum to Act amending the German Copyright Act (implementing the CDSM Directive) provides some guidance by emphasising that machine-readable reservation must enable automated processes because *"[...] the purpose of the regulation is to ensure that automated processes, which are typical criteria of text and data mining, can actually be automated in the case of content accessible online*".[44] Interestingly, the German Explanatory Memorandum mentions that the reservation can be included in the imprint of a given website (*Impressum*) or in its terms and conditions, *provided that it is machine-readable*.[45] On the contrary, Czech Explanatory Memorandum explained that the reservation may be easily implemented through standard metadata (e.g. by structuring the metadata to a format which automated tools are able to read) but noted that general statements on websites on in content terms of use are not a suitable mean to express the reservation.[46]

13 German court in *Robert Kneschke v. LAION* noted that while the term "*machine readability*" must be interpreted in light of the legislative intent underlying it — to enable automated queries by web crawlers — it should be understood in the sense of "*machine understandability*" whereas such question should always be answered based on the technical developments prevailing at the relevant time of use of the work. With reference to state-of-the-art technologies requirement stemming from the AI Act - which applies on providers of general-purpose AI models if intended to utilize TDM Exception - the court noted that "*these "state-of-the-art technologies" undoubtedly include, in particular, AI applications capable of comprehending text written in natural language".* The court further explained that CDSM Directive does not demand that a reservation needs to be declared "*in the simplest way possible*," but rather "*in an appropriate manner*" which suggests certain middle ground between the requirement of "*machine-readability*" enabling automated processes while at the same time granting the rightsholders the freedom to choose means available to them.[47]

and Data Mining: A Copyright Carol IN Rosati, E. The Routledge Handbook of EU Copyright Law. Ed. Eleonora Rosati. Abingdon. 2021. ISBN: 9780367436964. P. 30. Hugenholtz, B. The New Copyright Directive: Text and Data Mining (Articles 3 and 4) [online]. Kluwer Copyright Blog. 2019. Available at: http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ [Accessed on 31.12.2024].

43 Recital 8 and 18 CDSM Directive.

44 Explanatory memorandum (Gesetzesbegründung) of the German Government (Bundesregierung) to its legislative proposal implementing the CDSM Directive: Entwurf eines Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes, Gesetzesbegründung: Besonderer Teil. No. 19/27426. Page 95. Available at https://dip.bundestag.de/vorgang/.../273942 [Accessed on 31.12.2024].

45 Ibid.

46 Explanatory memorandum (Důvodová zpráva) of the Czech Government to the Act. No. 429/2022 Coll. (amending the Czech Copyright Act implementing the CDSM Directive). Section § 39c.

47 However, it is very important to highlight that – as already mentioned above – the question of "machine-readability" was only tackled by the court in obiter dictum of the judgement whereas although the court shared its legal opinion on the question at hand, it also explicitly noted that whether the defendant can rely on the TDM exception under Art. 4 CDSM Directive "*does not need to be conclusively determined*" which slightly undermines the precedential weight of the argumentation.

14 The court applied a rather *pro-rightsholder* interpretation as it set the benchmark of "*machine-readability*" relatively low which however imposes very high demands on the users relying on the TDM Exception when decoding such reservations. The court has however not tackled the issue of potential unreliability of Gen AI which may prevent such users from consistently and reliably identifying reservations in all cases.[48] As a result, while such reservations may *in most cases* be indeed decoded by generative AI capable of understanding natural language, the accuracy of decoding is unlikely to be flawless (for example reliability will likely vary depending on the specific generative AI model[49] or language of the reservation[50]). This uncertainty exposes generative AI developers to legal risks of potential copyright infringements despite applying their best efforts and state-of-the-art technologies. On the other hand, the failure to adequately present a reservation in a machine-readable form with sufficient reliability should not disadvantage users relying on the TDM exceptions who might not be able to reliably decode such reservation despite applying state-of-the-art technologies but should rather go to the detriment of the rightsholders who have the power and control as to how they implement and express their reservations.

15 Although he rightsholders to set the tone of the "*appropriate means*" as they decide how to implement their reservations, the recently adopted AI Act[51] obliges the providers of so-called general-purpose AI models[52] to put in place a policy to comply with Union copyright law, and in particular to "*identify*" ... "*through state of the art technologies*". reservations of rights.[53] In *Robert Kneschke v. LAION,* German court used a reference to the AI Act while assessing whether publicly available declarations in human language may constitute a machine-readable exception.[54] In the author's view, the interplay with Art. 53 of the AI Act could offer a valuable solution for addressing challenges under Art. 4 of the CDSM Directive. While Art. 53 of the AI Act applies specifically to providers placing general-purpose AI models on the EU market and may not cover all providers of generative AI

48 Not to mention that this interpretation creates a „chicken-and-egg" dilemma, as generative AI capable of understanding natural language cannot be developed without access to sufficiently broad high-quality datasets.

49 Iorliam, Aamo & Ingio, Joseph. (2024). A Comparative Analysis of Generative Artificial Intelligence Tools for Natural Language Processing. Journal of Computing Theories and Applications. Volume 2. 10.62411/jcta.9447.

50 Reliability of Gen AI decoding the reservation may for example largely depend on language of the reservation as some Gen AI models have higher reliability in English language but lower reliability in other languages.

51 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 or also AI Act).

52 Defined in Art. 3 AI Act as "*an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market*".

53 Providers of general-purpose AI models shall inter alia (i) draw up technical documentation (including also information on the data used for training, testing and validation and how the data was obtained and selected); (ii) put in place a policy to comply with Union copyright law, and in particular to identify and comply with, including through state of the art technologies, a reservation of rights expressed pursuant to Art. 4(3) CDSM Directive; and (iii) draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model as follows from Art. 53 AI Act. These requirements shall apply within 12 Months after the AI Act comes into force. Finally, respecting opt-outs from the TDM exception is an explicit part of the GPAI model providers' obligation to comply with EU copyright law as follow from Art. 53(1)(c) of the AI Act. As a result, GPAI models trained with material in violation of valid opt-outs are not compliant with the AI Act and may not be put into service or placed on the market in the EU. Recital 106 of the AI Act further justifies the requirement by competition grounds while explaining the necessity to ensure a level playing field among providers of general-purpose AI models where no provider should be able to gain a competitive advantage by applying lower copyright standards. Therefore, we can expect that within the upcoming 12 months, remaining developers of generative artificial intelligence shall follow the trend set by OpenAI and shall introduce their recommendations on implementation of the reservation from the TDM exception which shall make it easier for rightsholders to effectively implement their reservations. Concurrently, new obligations of publishing a sufficiently detailed summary about the content used for training shall make it easier for rightsholders to establish unlawful use of their content in case the reservation has not been duly complied with.

54 Specifically, the court assessed the question of whether and under what specific conditions a reservation of use expressed in "natural language" can also be considered "machine-understandable" and noted it must always be answered based on the technical developments prevailing at the relevant time of use of the work. Subsequently the court referred to "state-of-the-art technologies" under the AI Act and concluded that these "state-of-the-art technologies" undoubtedly include, in particular, AI applications capable of comprehending text written in natural language".

models utilizing copyright-protected content within the EU, the state-of-the-art technologies employed by these providers could set a precedent eventually influencing how courts interpret and apply Art. 4 CDSM Directive.

16 As explained above, the CDSM Directive aims to strike a balance between the interests of users of text and data mining (to be able to conduct automated analysis of data) and the interests of rights holders (to protect their rights). As a result, while users of text and data mining should indeed be expected to employ state-of-the-art technologies to decode reservations, rightsholders' "*express*" reservations in "*machine-readable*" formats should, in the author's view, achieve a *reliable level* of machine interpretability. This might require the reservation to be presented in a sufficiently binary form that enables such advanced technologies to *reliably* decode its content leaving no room for doubt. This may be reflected for example by a standardized formulas (despite being written in a natural human language) which could be for example similar to standardized open-source license terms. On the contrary, the author believes that vague terms and conditions generally prohibiting scraping or bot access without expressly invoking reservation of rights from the TDM Exception (mainly those applied prior to TDM Exceptions coming into effect) should in most cases in fact not be able to achieve the level of "*express*" reservation in "*machine-readable*" form fulfilling the required level of reliability of its decoding. For instance, in the case assessed by the German court, the plaintiff's reservation used a rather generic wording prohibiting *"use automated programs .. for purposes of ... scraping"* but did not expressly refer to text and data mining.[55] Moreover, the court noted that these terms were published on the websites as early as 13 January 2021, before the CDSM Directive was implemented in Germany on 20 May 2021. This timing suggests that the reservation may not have been intended to address the TDM exception. Such an interpretation, however, might conflict with the requirement for "*expressly*" reserving rights in "*machine-readable*" means, which in the author's view implies that a reservation should unequivocally be understood as the rightsholder's intention to prevent text and data mining, leaving no room for doubt.[56]

17 However, as of today, no such sufficiently unified language of such reservation exists despite some attempts to introduce unified formulas.[57] Such unification could be for example established by independent bodies having sufficient authority to influence the global market.[58]

III. Existing State-of-the-Art Technologies Enabling to Express Rightsholder's Opt-Out

18 Open Future research aptly differentiates between "*unit-based*" and "*location-based*" identifiers enabling to place the express rights reservations either at a high level, affecting all applicable content available for example under a given website, or reservations affecting each content item individually.[59] Among

55 Specifically, the court referred to the following wording on the defendant's website: *"RESTRICTIONS: YOU MAY NOT: (...) 18. Use automated programs, applets, bots or the like to access the XXX.com website or any content thereon for any purpose, including, by way of example only, downloading content, indexing, scraping, or caching any content on the website."*

56 For example, when requesting ChatGPT (version 4o) using various prompts to provide an answer whether the wording applied in the case at hand presents a valid reservation within the meaning of Art. 4 CDSM Directive, ChatGPT was not able to provide clear answer – out of three prompts, in one case ChatGPT responded positively, in one case provided vague answer and in one case responded negatively. Although this itself does not exclude the machine-readability, it somewhat underlines the possibility of Gen AI providing different conclusions. For example:
Prompt: "*Website published terms and conditions containing wording below. Has the owner of the website expressly reserved its rights by machine-readable means under article 4 para 3 of the EU CDSM Directive? "RESTRICTIONS: YOU MAY NOT: (...) 18. Use automated programs, applets, bots or the like to access the XXX.com website or any content thereon for any purpose, including, by way of example only, downloading content, indexing, scraping, or caching any content on the website."*
Answer: "*The wording you provided restricts the use of automated tools to access the website but does not seem to expressly reserve rights through machine-readable means, as required under Article 4(3) of the EU Directive 2019/790 on Copyright in the Digital Single Market (CDSM Directive). ... Based solely on the provided text, the website owner has not expressly reserved their rights under Article 4(3) by machine-readable means. To comply with the Directive, the owner would need to implement additional technical measures beyond this contractual language."*

57 Keller/Warso, 'Defning Best Practices for Opting Out of ML Training' (29 Sep 2023), OpenFuture Policy Brief #5; Available online at: www.openfuture.eu/wp-content/uploads/2023/09/Best-_practices_for_optout_ML_training.pdf [Accessed on 31.12.2024]. Keller/Warso, 'Considerations for Opt-out Compliance Policies' (16 May 2024), Open Future Policy Brief #6 (2024), available at https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf [Accessed on 31.12.2024].

58 For example, German Explanatory memorandum proposed to incorporate such wording to Impressum. Czech SPIR recommended standardized wording for website header.

59 Keller/Warso, 'Considerations for Opt-out Compliance Policies' (16 May 2024), Open Future Policy Brief #6

those location-based identifiers is the mostly cited method of implementing the reservation from TDM exception is Robots.txt.[60] Alternatively, TDM fields in the HyperText Transfer Protocol (HTTP) Response header, TDM Metadata in HTML Content,[61] or various forms of access restrictions denying access to automated bots also come into consideration or expressions via terms and conditions of the website.[62] In addition, there can be numerous types of "*unit-based*" identifiers depending on type of content – for example TDM Metadata in EPUB files or metadata or watermarking of various types of media files.[63] Location-based identifiers are suitable mainly for those rightsholders who manage their own domains or sites, while those unit-based may be suitable for independent files especially when expecting subsequent spreading the respective files on the internet.[64]

19 Technical measures continuously evolve and will continue to evolve in the future. For example, Goole announced its plan to explore additional machine-readable means for web publishers[65] and Spawning AI created a Do Not Train registry and recently published the new option of ai.txt[66] which

(2024), available at https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf [Accessed on 31.12.2024].

60 Robots.txt is based on principles of good faith not technically preventing a robot from accessing the site, but merely expressing the intention not to allow automated robots access (primarily the case of Robots.txt or information embedded in the website header). Since the CDSM Directive solely requires that such reservation must (i) be machine-readable and (ii) express the rightsholder's will not to allow text and data mining, even voluntary expression should be sufficient. Nevertheless, the question whether voluntary measures can be considered as effectively expressing such reservation is controversial. Hugenholz names Robots.txt as a typical example of technical restrictions expressing reservation within the meaning of Art. 4 of the CDSM Directive, while Ducato and Strowel express arguments based on the InfoSoc Directive against such interpretation. Hugenholtz, B. The New Copyright Directive: Text and Data Mining (Articles 3 and 4) [online]. Kluwer Copyright Blog. 2019. Available at: http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ [Accessed on 31.12.2024]. R. Ducato and A. Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility"', CRIDES Working Paper Series (2018) 10.13140/RG.2.2.15392.84482. Available at SSRN: https://ssrn.com/abstract=3278901 or http://dx.doi.org/10.2139/ssrn.3278901 [Accessed on 31.12.2024].

61 See for example W3C TDMRep Final Community Group Report of 2 Feb 2024. Available at: https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240202/ [Accessed on 31.12.2024]. Keller/Warso, 'Considerations for Opt-out Compliance Policies' (16 May 2024), Open Future Policy Brief #6 (2024), available at https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf [Accessed on 31.12.2024]. Hanjo Hamann, Artificial Intelligence and the Law of Machine-Readability: A Review of Human-to-Machine Communication Protocols and their (In)Compatibility with Art. 4(3) of the Copyright DSM Directive, 15 (2024) JIPITEC 102 para 1.

62 Other measures, on the contrary, may directly block access to the given website when identifying automated crawlers through various bot-detection measures (namely CAPTCHA, browser challenges, browser fingerprinting, etc.) or enable access solely to verified human users accessing the content (namely password protections or similar access restrictions). Explicit denial of access to the given website e.g. by displaying error window (either after previous recognition of automated user based on bot-detection measures or after failure to pass log-in or registration path) could possibly also serve as a means of expressing such reservation within the meaning of Art. 4 of the CDSM Directive. However, implementation of these measures is not always user-friendly and desirable for the rightsholders. On the other hand, the sole implementation of bot-detection measures (for example CAPTCHA or browser challenges) without subsequently disabling access or expressing the intention not to grant such access in any way, could hardly have such legal relevance due to the absence of expression of rightsholder's will. There are further technical restrictions used to recognize bots and tactics aimed to make bot access more complicated, such as for example rate-limiting or crawl delay. However, since such measures solely to indirectly complicate bot access but do not clearly express the website holder's intention not to allow access via automated means, such could accordingly hardly have such legal relevance.

63 See for example W3C TDMRep Final Community Group Report of 2 Feb 2024. Available at: https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240202/ [Accessed on 31.12.2024]. Open Future, Open Future policy brief #6: Considerations for opt-out compliance policies by AI model developers. Available at: https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf [Accessed on 31.12.2024].

64 Open Future, Open Future policy brief #6: Considerations for opt-out compliance policies by AI model developers. Available at: https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf [Accessed on 31.12.2024].

65 In June 2023, Google suggested an option to explore additional machine-readable means for web publishers and to attempt finding new alternatives to robots.txt in connection with artificial intelligence and other emerging technologies. A principled approach to evolving choice and control for web content. Google Blog. Available at: https://blog.google/technology/ai/ai-web-publisher-controls-sign-up/ [Accessed on 31.12.2024].

66 Spawning is an independent third party that created a Do

was already cited by the French Data Protection Authority in terms of scraping publicly available personal data.[67] Other examples of such new means could be the TDM Reservation protocol (TDMRep)[68] or DeviantArt's noai meta-tags.

20 In addition, there may be other means specific for various member states within the EU. For example, German explanatory memorandum suggests that the reservation can be included in the imprint of a given website (*Impressum*) - which is a section typical for German websites – or terms and conditions, as long as such reservation is machine-readable.[69] As explained therein, the purpose and intention of the regulation is to give the rightsholders the opportunity to prohibit such use while at the same time ensuring that automated processes, which are a typical for text and data mining, can truly be carried out automatically for content that is accessible online.[70] Czech Association for Internet Development[71] recommends – besides Robots.txt – to place opt-out related wording to website footer which has been followed by some rightsholders in the Czech Republic.[72] French collective management society SACEM announced in its statement dated 12 October 2023 that it is opting out of machine learning training for the works in its repertoire.[73] Interestingly, Spanish Ministry of Culture and Sport recently published for public consultation a draft Royal Decree (*Proyecto de Real Decreto*) on Extended Collective Licensing introducing the idea of collective management of copyright-protected works in the development of AI models.

IV. Robots.txt and its Technical Limitations

21 Robots.txt is often cited as a typical example of technical restrictions expressing reservation within the meaning of Art. 4 of the CDSM Directive.[74] However, there are numerous practical constrains associated with using Robots.txt to express reservation from the TDM exception (especially for purposes of preventing use of data for generative AI training).

22 Robots.txt (or also called the *Robots Exclusion Protocol*) is a simple text file containing rules on which crawlers may access which parts of a site.[75] Robots.txt is based on voluntary basis meaning it does not technically block the automated access, but merely expresses the rules for access introduced by the given website. Robots.txt consists of set of rules stipulating the following information: (i) to whom the rule applies (the "user agent"); (ii) which directories or files that agent can access; and (iii) which directories or files that agent cannot access.[76] Interestingly, Robots.txt has been published in 1994[77] and *defacto* become

Not Train registry intended to provide machine readable opt-outs to AI model trainers.

67 Commission nationale de l'informatique et des libertés (CNIL) ; La base légale de l'intérêt légitime: fiche focus sur les mesures à prendre en cas de collecte des données par moissonnage (web scraping); Guidance issued on 10 July 2024, Available at: https://www.cnil.fr/fr/focus-interet-legitime-collecte-par-moissonnage [Accessed on 31.12.2024].

68 TDM Reservation Protocol (TDMRep); Available online at: https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240202/ [Accessed on 31.12.2024].

69 Explanatory memorandum (Gesetzesbegründung) of the German Government (Bundesregierung) to its legislative proposal implementing the CDSM Directive: Entwurf eines Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes, Gesetzesbegründung: Besonderer Teil. No. 19/27426. Page 88. Available at https://dip.bundestag.de/vorgang/.../273942 [Accessed on 31.12.2024].

70 Ibid.

71 Czech Association for Internet Development – in *Czech* as Sdružení pro internetový rozvoj (abbreviated as "SPIR").

72 SPIR press release: Online vydavatelé se vymezují proti vytěžování dat umělou inteligencí. [online]. Spir.cz. Available at: https://www.spir.cz/online-vydavatele-se-vymezuji-proti-vytezovani-dat-umelou-inteligenci/ [Accessed on 31.12.2024].

73 Although in the author's view such CMO's declaration placed on its own website can hardly fulfill the requirements of a valid express reservation in machine-readable means (without appropriate legal basis in the law). SACEM press release: Pour une intelligence artificielle vertueuse, transparente et équitable, la Sacem exerce son droit d'opt-out. [online]. societe.sacem.fr. Available at: https://societe.sacem.fr/actualites/notre-societe/pour-une-intelligence-artificielle-vertueuse-transparente-et-equitable-la-sacem-exerce-son-droit [Accessed on 31.12.2024].

74 Hugenholtz, B. The New Copyright Directive: Text and Data Mining (Articles 3 and 4) [online]. Kluwer Copyright Blog. 2019. Available at: http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ [Accessed on 31.12.2024]. As will be further outlined below, Robots.txt is currently recommended by key market players as a means to avoid being scraped in connection with AI training.

75 As follows from the Google guidelines for developers accessible online at https://developers.google.com/search/docs/crawling-indexing/robots/robots_txt or also at http://www.robotstxt.org/robotstxt.html [Accessed on 31.12.2024]

76 Google Developers: Introduction to Robots.txt. Available at: https://developers.google.com/search/docs/crawling-indexing/robots/intro [Accessed on 31.12.2024].

77 The standard, initially RobotsNotWanted.txt, allowed web developers to specify which bots should not access their website or which pages bots should not access. The internet was small enough in 1994 to maintain a complete list of all

a standard shortly after. There are the following historical descriptions of Robots.txt.: (i) the original 1994 A Standard for Robot Exclusion document[78]; and (ii) a 1997 Internet Draft specification A Method for Web Robots Control[79], further expanded by standard RFC 9309 Robots Exclusion Protocol.[80] As David Pierce said for the Verge, "*GPTBot has become the main villain of robots.txt because OpenAI allowed it to happen*" whereas "*it did all of this after training the underlying models that have made it so powerful*".[81] However, Robots.txt is not further actively developed.[82] In terms of potential future development, Google, while officially supporting Robots.txt as the means of expressing bot access rules, last year noted via its VP of trust Danielle Romain that "*We recognize that existing web publisher controls were developed before new AI and research use cases .... We believe it's time for the web and AI communities to explore additional machine-readable means for web publisher choice and control for emerging AI and research use cases.*"[83]

1. Generally Prohibiting all Text and Data Mining via User Agent Line Blocking all Bot Access?

23 Robots.txt differentiates specific terms for selected users (in the "*User-agent*" line of the Robots.txt) and URLS which may or may not be accessed (in the "*Disallow/Allow*" line of the Robots.txt).[84] The default rule usually is that a user agent can crawl any page or directory not blocked by a disallow rule. However, by generally blocking all automated access via robots.txt, such website could prevent Google[85] and other search engines from accessing and indexing the given website or could negatively impact how such website appears in search results in search engines, which considering the functioning of the internet might an undesirable scenario. As a result, rightsholders only rarely choose to disallow all bot access via Robots.txt.

24 "*User-agent*" line of Robots.txt allows to apply different reservation on various users (e. g. by allowing Google to crawl and index a website and prohibiting specific crawlers to scrape the website). The rightsholder may choose a "*whitelist*" of crawlers who may access the site[86] or *vice versa* a "*blacklist*" of crawlers who may not access the site. Such approach could be a reasonable solution for rights holders. However, such approach requires knowing the list of whitelisted or blacklisted users and knowing how to specifically identify such users in the "*User-agent*" line (to establish the machine-readability of the information for potential bots accessing such Robots.txt).

25 Robots.txt however does not enable prohibiting a specific purpose or means of use, i.e. prohibit any kind of text and data mining by any crawlers. Theoretically the user agent line could also identify group of crawlers, nevertheless such approach makes it even more difficult to decode the Robots. txt and could thus prevent the machine-readability of the "*User-agent*" line. An example may be the recommendation of the Czech Association for Internet Development recommending adding "*Machine Learning*" to "*User-agent*" line to prohibit large language models from accessing the site for AI training.[87] This approach has been subsequently implemented by some rightsholders in the Czech Republic[88], however, there are no available data as to whether this approach has been followed by AI companies.

bots; server overload was a primary concern.

78 A Standard for Robot Exclusion, document dated 30 June 1994 published at: http://www.robotstxt.org/orig.html [Accessed on 31.12.2024].

79 A Method for Web Robots Control; document dated 4 December 1994; published at: http://www.robotstxt.org/norobots-rfc.txt [Accessed on 31.12.2024].

80 Koster/Illyes/Zeller/Sassman, 'Standard RFC 9309: Robots Exclusion Protocol', as of Sep 2022, Available at: Rfc-editor.org/rfc/rfc9309.html [Accessed on 31.12.2024].

81 Pierce, D. The text file that runs the internet. The Verge (2024) [online]. Available at: https://www.theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders [Accessed on 31.12.2024].

82 What about further development of /robots.txt? Robots.org. [online]. Available at: http://www.robotstxt.org/faq/future.html [Accessed on 31.12.2024].

83 Romain, D. A principled approach to evolving choice and control for web content. Google Blog. [online]. Available at: https://blog.google/technology/ai/ai-web-publisher-controls-sign-up/ [Accessed on 31.12.2024].

84 Koster/Illyes/Zeller/Sassman, 'Standard RFC 9309: Robots Exclusion Protocol', as of Sep 2022, Available at: Rfc-editor.org/rfc/rfc9309.html [Accessed on 31.12.2024].

85 As follows from Google guidelines for developers accessible online at https://developers.google.com/search/docs/crawling-indexing/robots/intro [Accessed on 31.12.2024].

86 Nevertheless, rightsholders should, where sought, allow automated crawling through a website containing terms and conditions (especially where websites are protected by such technical restrictions) in order to enable a search through a website containing terms of use via automated means if the rightsholder wishes to apply these.

87 SPIR press release: Online vydavatelé se vymezují proti vytěžování dat umělou inteligencí. [online]. Spir.cz. Available at: https://www.spir.cz/online-vydavatele-se-vymezuji-proti-vytezovani-dat-umelou-inteligenci/ [Accessed on 31.12.2024].

88 See for example official Czech Press Agency under ctk.cz/robots.txt or also some Czech media platforms including idnes.cz/robots.txt. [Accessed on 31.12.2024].

26 In addition, there are numerous other practical constrains associated with proper implementation and proper decoding of rightsholder's opt out. For example, it is market-standard that crawlers search for Robots.txt solely on the top-level directory of a site.[89] However, the CDSM Directive does not introduce any such requirement and thus even files and information hidden in lower levels can be legally effective.

2. Identifying Scrapers in User-Agent Line?

27 Another issue associated with proper decoding of Robots.txt is the standardisation of its content as Robots.txt requires identification of the scraper in the User-agent line to be effectively implemented. However, it is the scrapers themselves who set their own name.[90] After strike of lawsuits in the USA, top AI market players have set the trend of publishing the recommended way to opt-out from their AI training.[91] This approach however requires rightsholders to monitor instructions published by all viable scrapers and currently also significantly disadvantages those AI developers, who take this step of proactively publishing their recommendations on their websites against those who do not do so (since as follows from the Originality.AI analysis explained below, websites tend to follow such recommendations and restrict use of their data to such user agents).

28 For example, on 7 August 2023 OpenAI published on its website a recommendation on how to disallow their GPTbot from accessing a website as follows:

> "*To disallow GPTBot to access your site you can add the GPTBot to your site's robots.txt:*
> *User-agent: GPTBot*
> *Disallow: /*"[92]

29 On 28 September 2023, Google announced a Google-Extended, a new control for web publishers[93] which enables to place "*Google-Extended*" to user-agent line of Robots.txt of rightsholder's websites to prevent its content to be used to train Bard (later re-named to Gemini) and Vertex AI generative APIs and future generations of models that power those products.

30 Common Crawl, non-profit foundation producing and maintaining an open repository of web crawl data,[94] published its recommended structure of Robots.txt to prevent Common Crawl from crawling a website and recommended implementing "*CCBot*" to the user-agent line.[95] According to a study published in 2020, OpenAI's GPT-3 was trained using data mostly collected from Common Crawl.[96] On the other hand, Common Crawl is used for a variety of other purposes unrelated to generative artificial intelligence.[97]

31 In June 2024, another key AI market player Anthropic AI[98], developer of large language model called Claude, published its recommendation for placing "*ClaudeBot*" to the user-agent line of Robots.txt.[99]

32 Shortly prior to the above, on 7 July 2023 Czech Association for Internet Development[100] issued a recommendation to rightsholders on how to implement the reservation from the TDM exception within Robots.txt as follows:

> "*User-agent: MachineLearning*

89 See, for example, a recommendation in the Google guidelines for developers accessible online at https://developers.google.com/search/docs/crawling-indexing/robots/robots_txt [Accessed on 31.12.2024].

90 Koster/Illyes/Zeller/Sassman, 'Standard RFC 9309: Robots Exclusion Protocol', as of Sep 2022, Available at: Rfc-editor.org/rfc/rfc9309.html [Accessed on 31.12.2024].

91 No Robots(.txt): How to Ask ChatGPT and Google Bard to Not Use Your Website for Training. Electronic Frontier Foundation. [online]. Available at https://www.eff.org/deeplinks/2023/12/no-robotstxt-how-ask-chatgpt-and-google-bard-not-use-your-website-training [Accessed on 31.12.2024]. How to block AI crawlers with robots.txt. Netfuture. [online]. Available at https://netfuture.ch/2023/07/blocking-ai-crawlers-robots-txt-chatgpt/ [Accessed on 31.12.2024].

92 GPTBot. Available at https://platform.openai.com/docs/gptbot [Accessed on 31.12.2024].

93 An update on web publisher controls. Google Blog. [online]. Available at: https://blog.google/technology/ai/an-update-on-web-publisher-controls/ [Accessed on 31.12.2024].

94 Common Crawl is a non-profit foundation founded with the goal of democratizing access to web information by producing and maintaining an open repository of web crawl data that is universally accessible and analyzable by anyone. Common Crawl, CCBot. [online]. Available at: https://commoncrawl.org/ccbot [Accessed on 31.12.2024].

95 Common Crawl, CCBot. [online]. Available at: https://commoncrawl.org/ccbot [Accessed on 31.12.2024].

96 Brown, T.T., et al., Language Models are Few-Shot Learners. Available at: https://arxiv.org/pdf/2005.14165 [Accessed on 31.12.2024].

97 Common Crawl, Use cases. [online]. Available at: https://commoncrawl.org/use-cases [Accessed on 31.12.2024].

98 Anthropic has developed a family of large language models (LLMs) named Claude as a competitor to OpenAI's ChatGPT and Google's Gemini.

99 Does Anthropic crawl data from the web, and how can site owners block the crawler? [online]. Available at: https://support.anthropic.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler [Accessed on 31.12.2024].

100 Czech Association for Internet Development – in *Czech* as Sdružení pro internetový rozvoj (SPIR) -

Disallow: /"[101]

33 As can be seen from Robots.txt implemented by some media companies[102], many have implemented these solutions recommended by these AI companies. In 2023, Originality.AI analysed the top 1000 websites in the world to identify which sites are already blocking GPTBot[103] and later added also the CCBot, Google-Extended bot and anthropic-ai. As of June 2024, OriginalityAI found that 350 out of the 1000 websites, i.e. 35 %, block GPTBot, 216 out of the 1000 websites, i.e. 21,60% block CCBot, 126 out of the 1000 websites, i.e. 12.60 % block Google-Extended bot and 84 websites out of 1000 websites, i.e. 8.40% block anthropic.ai. As Originality.AI originally noted, "*it is not clear if "anthropic-ai" and "claude-web" would be effective as there has been no documentation from Anthropic*." (although in the meantime Anthropic published its recommendation). [104]

34 As a result, technical limitations of Robots.txt solution inevitably lead to the consequence that those companies which take this step of proactively publishing the identification of their scrapers are more likely to be excluded by rightsholders from use of their data. On the contrary, those scrapers who are not known to the rightsholders are less likely to be covered in rightsholders reservations. This result however does not seem to be fair as it is disadvantageous for those companies who publish their User agent instructions and motivates the other not to voluntarily publish this information.

35 Potential solution to the above technical limitations could be either a completely new solution designed to implement TDM exception and express rightsholder's rules for use of content for AI training. For example, the European Commission recently announced its plan to conduct a feasibility study on the creation of a central registry where rights holders could opt out from TDM. The purpose of the study is to assess both the opportunity and feasibility of developing a work-based registry of content identifiers and associated metadata that would support – whether centrally or within a federated network– the effective expression of TDM opt-outs and facilitate their identification by AI developers. This could be a possible solution which might however require robust technical solution (which is to be explored by the aforementioned feasibility study).[105]

36 Alternatively, if Robots.txt is to be used, generic wording of User Agent line enabling to express reservation from TDM exception without applying differing rules for various scrapers could appear to be fair and workable solution. For example, as Open Future Policy Brief suggests, these could take the form of wildcard user-agent names such as **-genai, *-tdm, *-aiuser*[106] or the form of *MachineLearning* as suggested by Czech SPIR.[107] Alternatively - instead of such binary opt-out/non-opt-out approach allo - such unified vocabulary could introduce even more granular taxonomy of use cases for rightsholders to opt out from.[108] This solution could for example enable rightsholders to prohibit TDM for generative AI training but allow use for other forms of AI.[109] However, such solution could be even more complicated to unify which is the main issue in the existing technological landscape.

101 Recommendation of Czech Association for Internet Development. Online vydavatelé se vymezují proti vytěžování dat umělou inteligencí. [online]. Available at https://www.spir.cz/online-vydavatele-se-vymezuji-proti-vytezovani-dat-umelou-inteligenci/ [Accessed on 31.12.2024].

102 Media companies' websites are typically those publicly available websites who can be expected to publish copyright-protected content.

103 AI Bot Blocking. OriginalityAI. [online]. Available at https://originality.ai/ai-bot-blocking [Accessed on 31.12.2024].

104 Note: In the meantime, Anthropic AI published its recommendation on implementing „ClaudeBot" within Robots.txt.

105 Study to assess the feasibility of a central registry of Text and Data Mining opt-out expressed by rightsholders, Accessible under File No. EC-CNECT/2025/OP/0002 in the EU Funding & Tenders Portal. Available online at: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/tender-details/8726813a-bd9b-4f58-8679-01c80f7a1abf-CN?isExactMatch=true&order=DESC&pageNumber=1&pageSize=50&sortBy=startDate [Accessed on 20.03.2025].

106 Keller/Warso, 'Considerations for Opt-out Compliance Policies' (16 May 2024), Open Future Policy Brief #6 (2024), available at https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf [Accessed on 31.12.2024].

107 Recommendation of Czech Association for Internet Development. Online vydavatelé se vymezují proti vytěžování dat umělou inteligencí. [online]. Available at https://www.spir.cz/online-vydavatele-se-vymezuji-proti-vytezovani-dat-umelou-inteligenci/ [Accessed on 31.12.2024].

108 Ibid.

109 For example, C2PA approach distinguishes between data_mining, ai_training, ai_generative_training, and ai_inference. See standards introduced by the Coalition for Content Provenance and Authenticity (C2PA). Available at: https://c2pa.org/. [Accessed on 31.12.2024]. Other approaches (such as Spawning's products and the DeviantArt no-ai meta tag) are specifically targeted at (generative) AI training, while others (such as TDMRep) are explicitly aimed at the full spectrum of text and data mining. See Keller/Warso, 'Considerations for Opt-out Compliance Policies' (16 May 2024), Open Future Policy Brief #6 (2024), available at https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf [Accessed on 31.12.2024].

37 Once the AI Act comes into effect, all providers placing general-purpose AI models on the market in the EU will be obliged to publish their policies on how they comply and identify with rightsholder's opt-out. It is not yet clear whether these policies will follow the same path taken by OpenAI, Common Crawl, Google and others and will contain instructions for User Agent line to prevent their scrapers accessing the respective content. Recently the AI Office published the first draft Code of Practice for general-purpose AI models for public consultation which however solely suggests that "*Signatories will only employ crawlers that read and follow instructions expressed in accordance with the Robot Exclusion Protocol (robots.txt)"*. The Code of Practice undergoes multiple rounds of consultations and is expected to be finalized before May 2025.

V. Burden of Proof & Logging Evidence of Valid Opt-Out

38 The reservation from the TDM exception should in the author's view be effective after being placed at the respective website. Prior to that moment the TDM exception applies without such condition that rightsholders expressly reserved their rights. Although as for example Peter Mézei aptly points out "*the directive neither prompts nor excludes that such reservations should be carried out ex ante (preceding the mining) or ex post (following the mining)*." while noting that "*TDM might happen quicker than an ex-ante reservation could have been expressed. Consequently, ex post reservations shall not be automatically excluded from the scope of Art. 4(3).".*[110] On the contrary, for example Czech Explanatory Memorandum explicitly highlights that reservation applies solely for future use and cannot apply retrospectively.[111] Such conclusion may follow also from past tense forms used in some member state laws implementing the CDSM Directive - for example in the German[112], Czech[113], Austrian[114] implementation. In addition, requiring the developer to do so does not seem to be proportionate in case the developer has lawfully relied on an exception from copyright protection allowing to *retain reproductions for as long as is necessary for the purposes of text and data mining*.[115] *Ex-ante* reservations also correspond to technological reality as once an AI model is trained, the copyright protected content can hardly be retrospectively removed from the original training data. As Open Future Policy Brief notes, for each version of AI model, there could be some sort of *opt-out cut-off date, after which new opt-outs will no longer affect the model's training* whereas such cut-off date could be transparently communicated once AI model is released.[116]

39 However, the existence of a reservation as of

110 Mezei, Péter, A saviour or a dead end? Reservation of rights in the age of generative AI (January 15, 2024). European Intellectual Property Review, 2024, 46(7), p. 461-469. Available at SSRN: https://ssrn.com/abstract=4695119 or http://dx.doi.org/10.2139/ssrn.469511. [Accessed on 31.12.2024]. Page 8.

111 Explanatory memorandum (Důvodová zpráva) of the Czech Government to the Act. No. 429/2022 Coll. (amending the Czech Copyright Act implementing the CDSM Directive). Section § 39c.

112 § 44b (3) of the German Urheberrechtsgesetz stating that *„(3) Nutzungen nach Absatz 2 Satz 1 sind nur zulässig, wenn der Rechtsinhaber sich diese nicht vorbehalten hat. Ein Nutzungsvorbehalt bei online zugänglichen Werken ist nur dann wirksam, wenn er in maschinenlesbarer Form erfolgt.*" or as translated to English: *„(3) Uses pursuant to paragraph 2 sentence 1 shall only be permitted if the rightholder has not reserved the right of use. A reservation of use in the case of works accessible online shall only be effective if it is made in machine-readable form.*"

113 § 39 c (2) of the Czech Copyright Act stating that *"(2) Ustanovení odstavce 1 se nepoužije pro rozmnoženiny díla, jehož autor si užití podle odstavce 1 výslovně vyhradil vhodným způsobem; v případě díla zpřístupněného podle § 18 odst. 2 strojově čitelnými prostředky*." or as translated to English: *„2) The provision of paragraph 1 does not apply to reproductions of the work, the author of which has expressly reserved the use according to paragraph 1 in an appropriate manner; in the case of a work made available in accordance with § 18 paragraph 2 by machine-readable means"*

114 § 42 h of the Austrian Urheberrechtsgesetz stating that "*(6)Jedermann darf für den eigenen Gebrauch ein Werk vervielfältigen, um damit Texte und Daten in digitaler Form automatisiert auszuwerten und Informationen unter anderem über Muster, Trends und Korrelationen zu gewinnen, wenn er zu dem Werk rechtmäßig Zugang hat. Dies gilt jedoch nicht, wenn die Vervielfältigung ausdrücklich verboten und dieses Verbot in angemessener Weise durch einen Nutzungsvorbehalt, und zwar etwa bei über das Internet öffentlich zugänglich gemachten Werken mit maschinenlesbaren Mitteln, kenntlich gemacht wird. Eine Vervielfältigung nach diesem Absatz darf aufbewahrt werden, solange dies für die Zwecke der Datenauswertung und Informationsgewinnung notwendig ist.*" or as translated to English *"(6) Anyone may reproduce a work for their own use in order to automatically evaluate texts and data in digital form and to obtain information on patterns, trends and correlations, among other things, if they have lawful access to the work. However, this shall not apply if reproduction is expressly prohibited and this prohibition is appropriately indicated by a reservation of use, for example in the case of works made publicly accessible via the Internet by machine-readable means. Reproduction in accordance with this paragraph may be retained as long as this is necessary for the purposes of data analysis and information retrieval."*

115 Art. 4 (2) CDSM Directive.

116 Keller/Warso, 'Considerations for Opt-out Compliance Policies' (16 May 2024), Open Future Policy Brief #6 (2024), available at https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf [Accessed on 31.12.2024].

certain moment in time may be practically difficult to prove in potential dispute without for example time-stamped evidence proving the existence of reservation from the TDM exception as of certain specific moment in time. In the event of a potential dispute, rightsholders as potential plaintiffs might be claiming copyright infringement whereas scrapers as potential defendants might be claiming that TDM exception applies. Therefore, the rightsholders will likely bear the burden of proof that copyright infringement occurred whereas scrapers will likely bear the burden of proof of lawful use of content and thus proving that TDM exception applies. German explanatory memorandum suggests that the burden of proof for the absence of a reservation shall be born the user who is relying on such exception.[117] Löbling, Handschigl, Hofmann and Schwedhelm are of the view that "*TDM user bears the onus of proof, mandated by the phrasing of paragraph 3 ("are permitted only if they have not been reserved*"), although acknowledge that "*copyright holder is accountable for properly expressing their opt-out decision*".[118] This question will remain to be addressed by civil procedural rules which differ in EU member states.

D. Remarks on the Interplay between TDM Exceptions under Art. 3 and 4 CDSM Directive Considering Practicality of Gen AI Development

40 Datasets are not always created by the same legal entities which are developing artificial intelligence. On the contrary, datasets are often populated by various third parties or non-profit organisations and only subsequently cleansed, adjusted and used by AI companies to train Gen AI.[119] This follows for example from limited publicly available information suggesting that some large language models might have been trained on datasets such as Common Crawl, LAION, BookCorpus, Wikipedia, WebText.[120] Each of these examples implements different purpose and *modus operandi* – for example Common Crawl is a non-profit organisation publishing a dataset consisting of raw web page data, metadata extracts, and text extracts collected from publicly available websites since 2008[121], LAION on the other hand provides publicly and free of charge a dataset for image-text pairs consisting of hyperlinks to images or image files publicly accessible on the Internet as well as other information related to the respective images, including an image description.[122] However, while these repositories – due to their non-profit nature - might themselves rely on TDM Exception for research purposes – those AI companies using their data might not. Although the Common Crawl Foundation proclaims to comply with Robots.txt and *no follow policies* of the scraped websites (for these purposes the Common Crawl Foundation even issued its own Robots.txt guidance recommending implementing "*CCBot*" to the user-agent line[123]), at the same time Common Crawl's publicly available Terms of use explicitly limit Common Crawl's liability for third party IP infringements and explicitly state that Crawled Content may be subject to separate terms of use or terms of service from the owners of such Crawled Content.[124] These aspects add additional layer of complexity in potential disputes over lawfulness of text and data mining. For example, in *Robert Kneschke v. LAION* the court tackled solely the use of protected content by LAION (as the defendant) but subsequent use of LAION datasets by AI developers was not part of the case.[125]

41 Both TDM exceptions are by virtue of the definition of text and data mining limited to actions aiming to *generate information.* Such requirement is stemming from the legal definition of text and data mining as a legal term defined in the CDSM Directive as "*any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations*".[126] As indicated in the preamble of the

117 Explanatory memorandum (Gesetzesbegründung) of the German Government (Bundesregierung) to its legislative proposal implementing the CDSM Directive: Entwurf eines Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes, Gesetzesbegründung: Besonderer Teil. No. 19/27426. Page 88. Available at https://dip.bundestag.de/vorgang/.../273942 [Accessed on 31.12.2024].

118 Löbling, L., Handschigl, Ch. Hofman, K., Schwedhelm, J. Navigating the Legal Landscape: Technical Implementation of Copyright Reservations for Text and Data Mining in the Era of AI Language Models. 14 (2023) JIPITEC 499 para 12.

119 Generally, preparation of dataset for AI training involves very thorough process involving data cleansing, de-duplication and other measures aiming to enhance dataset quality.

120 Brown, T.T., et al., Language Models are Few-Shot Learners. Available at: https://arxiv.org/pdf/2005.14165 [Accessed on 31.12.2024]. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I., Language Models are Unsupervised Multitask Learners, OpenAI, 2019.

121 Common Crawl, *Common Crawl Overview*, available at: https://commoncrawl.org/overview [Accessed on 31.12.2024].

122 LG Hamburg, Urteil vom 27. September 2024 – 310 O 227/23 (Robert Kneschke v. LAION).

123 Common Crawl, CCBot. [online]. Available at: https://commoncrawl.org/ccbot [Accessed on 31.12.2024].

124 Terms of Use of Common Crawl, available online at: https://commoncrawl.org/terms-of-use [Accessed on 31.12.2024].

125 Although partially mentioned in the *obiter dictum.*

126 Article 2 (2) CDSM Directive.

CDSM Directive, the text and data mining exceptions aim to encourage innovation in both the public and private sectors as legislators acknowledge its benefits in enabling the processing of large amounts of information with a view to "*gaining new knowledge and discovering new trends*".[127] Narrowing the definition of text and data mining solely to the purpose of *generating information* reflects the overarching goal of the CDSM Directive. For example, as noted in the German explanatory memorandum, the purpose of the text and data mining covered by the exception does not cover actions aimed at collecting and storing content to create parallel digital archives.[128] German court in *Robert Kneschke v. LAION* offered interesting perspective and interpreted the requirement of generating new information very broadly. The court applied TDM Exception with an explanation that the defendant undertook the reproduction action for the purpose of extracting information about "*correlations*" to compare the image content with the image description already stored in the text using an available software application. The court noted that although the creation of the dataset itself may not yet be associated with a knowledge gain, it is a fundamental step aimed at using the dataset for the purpose of later knowledge acquisition. The court held as sufficient that the dataset was undisputedly published for free and thus made available, particularly to researchers working in the field of artificial neural networks. However, the court considered as irrelevant whether such other researchers are commercial enterprises or non-profit undertakings.

42 However, although such interpretation has positive impact on innovation allowing such organisations to create and publish open-source datasets, such interpretation might not hold up. As explained above, some organisations might be merely populating publicly available data and publishing the respective datasets for non-profit research purposes, however, not train AI or generate new information themselves. On the contrary, such dataset created for non-profit purposes may be subsequently used by companies developing Gen AI on a for-profit basis. That could however mean that strictly speaking, companies creating datasets are *stricto sensu* not generating new information and on the contrary, reproductions made by companies developing Gen AI on a for-profit basis cannot be covered by the research exception. In addition, such argumentation had justification with respect to LAION as it does not publish the original works but solely hyperlinks and concurrently indeed provides analysis of the correlations. The same modus operandi however might not apply to other dataset publishers.

43 In instances where an AI model is initially developed under a non-profit framework, adheres to removing original datasets post-training, and later transitions into commercial use, the initial reproduction or extraction activities could technically still fall within the TDM exception under the CDSM Directive for non-commercial research purposes. However, this exception would strictly apply only to those preliminary reproduction and extraction actions within the non-profit stage. Any subsequent activities, including storage of original raw data or dissemination of copyrighted material within AI outputs that might arise due to data memorization, fall outside this exception as further described below. Lastly, if companies that create datasets are found to infringe on copyright, such infringement could potentially compromise the legality of AI companies' subsequent use of the datasets. Even if these AI companies duly rely on the TDM exception under Art. 4 of the CDSM Directive, initial copyright infringement might lead to unlawful access, conflicting with the lawful access requirement outlined in Articles 3 and 4 of the Directive.

E. Does the TDM Exception Really Provide an Answer? Is it Technically Possible to Train Gen AI but Prevent Verbatim Extracts of Training Data in Gen AI Outputs?

44 Due to the limited scope of 3 and 4 CDSM Directive, both TDM Exceptions cover solely the acts of reproduction but not subsequent modifications or communication to the public / reutilization of the original data.[129] Specifically, TDM exception covers

The following key elements of text and data mining can be derived from this legal definition: (i) automated analytical techniques; (ii) analysis of text and data in digital form; (iii) aim intended to generate information (including patterns, trends and correlations).

127 Recital 8 and 18 CDSM Directive.

128 Explanatory memorandum (Gesetzesbegründung) of the German Government (Bundesregierung) for a legislative proposal implementing the CDSM Directive: Entwurf eines Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes, Gesetzesbegründung: Besonderer Teil. Page 88. Available at https://dip.bundestag.de/vorgang/.../273942 [Accessed on 31.12.2024]. Page 88.

129 E Rosati, The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspects, available at http://www.europarl.europa.eu/RegData/etudes/BRIE/2018/604942/IPOL_BRI(2018)604942_EN.pdf. [Accessed on 31.12.2024]. Novelli, Claudio and Casolari, Federico and Hacker, Philipp and Spedicato, Giorgio and Floridi, Luciano, Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity (January 14, 2024). Available at SSRN: https://ssrn.com/abstract=4694565 or http://dx.doi.org/10.2139/ssrn.4694565 [Accessed on 31.12.2024].

(i) the *right of reproduction* of copyrighted works[130], databases[131], and on-demand press publications[132]; (ii) *the right of extraction* of a whole or a substantial part of databases covered by the sui generis database rights[133]; and (iii) the right to reproduction and the right to adaptation of computer programs[134].[135]

45 As follows from claims filed in the US and UK[136], plaintiffs often claim not only use of their works in connection with AI training but also further dissemination of their works within AI outputs which in terms of EU law would exceed the scope of right of reproduction and may constitute a communication to the public (as for example follows from the complaint filed by The New York Times Company against Microsoft and OpenAI or class action complaint filed by the US Authors Guild against Microsoft and OpenAI).

46 The act of text and data mining occurs at the early stage of model development. During this phase, the model is trained on such datasets. Although large language models might not be technically storing the original datasets and raw data used for training; such models may sometimes retain and produce verbatim snippets or other identifiable data elements due to a phenomenon known as data memorization. Data memorization occurs for example when specific data points, such as text or images, are repeatedly encountered during training, leading the model to "*memorize*" these elements, sometimes resulting in output that closely resembles or directly mirrors segments of the original data.[137] As Carlini concluded "*Memorization significantly grows as we increase (1) the capacity of a model, (2) the number of times an example has been duplicated, and (3) the number of tokens of context used to prompt the model*".[138]

47 Although TDM exceptions may serve as a legal basis authorizing use of protected content for purposes of AI training, they might not justify subsequent reuse the respective content in case generative AI models produce verbatim snippets of original works.[139] Practical solution may be implementation of additional measures. For example, de-duplication[140] of training data which is considered to be one of available countermeasures against data memorization[141] whereas "*the core idea is to remove any duplicated content—e.g., repeated documents—because duplicated content is much more likely to be memorized. However, deduplication does not guarantee that a model will not still memorize individual (deduplicated) examples*. In addition, applying various types of output filters may prevent further dissemination of the protected content within AI outputs such as retroactive censoring or memfree decoding which explicitly "*prohibit the model from emitting a sequence if it is contained (entirely or partially) in the training dataset*".[142] For example, GitHub's Copilot, a language

130 Article 2 InfoSoc Directive.

131 Article 5(a) Database Directive.

132 Article 15(1) CDSM Directive.

133 Article 7(1) Database Directive.

134 Articles 4(1)(a) and (b) InfoSoc Directive.

135 The scope of exception under Article 4, CDSM Directive is broader than the exception under Article 3 of the CDSM Directive (i.e. TDM for scientific purposes), which unlike Article 4 of the CDSM Directive does not cover the right to reproduction and the right to adaptation of computer programs.

136 See footnote 5.

137 Biderman, S., Prashanth, U. S. S., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., & Raff, E., 2023. *Emergent and Predictable Memorization in Large Language Models.* arXiv preprint arXiv:2304.11158v2 [cs.CL]. Available at: https://doi.org/10.48550/arXiv.2304.11158 [Accessed on 31.12.2024]. Huang, J., Yang, D., & Potts, C., 2023. Demystifying Verbatim Memorization in Large Language Models. Stanford University. Available at: https://arxiv.org/ [Accessed on 31.12.2024]. Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., & Zhang, C., 2023. Quantifying Memorization Across Neural Language Models. Google Research and Cornell University. Available at: https://arxiv.org/ [Accessed on 31.12.2024]. Albert Ziegler, "GitHub Copilot research recitation" Github blog, 30 June 2021; Available at: https://github.blog/2021-06-30-github-copilot-research-recitation [Accessed on 31.12.2024]. Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini, 'Preventing Verbatim Memorization in Language Models Gives a False Sense of Privacy', arXiv (2023), arXiv:2210.17546v3 [cs.LG], pp. 1–26. Gowthami Somepalli, Vasu Singla, Micah Goldblum, Joans Geiping & Tom Goldstein, "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models" (2023) https://arxiv.org/abs/2212.03860 [Accessed on 31.12.2024]. Nicholas Carlini. Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito & Eric Wallace, "Extracting Training Data from Diffusion Models" (2023).

138 Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., & Zhang, C., 2023. Quantifying Memorization Across Neural Language Models. Google Research and Cornell University. Available at: https://arxiv.org/ [Accessed on 31.12.2024].

139 Rosati, Eleonora, Infringing AI: Liability for AI-generated outputs under international, EU, and UK copyright law (August 31, 2024). European Journal of Risk Regulation, Available at SSRN: https://ssrn.com/abstract=4946312 or http://dx.doi.org/10.2139/ssrn.4946312 [Accessed on 31.12.2024].

140 Data deduplication has arisen as a pragmatic countermeasure against data memorization (Lee et al., 2021; Kandpal et al., 2022; Carlini et al., 2022). The core idea is to remove any duplicated content—e.g., repeated documents—because duplicated content is much more likely to be memorized.

141 Lee et al., 2021; Kandpal et al., 2022; Carlini et al., 2022.

142 Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini, 'Preventing

model-based code assistant, adopts similar measures and offers users to "*block suggestions matching public code*".[143] However, previous research indicates that even when a model is restricted from emitting any output with snippets of verbatim memorization, the model might still leak some parts of training data.[144] For example, research testing GitHub Copilot which implemented retroactive censoring shows that "*Copilot's filter can easily be bypassed by prompts that apply various forms of "style-transfer" to model outputs, thereby causing the model to produce memorized (but not verbatim) outputs*".[145] On the other hand, such "style-transfer" outputs may significantly less likely constitute copyright infringement than verbatim snippets depending on the level of autonomy of the creation and dependency on the pre-existing content.[146] Such assessment however depends on case-by-case basis taking into account also involvement of the user prompting the LLM.[147] in such case the burden of proof of the respective copyright infringement lies with the rightsholders potentially claiming such infringement.

48 As a result, even when duly and lawfully applying TDM exception for purposes of text and data mining to facilitate generative AI training, AI models may still face significant challenges and difficulties to rely on text and data mining within the legal borderlines of copyright laws.

F. Concluding Remarks

49 This paper highlighted practical challenges tied to TDM exceptions, which may inevitably come up in disputes over AI-related copyright infringements. For example:

- Machine-readable reservation allowing rightsholders to opt-out from for-profit TDM exception may hit the barrier of lacking standardisation.

- The CDSM Directive does not define the required level of "*machine-readability*" for rightsholders' reservations. German court noted that "*machine-understandability*" may be sufficient depending on technical developments at the relevant time of use. With such justification the court considered even terms and conditions in human language as machine-readable since such terms may be decoded by generative AI. German court in Robert Kneschke v. LAION noted that "these "state-of-the-art technologies" undoubtedly include, in particular, AI applications capable of comprehending text written in natural language" which might however not achieve sufficient level of reliability and thus applying these conclusions would pose significant risks for AI companies relying on such technologies to decode rightsholders' opt out.

- However, in order to strike a balance between the interests of users of text and data mining (to be able to conduct automated analysis of data) and the interests of rights holders (to protect their rights), this rightsholders' "*express*" reservations in "*machine-readable*" formats should, in the author's view, achieve *sufficiently reliable level* of machine interpretability which might not be achieved when relying on Gen AI decoding terms and conditions written in natural language. This might require the reservation to be presented in a sufficiently *standardized form* that enables such advanced technologies to *reliably* decode its content leaving no room for doubt. This may be reflected for example by standardized formulas (despite being written in a natural human language) – for example similarly as open-source licensing terms.

- Robots.txt is a key tool for expressing reservations but its simplicity can lead to technical limitations and unintended side effects. Prohibiting all bot access via Robots.txt affects website indexing by search engines, making it largely impractical.

- Currently, Robots.txt cannot block specific uses like text and data mining; it only allows naming specific scrapers in the user-agent line. Some AI market players set the trend of publishing instructions for the user-agent line to block their scrapers and opt-out from their AI training. However, this requires rightsholders to monitor all viable scrapers and disadvantages those AI companies who publish these instructions (since websites typically follow these recommendations if published and restrict data use for specified user agents) while practically favouriting those who do not (as the rightsholders do not know how to identify them in the User agent line).

Verbatim Memorization in Language Models Gives a False Sense of Privacy', arXiv (2023), arXiv:2210.17546v3, pp. 1–26. [Accessed on 31.12.2024].

143 Ibid.

144 Ibid.

145 Ibid.

146 Novelli, Claudio and Casolari, Federico and Hacker, Philipp and Spedicato, Giorgio and Floridi, Luciano, Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity (January 14, 2024). Available at SSRN: https://ssrn.com/abstract=4694565 or http://dx.doi.org/10.2139/ssrn.4694565 [Accessed on 31.12.2024].

147 Ibid.

- AI Act implementations might bring clarity by once providers of general-purpose AI models publish TDM compliance policies following state-of-the-art technologies, though this applies only to companies marketing such models in the EU.

- Potential disputes will also inevitably involve practical and procedural challenges, such as determining the extent of each party's burden of proof and how to demonstrate that a reservation was made at a specific point in time.

50 Consequently, given the practical and technical limitations discussed in this paper, developing a clear market standard solution that both AI developers and rightsholders can adhere to would be highly beneficial. Standardized TDM identifiers will enable to streamline opt-out processes and will reduce costs and increase legal certainty for both rightsholders and AI companies.

51 Nevertheless, since TDM exceptions allow solely the acts of reproduction / extraction but not subsequent modification and use – even if TDM part of AI training is resolved, AI companies will still have to carefully tackle the risks of any data memorization which may lead to producing verbatim snippets of training data which would not be legitimized by the TDM exceptions. As a result, even when duly and lawfully applying TDM exceptions to legitimize use of data for generative AI training, AI models may still face significant challenges and difficulties when scraping copyright protected content without a license from the rightsholders.

52 To the very end, machine-readable reservations allowing rightsholders to opt out of for-profit TDM exceptions could grant the rightsholders significant power, potentially leading to widespread withdrawal from AI training. This might deprive the EU public of future AI innovations using high quality datasets while at the same time not enabling the authors from benefitting therefrom (for example by offering their content in exchange for remuneration). Solutions such as machine-readable licensing models or collective management, could offer a balanced compromise between protecting rightsholders' rights and fostering AI development. Such solutions would however either require significant legislative changes or robust licensing frameworks and data spaces enabling to acquire license via automated means.