Generative AI and Creative Commons Licences

The Application of Share Alike Obligations to Trained Models, Curated Datasets and AI Output ***

by **Kacper Szkalej * and Martin Senftleben ****

Abstract: This article maps the impact of Share Alike (SA) obligations and copyleft licensing on machine learning, AI training, and AI-generated content. It focuses on the SA component found in some of the Creative Commons (CC) licences, distilling its essential features and layering them onto machine learning and content generation workflows. Based on our analysis, there are three fundamental challenges related to the life cycle of these licences: tracing and establishing copyright-relevant uses during the development phase (training), the interplay of licensing conditions with copyright exceptions and the identification of copyright-protected traces in AI output. Significant problems can arise from several concepts in CC licensing agreements ('adapted material' and 'technical modification') that could serve as a basis for applying SA conditions to trained models, curated datasets and AI output that can be traced back to CC material used for training purposes. Seeking to transpose Share Alike and copyleft approaches to the world of generative AI, the CC community can only choose between two policy approaches. On the one hand, it can uphold the supremacy of copyright exceptions. In countries and regions that exempt machine-learning processes from the control of copyright holders, this approach leads to far-reaching freedom to use CC resources for AI training purposes. At the same time, it marginalises SA obligations. On the other hand, the CC community can use copyright strategically to extend SA obligations to AI training results and AI output. To achieve this goal, it is necessary to use rights reservation mechanisms, such as the opt-out system available in EU copyright law, and subject the use of CC material in AI training to SA conditions. Following this approach, a tailor-made licence solution can grant AI developers broad freedom to use CC works for training purposes. In exchange for the training permission, however, AI developers would have to accept the obligation to pass on – via a whole chain of contractual obligations – SA conditions to recipients of trained models and end users generating AI output.

Keywords: Copyright, AI, Machine Learning, Licensing, Creative Commons, Share Alike

Recommended citation: Kacper Szkalej and Martin Senftleben, Generative AI and Creative Commons Licences The Application of Share Alike Obligations to Trained Models, Curated Datasets and AI Output, 15 (2024) JIPITEC 313 para 1

* Dr, Researcher in Intellectual Property Law, Institute for Information Law (IViR), University of Amsterdam, The Netherlands.

** Professor of Intellectual Property Law and Director, Institute for Information Law (IViR), University of Amsterdam; Of Counsel, Bird & Bird, The Hague, The Netherlands.

*** This article is based on the study *Mapping the Impact of Share Alike/Copyleft Licensing on Machine Learning and Generative AI* for which the authors secured funding from the Open Future Foundation, available at: https://openfuture.eu/publication/the-impact-of-share-alike-copyleft-licensing-on-generative-ai/. Both the study and this article have been established in complete academic independence.

A. Introduction

1 The increasing impact of AI on the copyright system has led to a multi-faceted discussion ranging from the creation of breathing space for text and data mining (TDM)[1] to the potential displacement of human creative labour.[2] A further facet of this debate is the impact of contractual obligations relating to copyright-protected training material on machine learning (ML) and, more broadly, the development and exploitation of generative AI systems. Copyleft licensing strategies and ShareAlike clauses found in some CC licences (we will call them collectively CLSA) impose obligations on the recipient to use the same licensing model that underlies the original license for downstream use.[3] In this way, a network effect is ensured that preserves and extends the commons. However, when copyright-protected knowledge resources are released under CLSA clauses and subsequently used for AI training, the question arises whether such licences provide the same safeguards for commons-based projects that they provide in case of more traditional uses. In the following analysis, we examine this question step-by-step. First, we explain the anatomy of CLSA licences and shed light on main features relevant to downstream use (section B). Second, we identify acts of use with copyright relevance in AI training processes (section C) before turning to the sensitive question of whether TDM exceptions are capable of prevailing over SA obligations and rendering corresponding licensing terms inapplicable (section D). On this basis, we explore more closely the application of the CLSA concept of 'adapted materials' to generative AI development and exploitation (section E). Finally, we discuss different strategies to ensure that SA obligations remain intact and can be passed on to AI developers, recipients of trained AI models and end users generating AI output (section F).

2 Throughout the analysis we will refer to the *development phase* by which we mean the entire ML-process culminating in the creation of a generative AI system, and the *exploitation phase* by which we mean the subsequent use of the generative AI system by a user who gives instructions (prompts) that result in the generation of material on the basis

1 Cf. S.M. Fiil-Flynn and others, 'Legal Reform to Enhance Global Text and Data Mining Research – Outdated Copyright Laws Around the World Hinder Research', *Science* 378 (2022) 951, 951 ; T. Ueno, 'The Flexible Copyright Exception for 'Non-Enjoyment' Purposes Recent Amendment in Japan and Its Implication', *Gewerblicher Rechtsschutz und Urheberrecht International* 70 (2021), 145 (150-151); M.W. Carroll, 'Copyright and the Progress of Science: Why Text and Data Mining Is Lawful', *U.C. Davis Law Review* 53 (2019) 893, 954; C. Geiger, G. Frosio, O. Bulayenko (2019), 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU', *Centre for International Intellectual Property Studies Research Paper* 2019/08, (Strasbourg: CEIPI 201), 5 and 31; T. Margoni and M. Kretschmer, 'A Deeper Look Into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology', *CREATe Working Paper* 2021/7 (Glasgow: CREATe Centre 2021), 10; M.A. Lemley and B. Casey, 'Fair Learning', *Texas Law Review* 99 (2021) 743, 770-771; R.M. Hilty and H. Richter, 'Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules – Part B: Exceptions and Limitations – Art. 3 Text and Data Mining', *Max Planck Institute for Innovation and Competition Research Paper Series* 2017-02, 1.

2 Cf. G. Westkamp, 'Borrowed Plumes: Taking Artists' Interests Seriously in Artificial Intelligence Regulation', 1 (19-26), forthcoming; K. de la Durantaye, 'Nutzung urheberrechtlich geschützter Inhalte zum Training generativer künstlicher Intelligenz – ein Lagebericht', *Archiv* für *Presserecht* 55 (2024), 9 (21-22); M.R.F. Senftleben, 'AI Act and Author Remuneration – A Model for Other Regions?' (2024), 6-23 <https://ssrn.com/abstract=4740268>; C. Geiger, 'Elaborating a Human Rights Friendly Copyright Framework for Generative AI', *International Review for Intellectual Property and Competition Law* 2024, forthcoming, 29-33 <https://ssrn.com/abstract=4634992>; D. Friedmann, 'Creation and Generation Copyright Standards', *NYU Journal of Intellectual Property and Entertainment Law* 14 (2024), forthcoming, 7-8; C. Geiger, 'When the Robots (Try to) Take Over: Of Artificial Intelligence, Authors, Creativity and Copyright Protection', in F. Thouvenin and others (eds), *Innovation – Creation – Markets, Festschrift für Reto M. Hilty* (Berlin: Springer 2024), 67-87; M.R.F. Senftleben, 'Generative AI and Author Remuneration', *International Review of Intellectual Property and Competition Law* 54 (2023) 1535, 1542-1556; G. Frosio, 'Should We Ban Generative AI, Incentivise It or Make It a Medium for Inclusive Creativity?', in E. Bonadio and C. Sganga (eds), *A Research Agenda for EU Copyright Law* (Cheltenham: Edward Elgar 2024), 19-21 <https://ssrn.com/abstract=4527461>; C. Geiger and V. Iaia, 'The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI', *Computer Law and Security Review* 52 (2024), forthcoming, 10-16 <https://ssrn.com/abstract=4594873>.

3 Copyleft licensing was originally developed within the free and open source software movement as an alternative to so-called permissive licensing. See generally P. McCoy Smith, 'Copyright, Contract, and Licensing in Open Source', in: A. Brock (ed.), *Open Source Law, Policy and Practice* (2nd ed., Oxford: Oxford University Press 2022) 83-97.

of those instructions. The terms *development* and *exploitation*, as used in the following discussion, are roughly equivalent to *training* and *inference* as used in technical literature.

B. Copyleft Licences And Their Applicability To Machine Learning

3 When a literary or artistic work is created by a human author, copyright law confers a set of exclusive rights to the creator.[4] The inevitable consequence of this is that use falling within the scope of exclusive rights is only permitted where authorisation for each relevant use exists. Such authorisation may either come from a licence given by the copyright holder or be based on a statutory permission such as a copyright exception or fair use provision.[5] In any other case the use is prohibited because of the exclusive nature of the conferred rights.

4 Within this matrix, CLSA licencing is based on the idea of *relying* on copyright as a mode to promote access to content by making the work available under specified conditions. For this reason, it is essential to determine the manner in which the exclusive rights are exercised in the case of this specific licensing model. Typically CLSA licences distinguish between two types of material: on the one hand, the original material protected by copyright, often denoted as 'licensed material'; on the other hand, derivative material, denoted in CC licences as 'adapted material', created by the licensee and derived from, or based on, the original, licenced material.[6] The two concepts are not mutually exclusive but merely denote, from the perspective of the rightholder (licensor) whether the licensed material has undergone further modifications.

5 CLSA licences rely on copyright, essentially, to achieve two central goals. First, the licence describes uses that are permitted, often in broad terms. By way of example, the CC BY-SA 4.0 licence enables the recipient to reproduce and share the licensed material, in whole or in part, and to produce, reproduce, and share adapted material.[7] Moreover, the recipient is authorised to exercise the permissions in all media and formats (known and unknown) and make necessary technical modifications.[8] These use permissions depend on compliance with further conditions that are imposed on the licensee to ensure that a subsequent recipient downstream can enjoy similarly broad permissions. In this vein, the recipient of licensed material may be prevented from offering or imposing additional or different licensing terms, and applying technological protection measures.[9] The CC BY-SA 4.0 licence also clarifies that a subsequent downstream recipient of the material receives an automatic offer setting forth the same licence conditions (including in the licensee's later licence, denoted as 'adapter's licence' – at least to the extent to which the licence relates to material over which the original licensor has rights).[10] Depending on the needs of the licensor, a licence may also restrict commercial use.[11]

6 Second, CLSA licences introduce a set of requirements on which the operability of the granted use permissions depends. That is, failure to comply renders the granted permissions inapplicable. For example, where recipients share the licenced material, as is expressly permitted, they may be required to retain copyright information supplied with the licensed material (attribution) or indicate that they have modified the material or retain an already existing indication of previous modifications, or indicate that the licensed material is licensed under a specific CLSA licence and include the text of, or a reference to, the licence.[12] Moreover, and most

4 For instance, see ISD, Articles 2 to 4.

5 CDSMD, Articles 3 and 4. As to the US fair use system, see P. Samuelson, 'Fair Use Defenses in Disruptive Technology Cases', UCLA Law Review 72 (2024), forthcoming <https://ssrn.com/abstract=4631726>; M. Sag, 'The New Legal Landscape for Text Mining and Machine Learning', *Journal of the Copyright Society of the USA* 66 (2019), 291.

6 As another example of a CL licence that has wide application, the Free Art Licence 1.3 refers to 'subsequent works'.

7 CC BY-SA 4.0 Section 2(a)(1).

8 CC BY-SA 4.0 Section 2(a)(4)

9 CC BY-SA 4.0 Section 2(a)(5)(c).

10 CC BY-SA 4.0 Section 2(a)(5)(a)-(b).

11 For example CC BY-NC-SA 4.0 Section 2(a)(1).

12 Such as in the case of CC BY-SA 4.0 and CC BY-NC-SA 4.0,

importantly, for adapted material produced by the licensee, the licence may require that the adapted version be made available on the same terms (SA condition). For example, the CC BY-SA 4.0 licence requires the licensee to apply the same licence, with the same conditions, or a licence that is equivalent with the granted licence.[13]

7 With such a setting in mind, CLSA licensing essentially sets in motion a cascade of contractual arrangements that ensure, and maintain, open access to the protected material. Recipients are free to use the original, licensed material as long as they observe the specific requirements set forth in the licence. If they create adapted material, they must make it available under the same terms. The model, simply stated, passes on the CLSA obligation to every user of the material. This contractual mechanism works because copyright protection of the original, licensed material, as well as those portions of adapted material that are derived from the licensed material, will prevent uses outside of the licence. In other words, copyright protection of the licensed material serves as a basis for granting the permissions and enforcing the conditions that establish the SA scheme. A licensee who does not observe CLSA obligations steps outside of the use permission following from the licence and, thus, acts without rightholder authorisation. As a result, downstream use of adapted material that neglects CLSA obligations amounts to infringement of copyright in the original material offered under CLSA terms.[14] The use of CLSA material for generative AI development, thus ultimately boils down to the question whether, and if so where exactly, a ML workflow using CLSA training resources involves copyright-relevant acts that may trigger an obligation to observe the CLSA conditions accompanying the materials offered under CLSA terms.

C. Machine Learning And Copyright-Relevant Acts Of Use

8 In the development phase, the ML workflow – namely the training of foundation models – typically requires accumulating vast amounts of multi-modal data.[15] In the case of foundation models relating to literary or artistic expression, copyright-protected source material will serve as 'data' input for training purposes.[16] As this data collection and use may involve copying of individual expression enjoying protection, the question arises whether the training process falls within the scope of the reproduction right granted in copyright law (following sections C.I and C.II).[17] The supply of curated datasets also raises the question whether the right of communication and making available to the public may play a role (section C.III).

I. Right of Reproduction

9 Article 9(1) of the Berne Convention for the Protection of Literary and Artistic Works (BC) confirms that authors enjoy the exclusive right to authorise the reproduction of their works in any manner or form.[18] The right has also been

which make this clear in Section 3(a).

13 CC BY-SA Section 3(b).

14 Although there seems to be a view that non-conformity with a licence should "merely" be treated as breach of contract (for which the default statutory remedies are normally weaker than in case of infringement, or as agreed in the contract), the CJEU has made it clear in Case C666/18 *IT Development SAS v Free Mobile SAS*, that remedies and sanctions available to rightholders through the Enforcement Directive must be available also in case of breach of a copyright licence agreement. Inevitably that presumes that infringement of copyright has taken place..

15 R. Bommasani et al., *On the Opportunities and Risks of Foundation Models*, Centre for Research on Foundation Models (CRFM) (Stanford Institute for Human-Centred Artificial Intelligence (HAI), Stanford: Stanford University 2021), 146 <https://crfm.stanford.edu/assets/report.pdf> .

16 As to the distinction between use of literary and artistic works as 'works' and use of works as 'data', see R. Ducato/A. Strowel, 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out', *European Intellectual Property Review* 43 (2021), 322 (334). Cf. also M. Borghi/S. Karapapa (2011), 'Non-display Uses of Copyright Works: Google Books and Beyond', *Queen Mary Journal of Intellectual Property* 1 (2011), 21 (44-45).

17 For a more detailed discussion of this conceptual issue, see M.R.F. Senftleben, 'Compliance of National TDM Rules with International Copyright Law – An Overrated Nonissue?', *International Review of Intellectual Property and Competition Law* 53 (2022), 1477 (1493-1502).

18 T. Dreier, 'Berne Convention' in Dreier T and Hugenholtz

incorporated in the so-called WIPO 'Internet' Treaties of 1996 (the WIPO Copyright Treaty (WCT) and the WIPO Performances and Phonograms Treaty (WPPT)) which aimed to adapt copyright law to the digital environment.[19] As the scope of the reproduction right in the digital environment was a highly contentious issue during negotiations, particularly in respect of temporary copying, such as in the operating memory of computers, the WCT does not contain a self-standing right of reproduction but instead incorporates it from the Berne Convention[20] and includes an Agreed Statement indicating that the right fully applies in the digital environment and that storage of a work in digital form in an electronic medium constitutes a reproduction.[21] However, an Agreed Statement does not have the status of an adopted treaty article.[22] The interpretative value of the Agreed Statement addressing the right of reproduction in the WCT is further reduced by the fact that the sentence referring to storage was not adopted unanimously and fails to provide an agreed definition of storage,[23] thus leaving the scope of the reproduction right in the digital environment open and prone to 'highly variable interpretation' as far as temporary copying goes.[24] Consequently, international copyright law has been said to leave open the question of temporary reproduction.[25] With regard to AI development, it is important to note that the right of reproduction granted at the international level need not be understood to cover TDM for ML training purposes.[26] The applicability of the reproduction right depends on the individual national or regional transposition of the applicable international rules into domestic law.

10 In the EU, the question was settled through the adoption of the 2001 Directive on Copyright in the Information Society (ISD),[27] which introduced in its Article 2 a comprehensive reproduction right that covers everything from permanent to temporary reproduction, in whole or in part, in any form and by any means, covering both the reproduction of works as well as subject-matter protected by neighbouring rights.[28] Accordingly, at least in the EU copyright

B (eds), *Concise European Copyright Law* (Kluwer Law International 2016), 45; S. Depreeuw, *The Variable Scope of the Exclusive Economic Rights in Copyright* (Kluwer Law International 2014), 65. In case of neighbouring rights this is Article 3(e) Rome Convention, which laconically explains that a reproduction involves the making of a copy or copies of a fixation. The fact that it must be a copy of a fixation follows naturally from the category of subject matter – recordings of sound (phonograms), of performances, or of broadcasts. In case of the Berne Convention (works), commentators note that the language of the Convention is absent a fixation requirement; Z. Efroni, *Access-Right: The Future of Digital Copyright Law* (Oxford University Press 2010), 220. S. Ricketson and J. Ginsburg, *International Copyright and Neighbouring Rights: The Berne Convention and Beyond* (vol I, 2nd ed, Oxford University Press 2006) 645 observe that it is 'open to debate whether the Berne Convention also requires member states to interpret 'any manner or form' to extend to transient digital fixations'. At least in terms of subsistence of protection, the Berne Convention introduces a discretionary possibility to require fixation of the work in Article 2(2).

19 WCT and WPPT, Preamble.

20 WCT, Art. 1(4).

21 WCT, Agreed Statement concerning Article 1(4). An identical statement is present in WPPT. See Agreed Statement concerning Articles 7, 11 and 16. Cf. M.R.F. Senftleben, (n 17).

22 J. Reinbothe and S. von Lewinski, *The WIPO Treaties on Copyright: A Commentary on the WCT, the WPPT, and the BTAP* (2nd ed., Oxford: Oxford University Press 2015) 66.

23 As to the specific circumstances surrounding the adoption of the Agreed Statement, see furthermore M.R.F. Senftleben, 'Compliance of National TDM Rules with International Copyright Law – An Overrated Nonissue?', *International Review of Intellectual Property and Competition Law* 53 (2022), 1477 (1489 and 1500-1501).

24 Ricketson and Ginsburg (n 18), 687; also JAL Sterling and P. Johnson , 'WIPO Copyright Treaty (1996)' in T. Cook (ed), *Sterling on World Copyright Law* (4th edn, Sweet & Maxwell 2015), 929 noting the question is open.

25 M.R.F. Senftleben, 'WIPO Copyright Treaty', in: T. Dreier and P.B. Hugenholtz (n 18) 99. The same is understood to hold true in respect of the WPPT. See F. Brison, 'WIPO Performances and Phonograms Treaty', id., 201-205.

26 M.R.F. Senftleben (n 17).

27 Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001, on the harmonisation of certain aspects of copyright and related rights in the information society, *Official Journal of the European Communities* 2001 L 167, 10. As to an earlier recognition in case of specifically software, see Article 2 of the Software Directive.

28 In case of neighbouring rights, however, this only concerns fixations of performances, first fixations of films, phonograms (sounds recordings), and fixations of

acquis, it is settled that the exclusive rights of copyright and neighbouring right holders cover virtually any form of copying protected content. Whilst certain types of copying *may* be subject to a copyright exception, the fact remains that, as a starting point, the broad right of reproduction granted in the EU covers acts of copying in the digital environment.

11 Considering this scope and reach of the reproduction right in the EU, we consider that the AI development phase is likely to involve, as a default, reproductions within the meaning of EU copyright law, thus rendering CLSA licences relevant.[29] In this vein, Recital 105 of the AI Act[30] confirms that the use of literary and artistic works for AI training purposes has copyright relevance[31] and involves TDM activities that require the authorisation of rightholders in the absence of a copyright exception: '[a]ny use of copyright protected content requires the authorisation of the rightholder concerned unless relevant copyright exceptions and limitations apply.'

II. Impact on Machine Learning

12 The analysis of ML processes based on the EU position requires a closer look at the individual training and development steps as not every ML stage involves the making of copies. For the sake of simplicity we think of the development phase as involving five stages, consisting of (1) data corpus compilation (2) data corpus preprocessing, (3) data corpus annotation, (4) training of the model, and (5) permanent creation of an artefact (trained model).[32] The initial stages concerning the creation of a training dataset (stages 1 and 2) which involve data collection, for example through web scraping, and conversion of the data into desirable formats, involve reproductions, be it downloading and storage, or reproductions in the operating memory of the system. For a reproduction to take place in computer systems, human cognition (perception of the work) is not necessary. That is why storage of protected material on non-volatile storage media, such as a flash drive, or more fluidly in volatile memory, such as the operating memory of a computer, amounts to a reproduction in the sense of copyright law. Because of the breadth of the reproduction right granted in the EU, the individual acts carried out during the ML process are likely to amount to independent, separate, acts of reproduction determined by the particular needs of the entire process.[33] In other words, the act of transferring data to the operating memory of the system for the purpose of conversion does not, as such, remove the copyright relevance of the reproduction carried out previously to store a copy. If that already stored copy is later deleted from the storage resource, it raises the question whether the conversion process can be regarded as a permissible form of transient copying in the sense of the copyright exception laid down in Article 5(1) ISD.[34] Needless to say, any back-up copies created as a result of security diligence also amount to separate reproductions. The third stage, essentially, involves data labelling and is essential for supervised learning, while the fourth constitutes the actual training phase involving computational analysis, correction and validation.[35] In these instances the

broadcasts. Protection of non-original photographs remains a matter for national legislation in the Member States.

29 However, see also the analysis by R. Ducato and A. Strowel (n 16) 334, who propose to distinguish between use of copyrighted material 'as a work' and use of copyrighted material as mere data – with the result that use as mere data may fall outside the scope of the right of reproduction.

30 This numbering refers to the text of the AI Act, as adopted by the European Parliament on 6 March 2024.

31 As to the discussion about the applicability of Articles 3 and 4 CDSMD to the training of generative AI models, see M.R.F. Senftleben (n 2), 7-14; F. Hoffmann, 'Zehn Thesen zu Künstlicher Intelligenz (KI) und Urheberrecht', *Wettbewerb in Recht und Praxis* 2024, 11 (16-18).

32 See generally T. Margoni, 'Artificial Intelligence, Machine Learning and EU Copyright Law: Who Owns AI?', *CREATe Working Paper* 2018/12 <https://www.create.ac.uk/artificial-intelligence-machine-learning-and-eu-copyright-law-who-owns-ai/>.

33 Cf. Recital 105 AIA and M.R.F. Senftleben (n 2) 7-14.

34 Even if it has, the subsequent creation of a converted copy will amount to a new reproduction.

35 See generally on the training stage J.-M. Deltorn, 'The elusive intellectual property protection of trained machine learning models: A European perspective', in: R. Abbott (ed.), *Research Handbook on Intellectual Property and Artificial Intelligence* (Cheltenham: Edward Elgar 2022) 87.

same principle applies: *if* labelling and training involve copying of copyright-protected data, at least in the operating memory, these acts are likely to constitute reproductions within the meaning of the broad right granted in the EU,[36] even though they are merely a means to an end.

13 The broad EU approach to the reproduction right also raises the question whether the right may become relevant beyond the act of, strictly speaking, duplication of copyright-protected data collected during the corpus compilation phase, in particular in relation to stages 2 and 3 (data corpus preprocessing and annotation). Notably, the Court of Justice of the European Union (CJEU) has considered the process of canvas transfer (the removal of ink form a paper poster and its transfer to a canvass) as an act of reproduction because of the change of medium.[37] Although not concerning digital copies, this broad approach raises the question of whether electronic changes to a computer file containing a work that result in adaptation or conversion of the file to a desirable format could similarly involve an act of reproduction, which would be different and separate from the mere act of copying data. While CJEU jurisprudence points in this direction, the Canadian Supreme Court has reached a different conclusion for such acts.[38] Hence, the issue has not yet been settled. If the issue is brought before the CJEU, the Court may refrain from extending the Canvas approach to file conversions for TDM purposes.

14 Whether copyright-relevant acts of reproduction take place during stage five is not as straightforward. Although the applicable copyright principles are easy to explain, the model exists as a separate artefact: normally operating independently from its training pipeline.[39] It does not seem to retrieve the contents of the training dataset when generating outputs during the exploitation phase. Hence, it can be argued that the artefact exists and operates independently from the copyright-protected data, including 'licensed material' triggering CLSA obligations, that have been used as training resources in the preceding steps one to four. Following this line of argument, the artefact can be described as a giant collection of data points and vectors that have been derived from the training material.[40] It can also be assumed that the artefact is unlikely to contain copyright-protected traces of works that were used for training.[41] The adoption of this perspective leads to the conclusion that the creation of the trained model at stage five breaks the link with CLSA licensing obligations that may rest on training resources. If the artefact as such does not contain copyright-protected traces of CLSA works used for training purposes, copyright law does not offer tools for enforcing CLSA conditions: relevant acts of reproduction are sought in vain.

15 As so often in the legal debate, however, nuance is important. In the CJEU's jurisprudence, in particular the case law established in *Infopaq* and *Pelham*,[42] confirms that for assuming a relevant act of reproduction it would be sufficient that a fragment of a work is included in the artefact. In the case of copyright, this fragment would have to satisfy the originality test of free, creative choices (a text

36 ISD, Article 2.

37 Case C-419/13 *Art & Allposters v Stichting Pictoright*, para 43.

38 Compare *Théberge v. Galerie d'Art du Petit Champlain Inc* [2002] 2 S.C.R.

39 J.-M. Deltorn, (n 35) p. 88.

40 *See* for similar reasoning by American scholars P. Samuelson, C.J. Sprigman, M. Sag, *Comments in Response to the Copyright Office's notice of Inquiry on Artificial Intelligence and Copyright* (30 October 2023), 7-8 <https://www.regulations.gov/comment/COLC-2023-0006-8854>.

41 As discussed in more detail below, it cannot be ruled out that AI models memorise certain aspects of training data. Cf. I. Emanuilov and T. Margoni, 'Forget Me Not: Memorisation in Generative Sequence Models Trained on Open Source Licensed Code' <https://ssrn.com/abstract=4720990>>, 10-15; S. Biderman and others , 'Emergent and Predictable Memorization in Large Language Models' (*arXiv*, 31 May 2023) <https://arxiv.org/abs/2304.11158>; X. Gu and others, 'On Memorization in Diffusion Models' (*arXiv*, 4 October 2023) <https://arxiv.org/abs/2310.02664>. However, the central question from a copyright perspective is whether these memorised aspects contain protected traces of copyright-protected works or other protected subject matter, such as sound recordings. As discussed below, it seems to us that, at least in the majority of cases, it cannot generally be assumed that protected elements of CLSA material will be memorised and become part of trained models.

42 Case C-5/08, *Infopaq v DDF*, paras 38-39; Case C-476/17 *Pelham v Hütter and Schneider-Esleben*.

extract of 11 words may be sufficient).[43] This is why we referred to 'copyright-protected traces' above. In the case of neighbouring rights, the test for assuming protection may be even more relaxed. With regard to sound recordings, for instance, the CJEU has confirmed that the reproduction right of phonogram producers covers sound extracts 'even if very short' (a sound sample of 2 seconds may be sufficient).[44] These nuances might prove to be relevant in cases where memorisation, or overfitting or parroting, of data from the training dataset might actually take place.[45] If, therefore, protected fragments of a work or subject matter enjoying neighbouring rights protection are contained in the stage five artefact, a relevant act of reproduction takes place and the equation is different. Here the CJEU's judgment in *Allposters* mentioned above may prove to offer an additional relevant argument, if it is to be read as implying that the potentially different technological representation of such a fragment in the stage five artefact, compared to its representation in the training dataset, is indeed to be captured by the European concept of reproduction. However, whilst the decisions in *Infopaq* and *Pelham* confirm that an infringing exploitation can already be assumed in the case of text excerpts that are as short as 11 words, or extracts from sound recordings that are as short as 2 seconds, the copyright assessment is not quantitative but qualitative and therefore case-specific. In the case of works, the used fragment must be original and contain free, creative choices of the original work.[46] In the case of neighbouring rights, the CJEU has also developed additional criteria. Fragments taken from a protected sound recording, for instance, no longer amount to infringement if they are used in a derivative phonogram 'in a modified form unrecognisable to the ear.'[47]

16 Factoring this important nuance into the equation, it nevertheless seems to us that, at least in the majority of cases, we can uphold the above conclusion: with the creation of the stage five artefact, the link with CLSA obligations is broken and copyright is no longer available as a tool to enforce CLSA conditions. In practice, it will also be difficult to prove that protected traces of works or other subject matter made their way into the trained model, especially absent access to the training data for comparison. How can we provide evidence that free, creative choices of a human author have been woven into the fabric of the final artefact? How can we prove that sound snippets in the trained model are recognisable to a human ear?

17 These practical considerations need not always thwart copyright claims. Ultimately, copyright is a property right and the duty of care to ensure compliance lies not with the rightholder but the developer or adopter of the model. In an infringement case, the judge may reverse the burden of proof and impose the obligation on the artefact developer or adopter to show that the trained model does not contain copyright-protected traces of CLSA works. In the case of iconic works that a web crawler looking for training material is likely to collect very often, such as a famous quote[48] or drawings of famous fictional characters, the AI developer may even find it particularly difficult to provide this proof.[49]

43 Case C-5/08 *Infopaq v DDF*, paras 38-39.

44 Case C-476/17 *Pelham v Hütter and Schneider-Esleben*, para 29.

45 See generally D.J. Gervais and others, 'The Heart of the Matter: Copyright, AI Training, and LLMs' (2024), 11 https://ssrn.com/abstract=4963711>; I. Emanuilov and T. Margoni (n 41) pp. 10-15; N. Carlini and others, 'Extracting Training Data from Large Language Models', in: *30th USENIX Security Symposium (USENIX Security 21)* (USENIX Association 2021), 2633-2650 <https://www.usenix.org/system/files/sec21-carlini-extracting.pdf>; N. Carlini and others, 'Quantifying Memorization Across Neural Language Models' (*arXiv*, 6 March 2023) <https://arxiv.org/abs/2202.07646v3>; S.A. Taghanaki and J. Lambourne, 'Detecting Generative Parroting through Overfitting Masked Autoencoders' (*arXiv*, 19 June 2024) <https://arxiv.org/html/2403.19050v3>.

46 Case C-5/08, *Infopaq v DDF*, para 51. See also Joined Cases C-403/08 and C-429/08 *FAPL*, para 159; Case C-406/10 *SAS Institute v World Programming*, para 70.

47 Case C-476/17 *Pelham v Hütter and Schneider-Esleben*, paras 29-31 and 39.

48 For an example concerning the beginning of a chapter of J.K. Rowling's *Harry Potter and the Philosopher's Stone*, see I. Emanuilov and T. Margoni (n 41) p. 15.

49 I. Emanuilov and T. Margoni (n 41) p. 26. Cf U. Hacohen and N. Elkin-Koren, 'Copyright Regenerated: Harnessing GenAI to Measure Originality and Copyright Scope', Harvard Journal of Law and Technology (2024) 37 <https://ssrn.com/abstract=4530717>; U. Hacohen and others, 'Not All Similarities Are Created Equal: Leveraging Data-Driven

If the system has somehow stored all the information necessary to identify and reproduce a cat or dog, why should the system have refrained from doing the same with regard to Mickey Mouse, Spiderman, Lucky Luke and Nijntje?

18 However, even if we could assume that there is a statistical probability of copyrighted traces of iconic CLSA works finding their way into the trained model, we believe that this statistical probability of CLSA facets in the artefact is not a sufficiently solid basis for *routinely* enforcing SA conditions in AI development contexts, as the legal discussion is currently in its infancy and there is no series of court decisions providing established case law. Given the legal uncertainty surrounding copyright claims based on training material memorisation, it is important to explore alternative, potentially more robust solutions. To bring these alternative solutions to light, we focus on the above-described assumption that the trained model only contains unprotected data points and vectors which, in turn, leads to the conclusion that, in the majority of cases, the link with CLSA licensing obligations is broken.

19 Ascertaining the copyright status of the stage five artefact may also raise a challenge that goes to the core of the reproduction right and beyond the legal-technical questions of training material memorisation and the burden of proof. If we conceive of the model as having a capacity to evoke the image of an existing work (or parts thereof) following training, rather than a capacity to retrieve it from a repository of stacked copies (or fragments thereof) that are algorithmically selected and modified following a prompt, the manner in which the model operates may be more similar to how a human being is capable of imagining an object. If this is the feature of the stage five artifact, it may be difficult – if not impossible – to qualify the creation of the stage five artefact as involving the reproduction right from the outset.

Biases to Inform GenAI Copyright Disputes' (*arXiv*, 7 May 2024) <https://arxiv.org/abs/2403.17691>; M. Sag, 'Copyright Safety for Generative AI' Houston Law Review 61 (2023) 295, 321-337; A. Guadamuz, 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs', Gewerblicher Rechtsschutz und Urheberrecht International 73 (2024) 111, 121-122.

III. Right of Communication and Making Available To The Public

20 While the right of reproduction is certainly the centre of gravity in the debate on generative AI systems and CLSA conditions, we must not overlook interactions that may take place on the market for ML technology. Adding this broader context, other exclusive rights than the right of reproduction may also become relevant at the development phase, namely the 'right of communication' to the public, and particularly the 'making available' prong of the right.[50] In particular, offers to the public to obtain curated training datasets that include copies of protected content, whether annotated or not, may amount to an act of making available to the public in the sense of EU copyright law. This is ultimately a jurisdictional issue as copyright protection is limited by the principle of territoriality, but at least in the case of the EU the matter seems to be settled. Considering CJEU jurisprudence on the right of communication to the public granted in Article 3 ISD,[51] it cannot be ruled out that such an offer would involve copyright law and amount to an act of communication to the public/making available to the public that requires the authorisation of the rightholder. Accordingly, it would activate the obligations following from SA conditions in cases where CLSA knowledge resources are used to build a curated dataset. The offer and distribution of such a dataset would require compliance with CLSA terms. In the case of CC BY-SA 4.0, the use will also be governed by the prohibition to offer or impose additional or different terms than provided under that CLSA licence in respect of 'licenced material' (to which the licensor has exclusive rights).[52] It is also noteworthy that whilst the use permissions granted by copyleft licences are broad, they may also be limited to non-commercial use (CC BY-NC-SA 4.0).

50 ISD, Article 3.

51 Case C-263/18 *NU and GAU v Tom Kabinet*, establishing that the offer to buy an e-book (that could be purchased by one person only) amounts to a communication to the public. In the EU the same applies in case of offers of products that are not delivered online, see Case C-516/13 *Dimensione Direct Sales and Labianca v Knoll Internationall*, para 28-32; Case C-5/11 *Donner*, para 30.

52 CC BY-SA 4.0, Section 2(a)(5)(c).

Such a licence might prevent the sharing of material if it is done for a commercial purpose.

D. Copyright Exceptions Covering Machine Learning In The EU

21 An important aspect of CLSA licences is the manner in which they address the relationship to copyright exceptions. Certain copyleft licensing schemes explicitly give precedence. For example, the CC BY-SA 4.0 license states the following in Section 2(a)(2):

> *For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions.*

22 This clause, essentially, makes it necessary to identify uses permitted under relevant 'exceptions and limitations' (collectively referred to as 'copyright exceptions' or 'exceptions' in the following analysis) in a given copyright regime. To the extent to which ML workflows and related uses fall within the scope of exceptions in the EU, these statutory use permissions prevail and render the CLSA conditions inapplicable. Concomitantly, uses that are not covered by a copyright exception continue to instead be governed by the terms of the licence. For this reason, it is essential to determine the scope of copyright exceptions that can apply to ML workflows. Where legislators have introduced provisions that have the potential of covering the entire ML process, such as the TDM provisions in the EU, it is crucial to determine the impact of those provisions as the precedence given to copyright exceptions in copyleft licences is likely to affect the continued viability of CLSA terms.

23 The catalogue of exceptions in Article 5 ISD is quite diverse. In respect of ML processes, it is noteworthy that it includes the possibility to carry out temporary reproductions under certain further conditions (Article 5(1) ISD). With the adoption of the 2019 Directive on Copyright in the Digital Single Market (CDSMD),[53] the EU legislator has introduced two additional provisions that have given the debate an entirely new edge. Conditioned on lawful access to the material used for ML purposes, Articles 3 and 4 CDSMD provide for exceptions to the right of reproduction that enable TDM, which Article 2(2) CDSMD defines broadly as an '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'. With this broad definition, imposing no restriction on the type of information that should be generated, the provisions are apt candidates for covering various ML uses, and are indeed considered as such.[54] The European legislature has recently affirmed the relevance of the TDM provisions for the development and training of generative models in Recital 105 of the AI Act. Most important for our purposes is the exception in Article 4 CDSMD because it is not subject to a general purpose limitation but applies to any actor or purpose for which TDM is carried out, including commercial TDM projects. Article 3 CDSMD, by contrast, imposes both a purpose limitation and a beneficiary limitation: it applies only to research organisations[55] and cultural heritage institutions[56] and covers only TDM for the purpose of scientific research.[57] To complete the overview of copyright exceptions that play a role in

53 Directive 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, *Official Journal of the European Communities* 2019 L 130, 92.

54 T. Chiou, 'Copyright lessons on Machine Learning: what impact on algorithmic art?', *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 9 (2019), 398 (409).

55 Defined in CDSMD, Article 2(1), as 'a university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research: (a) on a not-for-profit basis or by reinvesting all the profits in its scientific research; or (b) pursuant to a public interest mission recognised by a Member State'.

56 Defined in CDSMD, Article 2(3), as 'a publicly accessible library or museum, an archive or a film or audio heritage institution'.

57 See further K. Szkalej, 'The paradox of lawful text and data mining? Some experiences from the research sector and where we (should) go from here' (2024), forthcoming <https://ssrn.com/abstract=5000116>.

ML contexts, we also address, at the end of this part, the aforementioned temporary copying exception provided under Article 5(1) ISD.

I. TDM Provisions

24 As explained, copyleft licences let statutory copyright exceptions prevail over the licencing terms. Against this background, the introduction of TDM exceptions in the CDSM Directive raises the question of whether it still makes sense to deploy CLSA licences as a mode to regulate downstream use. To the extent to which the TDM exceptions cover ML processes, they prevail over the SA condition and render it inapplicable under the current configuration of the relationship between CC licenses and copyright exceptions. Nonetheless, we believe that SA conditions can still play an important role. To explain this point, we must take a closer look at the TDM exceptions in EU copyright law.

1. Output Not Covered

25 First, the TDM exceptions laid down in Articles 3 and 4 CDSMD only concern the TDM *process* of collecting and analysing copyright-protected data to generate information relevant for creating a ML tool or foundational model.[58] Articles 3 and 4 CDSMD do not cover the reproduction of copyright-protected features in literary and artistic content which the fully trained AI model generates in the end. It is an entirely different question of who might be liable under copyright law in the event that such output could be deemed to infringe copyright in a pre-existing work. We return to this issue in part 5, highlighting here only the aspect that the applicability of a copyright exception covering TDM does not, as such, automatically render CLSA licence clauses inapplicable with regard to AI output even though a copyleft licence scheme such as CC BY-SA 4.0 states explicitly that copyright exceptions prevail. Instead, the precedence given to copyright exceptions only concerns the exempted form of use falling within the scope of the exception, namely the ML process leading to the establishment of the generative AI model in the case of the TDM provisions in Articles 3 and 4 CDSMD. Any other use, such as the subsequent content generation based on a user prompt, could in principle be regulated by the CLSA licence terms, to the extent that it involves copyright-relevant acts requiring the authorisation of the CLSA licensor.

2. Opt-out Mechanism

26 Second, whereas Article 3 CDSMD is mandatory by law and cannot be contracted out,[59] in case of Article 4 CDSMD, TDM can be carried out only if the rightholder has not reserved the use of the protected material in an appropriate manner. With this opt-out mechanism, Article 4(3) CDSMD, rather than staying silent on contractual overridability, affords rightholders the opportunity to determine whether they wish to make their works available for TDM. In other words, Article 4 CDSMD is merely a conditional exception. Once the rightholder has opted out in accordance with Article 4(3) CDSMD, the use privilege no longer applies:

> *The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.*[60]

27 With regard to CLSA licence terms, this means that rightholders can use the opt-out mechanism in Article 4(3) CDSMD when they wish to minimise the impact of the copyright exception in Article 4 CDSMD and maximise the scope of copyright as a basis for enforcing CLSA conditions. As a result of the opt-out, the use of the original material for TDM purposes requires authorisation unless it falls within the scope of the mandatory exemption of scientific TDM in Article 3 CDSMD. Hence, the rightholder has the opportunity to impose CLSA licensing terms and make the use dependent on compliance with these terms. This exercise of the opt-out possibility, admittedly, may give rise to a dilemma in the light of

58 See also the definition in CDSMD, Article 2(2).

59 See however K. Szkalej (n 57), 11.

60 CDSMD, Article 4(3).

the current configuration of CLSA licensing regimes: the opt out shuts down the exception. Current CLSA licensing schemes, however, take as a starting point that copyright exceptions ought to remain intact in order not to curtail user rights following from statutory use permissions.

28 Against this background, the crucial question is whether, from the perspective of the CLSA licensing approach, it can be deemed legitimate to use the opt-out mechanism in Article 4(3) CDSMD and curtail the TDM freedom following from Article 4(1) CDSMD for the purpose of imposing CLSA conditions. From the perspective of EU copyright law, a rightholder availing itself of the opt-out possibility is exercising a prerogative and limitation of the TDM freedom that is inherent in the copyright exception itself. From this perspective, it does not seem inconsistent to restrict TDM falling under Article 4 CDSMD in order to create the possibility of granting and enforcing a tailor-made CLSA licence (that may be broad and allow TDM as long as the SA condition is observed). The opt-out mechanism thus appears as an efficient tool to expand CLSA culture to the realm of AI-generated literary and artistic output.[61]

II. Temporary Copying

29 As already explained above, the EU has opted for the introduction of a broad, comprehensive right of reproduction in Article 2 ISD – a right of reproduction that applies regardless of whether the act of copyright is 'temporary or permanent'. As a counterbalance to this comprehensive exclusive right, the EU copyright system prescribes a mandatory exception that enables temporary copying in Article 5(1) ISD. The provision permits temporary reproductions, which are transient or incidental, and form an integral and essential part of a technological process, and the sole purpose of which is to enable lawful use of the content,[62] on condition that it does not have independent economic significance.

30 Although this temporary copying rule only applies on several further conditions – ranging from the transient nature of the reproduction to the absence of independent economic significance – it nevertheless can cover ML activities taking place during the development phase leading to a generative AI model. Importantly, the adoption of specific TDM exceptions has not made Article 5(1) ISD obsolete. Instead, Articles 3 and 4 CDSMD coexist with the temporary copying rule in Article 5(1) ISD.[63] All these copyright exceptions thus offer statutory use permissions for ML reproductions falling within their specific fields of application.

31 As to the specific scope of Article 5(1) ISD, it must be pointed out that the temporary copying rule is quite a complex provision with five central requirements that must be satisfied cumulatively in order to benefit from the use privilege.[64] As regards the first condition, the existence of a 'temporary' reproduction can be assumed, for example, when the copies are immediately deleted or replaced automatically.[65] A reproduction can be deemed 'transient' when the conservation period of copies is limited to the time necessary for the technical process of making the reproduction and the copies are automatically erased after completion of the process.[66] A reproduction is 'incidental' where it is not self-contained with respect to the technical process of which it forms part. Thus, copies resulting from temporary reproductions should have no purpose that is separate from the one for which they have been made in the framework of ML.[67]

32 These conceptual contours indicate clearly that Article 5(1) ISD only offers limited possibilities in ML

61 See also A. Lazarova and others, *Creative Commons Statement on the Opt-Out Exception Regime / Rights Reservation Regime for Text and Data Mining under Article 4 of the EU Directive on Copyright in the Digital Single Market* (Creative Commons 2021) <https://creativecommons.org/wp-content/uploads/2021/12/CC-Statement-on-the-TDM-Exception-Art-4-DSM-Final-updated.pdf>.

62 ...or a transmission in a network between third parties by an intermediary.

63 CDSMD, recital 9.

64 Case C-5/08 *Infopaq v DDF*, para 55.

65 Case C-360/13 *PRCA v NLA and Others (Meltwater)*, para 26.

66 Id., para 40.

67 Id., para 43.

contexts.[68] As copies based on Article 5(1) ISD cannot be retained for a longer period, the provision does not permit the creation of source data repositories. The transient nature of the copies excludes reuse from the outset.

33 Nonetheless, Article 5(1) ISD may play a role when ML concerns online sources that can be analysed directly and processed in the format in which they are available on webpages.[69] For a computational analyses based on web scraping, the requirements of a temporary and transient nature need not constitute insurmountable hurdles. The invocation of the use privilege in connection with ML also seems in line with the general objectives underlying the provision.[70] The CJEU has recognised that, in order to protect the effectiveness of the temporary copying rule and safeguard its purpose, Article 5(1) ISD must be understood to allow the development and operation of new technologies and ensure a fair balance between the rights and interests of rightholders and those of users.[71]

34 Against this backdrop, it seems consistent to assume that, as long as the individual requirements of the provision are fulfilled, AI trainers can belong to the circle of users who can benefit from Article 5(1) ISD in the context of ML. As CLSA licensing terms allow copyright exceptions to prevail over contractual SA conditions, this means that, to the extent to which the temporary copying rule covers reproductions carried out for ML purposes, SA obligations are rendered inapplicable.

E. Generative AI and The Concept Of 'Adapted Materials'

35 The concept of 'adapted materials' is an essential component of CLSA clauses with particular importance to downstream use. As it may be relevant to both the development phase and the exploitation phase, we treat the two phases separately in our analysis. However, it is useful to first define the term as it gives us an idea of the type of material we are dealing with. For the purpose of our analysis, we rely on the definition of 'adapted materials' in the CC BY-SA 4.0 licence, which defines the term as:

> *material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed relation with a moving image.*[72]

36 In light of this definition, it seems safe to assume that the term covers in any case material that (1) is protected by copyright; and (2) is derived from or based on licenced material (which too is protected by copyright). Importantly, the material has been modified *in a manner requiring permission* from the licensor. Seen from the perspective of the rightholder (the CLSA licensor), licence clauses that concern adapted material continue to operate in the sphere of copyright law, i.e., as explained above, the exclusive rights granted in copyright law serve as a basis for imposing CLSA obligations and enforcing these obligations. One initial question is nonetheless whether the definition of 'adapted material' is intended to fully align with copyright nomenclature.

68 Cf. C. Geiger, G. Frosio, and O. Bulayenko, 'Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data? Legal Analysis and Policy Recommendations', International Review of Intellectual Property and Competition Law (2018), 814 (821-822); R.M. Hilty and H. Richter, 'Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules – Part B: Exceptions and Limitations – Art. 3 Text and Data Mining', *Max Planck Institute for Innovation and Competition Research Paper Series* 2017-02, 2.

69 M.R.F. Senftleben, *Study on EU Copyright and Related Rights and Access to and Reuse of Data*, European Commission, Directorate-General for Research and Innovation (DG RTD) (Brussels: Publications Office of the European Union 2022) , 27-28 <https://data.europa.eu/doi/10.2777/ 78973>.

70 Cf. T. Margoni and M. Kretschmer, 'A Deeper Look Into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology', *CREATe Working Paper* 2021/7 (Glasgow: CREATe Centre 2021),18-19.

71 Joined Cases C-403/08 and C-429/08 *FAPL*, paras 163-164; Case C-360/13 *PRCA v NLA and Others (Meltwater)*, para 24.

72 CC BY-SA 4.0, Section 1(a).

We assume it need not strictly follow the concept of 'adaptation' in copyright law. As explained in part 1, the term merely seems to denote that the CLSA licensee has no objection against the licensed material undergoing further modifications.

37 In this context, the reference to 'material subject to Copyright and Similar Right' at the beginning of the definition indicates, in our view, that the CLSA licence is intended to cover material in which the original material (licensed material) is shimmering through to such an extent that the licensor can invoke copyright as a means to enforce the CLSA conditions because the adapted material still displays copyright-protected creative choices of the licensor.[73] In this scenario, the CLSA clause imposes obligations on what the CC BY-SA 4.0 licence denotes as 'Adapter's Licence', which is the licence that the licensee provides downstream. On the one hand, this additional aspect of the licensing scheme seems to presume that the licensee/adapter creates material that attracts copyright protection itself – copyright that can be used as a basis for passing on CLSA obligations downstream. On the other hand, considering the entire design of the CC BY-SA 4.0 licence, it is noteworthy that the licensee/adapter *does not* issue a sublicence to the original material. As indicated earlier, under clause 2(a)(5) of CC BY-SA 4.0, it is the original licensor who licenses the rights in the relevant portions of the adapted material:

> Every recipient of Adapted Material from You automatically receives an offer from the Licensor to exercise the Licensed Rights in the Adapted Material under the conditions of the Adapter's License You apply.

38 Arguably, this chain of licences granted by the original licensor offers room for arguing that SA obligations can survive modifications even if these modifications do not attract copyright protection themselves. The current wording of clause 2(a)(5) obscures this argument by referring to 'Every recipient of Adapted Material'. If 'Adapted Material' must be understood to require material which adds sufficient new creative choices to attract copyright protection, it becomes doubtful whether the offer – a licence by the original licensor – also covers cases where modifications of the original material are not eligible for copyright protection.

39 However, this potential doubt can be dispelled. First, the formulation 'material subject to Copyright' at the beginning of the definition of 'adapted materials' need not be understood to introduce a strict requirement of modifications attracting copyright protection themselves. It may simply reflect the fact that, because of takings from the copyrighted material offered under CLSA conditions, the adapted material is subject to the copyright in the original CLSA source. Interestingly, this more flexible interpretation is in line with CJEU jurisprudence. In *Deckmyn*, the CJEU clarified that it could not be inferred from the usual meaning of the term 'parody' in everyday language, that the concept was:

> *subject to the conditions set out by the referring court in its second question, namely: that the parody should display an original character of its own, other than that of displaying noticeable differences with respect to the original parodied work...*[74]

40 With regard to work adaptations in the guise of parody, the Court, thus, explicitly rejected an approach requiring the parodist to add free, creative choices[75] that attract copyright protection coming on top of the protection which the original source material enjoys. Following in the footsteps of *Deckmyn*, the requirement of 'material subject to Copyright' in the definition of 'adapted material' can be deemed satisfied whenever protected features of the original material are still present – regardless of whether the adaptation itself is also eligible for copyright protection. This flexible reading allows us to establish a CLSA licence chain which, under clause 2(a)(5), has its origin in the SA offer made by the licensor of the original, licensed material. As

73 For a discussion of the relatively low threshold for assuming this copyright relevance in EU law, see M.R.F. Senftleben, 'Flexibility Grave – Partial Reproduction Focus and Closed System Fetishism in CJEU, *Pelham*', International Review of Intellectual Property and Competition Law 51 (2020), 751 (751-769).

74 Case C-201/13, *Deckmyn Vrijheidsfonds VZW v Vandersteen and Others*, para 21.

75 Case C-5/08 *Infopaq v DDF*, para 45; Case C-145/10 *Painer v Standard Verlags and Others*, para 89.

long as sufficient copyright-protected features of the original work remain discernible in downstream productions – qualified as 'adapted material' regardless of whether they have fresh, original features of their own – the SA obligation (that can be traced back to the original work and the initial licence granted by the original licensor) remains intact and enforceable.

41 For the purposes of our present inquiry, the essential point is that the definition of 'adapted material' and the outlined licence design determine the extent to which CLSA licensing schemes impact ML processes (development phase, section E.I) and AI-generated output (exploitation phase, section E.II). We now turn to a more detailed analysis of these two dimensions.

I. Input/Development Phase

42 Considering the different stages of ML described above, it is clear that collected material undergoes certain modifications for the purpose of making the ML process possible and more efficient. From the perspective of the licence mechanism, which refers to 'adapted material' in the context of regulating downstream use, the crucial question is whether work results that are obtained during the training process constitute modifications of the original, licensed CLSA material that can be classed 'adapted material' in the sense of the CC definition. As explained, the test is whether protected traces of the original, licensed CLSA material are still present in modifications arising during the training process: protected traces that allow the licensor to rely on copyright as a vehicle to enforce CLSA obligations. As already discussed in section C.II, the final artefact – the trained model – is unlikely to constitute adapted material. Arguably, it is independent from copyright-protected CLSA resources that have been used for training purposes. If the trained model is primarily seen as a giant collection of data points and vectors,[76] it can be assumed that it does not contain copyright-protected traces of works used for training. Following this approach, the model as a whole and its components cannot be regarded as 'adapted material' in the sense of the CC definition and the copyright link with CLSA licensing obligations is broken. Hence, copyright law does not offer tools for enforcing CLSA conditions with regard to the final trained model: in the absence of copyright-protected traces, the model does not have copyright relevance. Neither the creation of the model nor its further distribution amount to copyright infringement if protected features of original CLSA material do not shimmer through.

43 As explained in section C.III, the equation is different in the case of CLSA works that become building blocks of curated datasets. It is conceivable that obligations regarding 'adapted material' in the sense of the CC definition cover curated training datasets that contain sources to which the SA obligation is attached. Subject to our caveat further below relating to technical modifications, the making available of such datasets to the public, which may fall under a separate 'Adapter's Licence',[77] may trigger obligations to comply with SA conditions. This also means that the provider of the curated dataset containing CLSA components would be under an obligation to pass on the SA obligation to recipients (model developers).

44 However, where content originally released under a CLSA licence, such as CC BY-SA, is used to curate a training dataset and this dataset is later offered to external model developers, the provider of the curated dataset would have to ensure compliance with the SA condition of the relevant licence that governed the development of the curated dataset. The inevitable consequence of providing the curated dataset in a manner that contradicts the SA conditions imposed by the licence might, additionally, be that neither the TDM provisions in Articles 3 and 4 CDSMD nor the temporary copying exception in Article 5(1) ISD can be invoked any longer. While the discussion on lawful access requirements in EU copyright law is ongoing,[78] the view might be held that the

76 as to the question of memorisation of copyright-protected traces, see section C.II above.

77 'Adapter's Licence' in the terminology of CC-BY-SA 4.0, as mentioned above.

78 See the broader discussion on lawful access requirements and the problem of circularity: lawful access requirements subjecting copyright exceptions to contractual terms that may erode the freedom of use which the legislator sought to create when introducing the copyright exception in the first place. Cf. T. Margoni, 'Saving Research: Lawful Access

requirements of 'lawful use',[79] 'lawful access'[80] or 'lawfully accessible'[81] set forth in these provisions are not satisfied if CLSA components in the training dataset are used by model developers who do not assume the SA obligation themselves. The making available of the curated dataset in a way that does not pass on the SA obligation to model developers would culminate in use of CLSA resources without authorisation and, therefore, amount to copyright infringement, rendering the source material used for ML purposes unlawful. If the dataset developer does not observe the SA obligation, this lack of compliance is thus likely to prevent the model developer from demonstrating lawful access to the CLSA material which, arguably, is a prerequisite for both the TDM exceptions and the temporary copying exception.

45 Considering the full spectrum of concepts in CC licences, however, it is important to point out that, next to the described approach focusing on the concept of 'adapted material', the CC BY-SA 4.0 offers room for an alternative solution based on the concept of 'technical modification'. The CC BYSA 4.0 makes it clear that in so far as mere technical modifications of licensed material are concerned, making these modifications for purposes that in any event would be covered by the licence (which does provide broad use permissions to reproduce and share material and includes making technical modifications to the material) 'never produces Adapted Material'.[82] In other words, technical modifications constitute, under the typology adopted in the licence, licensed material. Under this alternative approach, the question arises whether potential modifications made to establish a curated dataset can be regarded as 'technical modification' in the sense of the licence. If this question is answered in the affirmative, the clauses in the licence on technical modifications would apply to curated datasets – and not the clauses on adapted material. Importantly, this conclusion need not exclude contractual obligations to observe SA conditions. It only excludes the application of Section 3(b) of the licence which applies to adapted materials. However, Section 3(a) concerns sharing of licensed material. This includes technically modified versions of the material, as addressed here. That material must be shared in a manner that includes copyright information and the terms of the licence etc. Moreover, Section 2(b)(5)(c) prevents downstream restrictions on the licenced material. Combining Section 3(a) and Section 2(b)(5)(c), the conclusion seems inescapable that technically modified versions are automatically subject to a SA condition resting on the licenced material. Hence, even if the concept of 'adapted material' cannot be applied to curated datasets, SA conditions remain relevant because they are attached to technically modified versions of the licensed material.

46 Finally, we must recall that the definition of 'adapted material' requires that the material be *modified in a manner requiring permission*. Therefore, copyright exceptions, especially the new TDM provisions discussed above, enter the picture and reduce the applicability of the SA condition to activities and materials that are not covered by pertinent exceptions. When it is assumed (as we did above), that Article 4 CDSMD has the potential to cover all copyright-relevant acts carried out during the ML training process, the term 'adapted material' thus becomes moot at the development stage unless, as explained above, the CC licensor seeking to introduce CLSA obligations exercises the opt-out possibility available under Article 4(3) CDSMD.

47 If the opt-out mechanism is used, this leads to a reservation of copyright that offers far-reaching possibilities for preserving the SA obligation at the development phase. In particular, the reservation of copyright offers CC licensors the opportunity to make it a condition in the licensing contract

to Unlawful Sources Under Art. 3 CDSM Directive?' (Kluwer Copyright Blog, 22 December 2023) <https://copyrightblog.kluweriplaw.com/2023/12/22/saving-research-lawful-access-to-unlawful-sources-under-art-3-cdsm-directive/>; V. Stančiauskas and others, *Improving Access to and Reuse of Research Results, Publications and Data for Scientific Purposes – Study to Evaluate the Effects of the EU Copyright Framework on Research and the Effects of Potential Interventions and to Identify and Present Relevant Provisions for Research in EU Data and Digital Legislation, With a Focus on Rights and Obligations* (Brussels: Publications Office of the European Union 2024), 150-153 and 187-194 <https://data.europa.eu/doi/10.2777/633395>.

79 ISD, Article 5(1).

80 CDSMD, Article 3.

81 CDSMD, Article 4.

82 CC BY-SA 4.0, Section 2(a)(4).

that the final trained model be distributed under CLSA conditions – in the sense of imposing an obligation on AI trainers to pass on SA conditions to downstream recipients regardless of whether the artefact contains protected traces of copyright-protected CLSA material. As we will explain in the following section, this possibility of preserving CLSA conditions rests on the opt-out mechanism and contractual obligations which the CC licensor imposes on AI trainers using CLSA material for ML purposes. If the artefact does not contain copyright-protected traces of CLSA training material and, hence, does not constitute adapted material in the sense of the CC licence, the enforcement of the SA condition must be based on the contractual obligation that was established with the model developer (licensee) at the beginning of the development phase. Hence, the focus shifts from copyright enforcement to the enforcement of contract terms in the relationship with the model developer.

II. Output/Exploitation Phase

48 The exploitation phase (use of generative AI systems based on a model trained on CLSA content) raises complex issues relating to the existence of copyright-relevant acts that may trigger CLSA obligations. Generative AI output often remains limited to general ideas, concepts, styles etc. that the AI system has deduced from human training material during the development phase. According to the so-called idea/expression dichotomy recognised in international copyright law, these general ideas, concepts, styles etc. do not enjoy copyright protection as long as they do not contain copyright-protected creative choices of the author of knowledge resources used for training purposes.[83]

49 Hence, the question arises whether copyright law offers a sufficient basis for imposing CLSA licensing terms on AI output at all. We recall that for these obligations to apply, the licence requires the sharing of 'adapted material'. As explained above, the CC BY-SA 4.0 licence defines adapted materials as:

> *material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor.*[84]

50 Content produced by a generative AI system trained on CLSA resources, however, need not display protected traces of individual human expression that would require permission under copyright law.[85] Compared to the development phase, the situation is thus markedly different. During the development phase, protected human works are used as learning resources for the AI model. Hence, there is a direct link between the ML process and the use of protected human literary and artistic works made available under CLSA licensing terms. With regard to AI output (inference), however, the copyright basis for triggering CLSA obligations is less clear. Once again: instead of reproducing individual expression – protected free, creative choices by a human author[86] – AI output may merely reflect unprotected ideas, concepts and styles.

51 In light of the long-standing and well-established idea/expression dichotomy in copyright law, it is thus important to distinguish between two different types of AI output in the context of CLSA licensing: first, AI output that only contains unprotected ideas, concepts or styles (section E.II.1) and, second, AI output that displays traces of copyright-protected CLSA material on which the AI model was trained (section E.II.2). We now turn to a more detailed discussion of these scenarios.

83 Article 9(2) TRIPS; Article 2 WCT. As to the role of the idea/expression dichotomy in the generative AI debate, see M.A. Lemley and B. Casey (n 1), 772-776. With regard to the approach in the EU, see M.R.F. Senftleben, *The Copyright/Trademark Interface – How the Expansion of Trademark Protection Is Stifling Cultural Creativity* (The Hague, Kluwer Law International 2020), 27-28. See also Dutch Supreme Court, 29 March 2013, ECLI:NL:HR:2013:BY8661, *Broeren v Duijsens*, para. 3.5.

84 CC BY-SA 4.0, Section 1(a).

85 M.A. Lemley and B. Casey, 'Fair Learning', Texas Law Review 99 (2021), 743 772-776.

86 Case C-5/08, *Infopaq v DDF*, para 45; Case C-145/10 *Painer v Standard Verlags and Others*, para 89.

1. AI Output Consisting Of Unprotected Ideas, Concepts Or Styles

52 First, it is conceivable that AI output merely reflects unprotected ideas, concepts, styles etc. Due to the idea/expression dichotomy, it can be ascertained, as a default position, that this type of AI output falls outside the scope of copyright protection altogether. At the international level, Article 9(2) of the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) and Article 2 of the WIPO Copyright Treaty confirm this conclusion. Practically speaking, this means that copyright law does not offer a basis for extending CLSA obligations to this type of AI output. Considering the whole analysis conducted so far, it can even be said that the copyright link with CLSA obligations is broken twice:

- as explained in section C.II, the end result of the development phase (the final trained model) need not contain traces of copyright-protected CLSA training material. If the artefact only contains data points and vectors which, as such, no longer constitute copies of copyright-protected individual expression taken from CLSA works, the artefact does not constitute adapted material within the meaning of the CC licence and copyright is no longer available to enforce CLSA conditions;

- moreover, if the AI modelonly generates output consisting of unprotected ideas, concepts and styles, copyright relevance must also be denied with regard to this output. If AI output does not include protected features of original CLSA material used for training purposes, copyright is no longer available as a legal tool to attach SA obligations to AI output.

53 If this result is deemed unsatisfactory, it is important to explore a remaining avenue for placing SA obligations on AI output: the use of contractual stipulations. We hinted at this possibility already at the end of the preceding section. The unavailability of copyright as an enforcement tool need not lead to a situation where CLSA conditions can no longer be imposed on model recipients and end users altogether. It only means that an alternative legal tool must be employed, namely a chain of contractual obligations that starts when CLSA works are included in training resources for AI models. To develop the whole chain, the CC licensor must make sure that contractual CLSA obligations are consistently passed on from the model trainer using CLSA works to model recipients and end users.

54 To achieve this result, it is conceivable to require AI developers using CLSA works to introduce contractual terms that oblige recipients of the final AI model – the stage five artefact in our analysis in section C.II – to accept SA obligations. In this way, CLSA conditions can be passed on to model recipients who would then be bound to observe SA obligations when including the final, CLSA-trained model in AI systems and enabling end users to generate AI output. To ensure that the chain of contractual CLSA obligations is not broken, providers of AI systems (recipients of the final model) must also be obliged to make sure that end users who generate AI output are bound to observe CLSA conditions with regard to the content that results from their prompts. Implementing this chain of CLSA obligations on the basis of contractual agreements, it no longer matters whether the artefact contains copyright-protected traces of CLSA works. It also does not matter whether AI output displays copyright-protected features of CLSA training material. On the basis of contract law, the obligation to observe CLSA conditions can be extended to model recipients and end users regardless of copyright claims.

55 As indicated above, the opt-out mechanism in the general TDM provision laid down in Article 4 CDSMD could serve as a legal vehicle to forge this chain of contractual obligations starting with the acceptance of CLSA obligations by the AI developer who, then, would have to pass on these obligations to model recipients and end users. Following this approach, users of CC licences could reserve copyright in accordance with Article 4(3) CDSMD strategically to extend contractual SA obligations to recipients of trained models and end users generating AI output. To achieve this result, CC licensors must seize the opportunity to reserve copyright and subject the use of CLSA material in the world of AI-generated content to conditions, such as SA. Seeking to implement this approach, it is thus necessary to declare an opt out under Article 4(3) CDSMD and

employ copyright as a legal tool to make the use of CLSA material in TDM activities (falling outside the scope of the research rule in Article 3 CDSMD) dependent on compliance with conditions that allow the downstream maintenance of SA obligations.

56 This approach need not lead to a categorical exclusion of CLSA material from AI training datasets. By contrast, a tailor-made licence solution can grant AI developers broad freedom to use CLSA resources for training purposes. In exchange for the training permission, however, AI developers would have to accept CLSA obligations, including the obligation to create a whole chain of contractual agreements that binds model recipients and end users:

- *model recipients*: AI trainers using CLSA resources must be obliged to make the final trained model available only if the model recipient accepts SA conditions and agrees to pass on these obligations to end users. As a result, recipients of AI models trained on CLSA resources would be obliged to ensure that SA conditions are also attached to AI output generated by users;

- *end users:* to implement this in practice, model recipients must be obliged to embed SA conditions in the contractual terms governing the use of their AI systems and require users to accept these conditions.This could be enforced by refusing to respond to prompts unless the user agrees to be bound by the SA obligation. As this extension of SA conditions to users would follow from contractual terms accompanying the use of the AI system, it is immaterial whether the AI output displays copyright-protected features of original CLSA material or consists of unprotected ideas, concepts or styles. As the SA obligation follows from a contract, the copyright status of the output is not decisive.

57 The underlying legal-doctrinal machinery can be described as follows: the TDM opt out mechanism in Article 4(3) CDSMD is used as leverage to impose contractual CLSA obligations. The CC licensor invokes Article 4(3) CDSMD to opt out and exclude the statutory use permission that would otherwise follow from Article 4(1) CDSMD. As a result, the licensor can rely on copyright to impose specific CLSA licensing terms. On the one hand, the licence offers broad freedom to use the CLSA material for AI training purposes. On the other hand, the licence obliges the AI developer to make available the final trained model under SA conditions – regardless of whether the artefact contains copyright-protected traces of the CLSA training material. On its merits, the reservation of copyright is thus used to create a bargaining opportunity to conclude a regular contract with specific CLSA obligations.

58 If an AI developer refuses to accept the CLSA conditions, or does not comply with them, acts of reproducing CLSA material during the training stages one to three (see section C.II) fall outside the licence and amount to copyright infringement. If the final artefact (stage five) does not include copyright-protected traces of CLSA training material, the establishment and further distribution of the trained model is unlikely to constitute copyright infringement. However, it culminates in a breach of the contractual CLSA conditions which the CC licensor made a precondition for the initial use permission underlying the whole ML process. Including the obligation to pass on SA obligations to users of generative AI systems, the CLSA conditions are extended to the exploitation phase where AI output is produced.

59 In the AI licensing arena, the success of the described SA extension strategy will depend on the attractiveness and importance of CC resources for AI training. If alternative training resources are available that do not require the acceptance of CLSA obligations, AI developers may prefer these alternative materials. Finally, it must be considered that the chances of enforcing CLSA conditions in AI contexts may depend on the role of CLSA resources in the data amalgam applied for AI training. If CLSA material only plays a minor role, it may be difficult to trace AI output back to CLSA training resources and provide evidence of the violation of CLSA licence terms.

2. AI Output Displaying Protected Traces Of CLSA Training Material

60 The second scenario that we outlined above concerns the situation where AI output reproduces copyright-protected features of CLSA works that have been

used as training resources. This second scenario can hardly be described as a 'mainstream' scenario. As stated above, the first scenario – AI output that only displays unprotected ideas, concepts or styles – seems much more common. Nevertheless, considering the large volume of AI output – systems capable of producing a myriad of content items in a relatively short period of time – it simply cannot be ruled out that, perhaps even with high statistical probability, some AI-generated content items display copyright-protected features of CLSA works that were part of work repertoires used during the ML process. In this case, the equation is markedly different.

61 Using EU copyright law as a reference point, it can be said that, as a rule of thumb, the moment AI output contains copyright-protected features of source materials used for training purposes, copyright law provides a basis for introducing CLSA obligations. As already explained above, the CJEU has confirmed that, for takings from original works to amount to a relevant partial reproduction in the sense of copyright law, it is necessary that copied elements fulfil the originality test. That is only the case when these elements – scrutinised in isolation – reflect a sufficient degree of free, creative choices to qualify as their author's 'own intellectual creation'.[87] In other words: if copyright-relevant traces of CLSA training resources can be identified in AI output, this AI output offers a basis for arguing that the AI system has generated 'adapted material' in the sense of the CLSA approach. As already concluded above in the light of the CJEU's *Deckmyn* decision,[88] it seems overly restrictive and perhaps strategically undesirable to require, when drafting CC licenses, that adapted material have original features of its own – coming on top of protected elements of the original CLSA material. Even if the terms of a contract define the term as requiring that modifications of the original CLSA material be independently eligible for copyright protection, it may still be possible to demonstrate that sufficient human creative choices have been made during an iterative prompt writing process, or have been added after receiving the AI output to refine the final result.[89] Either way, if AI output contains traces of the original 'licensed material', this creates a possibility for CC licensors to argue that the use and further distribution of this AI output amounts to copyright infringement unless the user observes the CLSA conditions under which the licensor is willing to give a licence. More concretely, whilst the licensed material found in AI output may be reproduced and shared, in whole and in part, no terms or technological measures may be imposed to restrict these uses, acts of sharing the material must retain copyright information, indicate modifications and licence information, and any further recipient of the material must be subjected to the same SA terms.

62 At this point of our analysis, it seems important to point out that, in the case of AI output displaying copyright-protected features of CLSA works, a finding of copyright infringement does not necessarily depend on whether the user triggering the content with its prompt is actually aware of the fact that the AI output infringes a pre-existing work. While the CJEU has introduced a subjective knowledge criterion in hyperlinking cases,[90] other infringement situations, such as the further sharing and making available of AI output with copyright-protected features of CLSA works in social media or on online platforms, do not offer users the opportunity to routinely rebut

87 Case C-5/08 *Infopaq v DDF*, paras 38-39.

88 Case C-201/13 *Deckmyn and Vrijheidsfonds VZW v Vandersteen and Others*, para 21.

89 As to the traditional copyright originality test requiring free, creative choices of a human author, see once again Case C-5/08 *Infopaq v DDF*, para 45; Case C-145/10 *Painer v Standard Verlags, and Others*, para 89. As to the impact of this originality test on copyright protection for AI productions in the literary and artistic field, see P.B. Hugenholtz and J.P. Quintais, 'Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?', International Review of Intellectual Property and Competition Law 52 (2021), 1190 1212-1213; D. Burk, 'Thirty-Six Views of Copyright Authorship, by Jackson Pollock', Houston Law Review 58 (2020), 263 270-321; J.C. Ginsburg and L.A. Budiardjo, 'Authors and Machines', Berkeley Technology Law Journal 34 (2019), 343 395-396; M.-C. Janssens and F. Gotzen, 'Kunstmatige Kunst. Bedenkingen bij de toepassing van het auteursrecht op Artificiële Intelligentie', Auteurs en Media 2018-2019, 323 325-327; R. Pearlman, 'Recognizing Artificial Intelligence as Authors and Investors under U.S. Intellectual Property Law', Richmond Journal of Law and Technology 24 (2018), 1 4.

90 Case C-160/15, *GS Media v Sanoma Media Netherlands and Others*, paras 49-51.

infringement arguments by simply stating that they had no knowledge of traces of protected works in the AI output. In a litigation setting where two people created the same content (or roughly similar content), the defendant to an infringement claim (in our case the user triggering infringing AI output) would have to give a credible story of how they came up with the individual expression independently.[91] Demonstrating that the user was not aware of the use of copyright-protected CLSA material for AI training purposes might not suffice. While this is of course an issue which the CJEU might have to clarify at some point, the default position in current copyright law remains that someone appears to have exploited the pre-existing copyright-protected work whenever a copy of that work is created. The ball is then in the alleged infringer's court. In other words: the AI user would have to advance convincing arguments to rebut the infringement claim.

63 Arguably, this liability risk offers opportunities to infuse CLSA conditions. In principle, every user of CLSA resources (anyone further downstream) can receive an offer from the original CC licensor to use the licensed material and include traces of this licensed material in adapted material (such as portions of AI output that relate to the licensor's content). The mere availability of the licence and the offer of an authorisation under CLSA conditions, however, does not imply that every downstream user is aware of this opportunity to receive permission and escape the verdict of infringement. Hence, it is necessary that the downstream user triggering the production of AI output be informed about the licence offer and encouraged to accept this offer.

64 To achieve this result, we must navigate between two different contributions leading to AI output that contains protected features of original CLSA material: the AI provider makes available the system that produces this content. However, the final production of the AI output is triggered by a different person, namely the end user. With regard to this amalgam of system provider and user involvement, several considerations seem relevant. The user does not have access to the training dataset, nor is the user likely to be aware of what was part of the training dataset. An AI system provider using a CLSA-trained model, by contrast, may be aware of CLSA material used during the ML process – either because the provider conducted the AI training himself (same person), or because the AI trainer (being another person) passed on SA obligations in accordance with the contractual strategy developed in the preceding section. The AI system provider, however, does not enter the prompt.

65 Nonetheless, it may be possible to establish a sufficient link with the AI system provider when it is considered that this person exercises possessive control over the AI system and has designed the user interface enabling the user to request the generation of AI output, in accordance with the freedoms and limitations set by the system provider. From the perspective of EU copyright law, it is conceivable that this role is sufficient to impose an obligation to ensure observance of the CLSA terms with regard to the AI output. Arguably, a parallel can be drawn with the CJEU decision in *The Pirate Bay* where the Court considered that the operation of an online platform that indexed information about copyright-protected material without hosting that material, and which made it easier to locate that material, carried out an act of communication to the public within the meaning of Article 3 ISD.[92] The Court had paved the way for this broad application of the right of communication to the public – *de facto* collapsing the traditional distinction between primary liability of the user who uploads infringing content, and secondary, contributory liability of the platform – in the earlier decision in *Filmspeler*. In that case, the Court had dealt with the offer of multimedia players with pre-installed add-ons that specifically enabled purchasers to have access to protected works published illegally on streaming websites.[93] Instead of raising the question whether harmonised EU law provided a basis for assuming secondary, contributory liability to infringing content sharing, the CJEU held that the sale of such a multimedia player constituted a primary act of communication

91 Cf. N. Elkin-Koren and others, 'Can Copyright be Reduced to Privacy?' (*arXiv*, 24 March 2024), 1-2 <https://arxiv.org/abs/2305.14822>.

92 Case C-610/15 *Stichting Brein v Ziggo and XS4ALL Internet (The Pirate Bay)*, paras 36-39 and 47.

93 Case C-527/15 *Stichting Brein v Wullems (Filmspeler)*, para. 41.

to the public in the sense of Article 3(1) ISD.[94]

66 To support this remarkable extension of the concept of 'communication to the public' to the preparatory phase of offering and selling a multimedia player – a phase in which the purchaser has not yet set in motion the process of accessing illegal content – the Court focused on knowledge of infringing conduct and the aim to exploit illegal streaming content. The 'Filmspeler' multimedia player was sold with full knowledge that the add-ons, which included pre-installed hyperlinks gave access to works published illegally on the internet.[95] Following this approach, it cannot be ruled out that the AI system provider must be deemed the adapter, or co-adapter, in the case of AI output that displays protected features of CLSA material. In practice, this co-responsibility means that, even if a system user triggers the production of the AI output, the AI system provider is obliged to ensure that the SA conditions are observed. Otherwise, the CC licensing conditions are not fulfilled and the AI system provider exposes himself and users of the AI system to the described copyright infringement risk.

67 In line with the outlined CJEU approach, this responsibility of the AI system provider follows from the fact that, having included CLSA resources in the training dataset himself, or having been informed about this by the AI trainer, the AI system provider must be well aware that output produced by the AI system may contain protected traces of original CLSA works. Hence, it can be argued that the AI system provider offers the AI system in full knowledge of the fact that AI output with protected CLSA ingredients may result from the use of the system. To reduce this liability risk, the AI system provider should introduce the CLSA obligations accompanying the training material and pass on these obligations to users. As discussed in the preceding section, the AI system provider can, for instance, make the generation of AI output following a user prompt dependant on acceptance of the CLSA terms that are attached to the material used for training purposes.

68 The same strategy can be applied when the described parallel with the CJEU's *Filmspeler* approach is deemed unconvincing and the user entering the prompt for the AI output is regarded as the only person responsible for the AI production containing copyright-protected traces of CLSA training material. To reduce liability risks for users in this situation, it is desirable that AI system providers include CLSA obligations in the terms of use relating to AI systems that are based on CLSA-trained models. To pass on CLSA obligations to users of the final AI system and reduce their liability risk, it is advisable to follow the approach described in the preceding section and adopt additional contractual obligations, namely the obligation to include CLSA clauses in the terms of use accompanying the AI system. In this way, it can be ensured that users become aware of CLSA obligations. In addition, it can be stated that, by using the AI system and entering prompts, the user implicitly accepts the CLSA terms and the obligation to distribute AI output under SA conditions. As already proposed, users could be obliged to accept CLSA terms before the AI system produces output following a user prompt.

69 However, it is important to recall again that the concept of 'adapted materials', as defined in the CC BY-SA 4.0 licence, does not include material created on the basis of copyright exceptions and limitations. Therefore, any relevant copyright exception that could apply to AI output insofar as the copyright status of the material is concerned, will affect the status of the generated material. Even if a prompt leads to AI output with protected CLSA features, copyright exceptions, such as the exemption of quotations, parodies, caricatures and pastiches in EU copyright law,[96] may prevail over CLSA

94 id., para 52.

95 id., paras 50-51.

96 ISD, Article 5(3)(d) and (k). Cf. G. Westkamp, 'Borrowed Plumes: Taking Artists' Interests Seriously in Artificial Intelligence Regulation', 1 19-26, forthcoming; M.R.F. Senftleben, 'User-Generated Content – Towards a New Use Privilege in EU Copyright Law', in T. Aplin (ed), *Research Handbook on IP and Digital Technologies* (Cheltenham: Edward Elgar 2020), 136 (145-162); S. Jacques, *The Parody Exception in Copyright Law* (Oxford: Oxford University Press 2019), 91-133; E. Hudson, 'The pastiche exception in copyright law: a case of mashed-up drafting?', Intellectual Property Quarterly (2017), 346 362-364; F. Pötzlberger, 'Pastiche 2.0: Remixing im Lichte des Unionsrechts', Gewerblicher Rechtsschutz und Urheberrecht 2018, 675 681; J.P. Quintais, *Copyright in the Age of Online Access – Alternative Compensation Systems in EU Law* (Alphen aan den Rijn: Kluwer Law International

obligations in cases where, as a result of iterative prompt writing and use of the AI system as a tool for human expression, or the addition of human creative choices to AI output,[97] the AI system user can invoke these copyright exceptions.

F. ShareAlike/Copyleft Options In The Era Of Generative AI

70 Our analysis demonstrates that challenges concerning successful deployment of copyleft licences relate predominantly to the design of the licenses, which are bound to differ in scope because of a fragmented copyright framework across the globe. If it is deemed desirable to preserve the CLSA approach in the era of generative AI and attach SA obligations to AI output, it will be necessary to revise the licences. Ultimately, it may be inevitable to rely on the bargaining power that the reservation of copyright offers to ensure the continued viability of CLSA licences. Indeed, this is the very idea of copyleft licensing – to *rely* on the prerogatives that copyright law provides in order to ensure that downstream creations that are derived from the original material are made available on the same terms to others. Taking EU copyright law as a reference point, two markedly different policy options are available:

71 On the one hand, the CC community can uphold the supremacy of copyright exceptions. In countries and regions that exempt ML processes from the control of copyright holders, this approach leads to far-reaching freedom to use CC resources as training material for AI systems. At the same time, it is likely to marginalise SA obligations in the realm of literary and artistic AI productions. In the EU, for instance, an approach that allows TDM exceptions to prevail over CLSA licensing conditions implies that AI developers are free to invoke Articles 3 and 4 CDSMD and use original CLSA material for AI training purposes without seeking permission – and without accepting SA obligations. In consequence, it seems particularly difficult, if not impossible, to impose SA obligations with regard to output generated by the fully trained AI system. As AI developers need not subscribe to CLSA conditions, there is hardly any possibility of requiring them to observe these conditions when generating AI output themselves, or pass on CLSA obligations to users who trigger the production of AI output with their prompts. In sum, supremacy of copyright exceptions can easily lead to a situation where SA obligations play hardly any role in the context of generative AI systems and literary and artistic output produced by these systems.

72 On the other hand, the CC community can use copyright strategically to extend SA obligations to AI training results and AI output. To achieve this goal, it is necessary to seize opportunities to reserve copyright and subject the use of CC material in the world of AI development and exploitation to conditions, such as SA. Following this approach, it is advisable to declare an opt out under Article 4(3) CDSMD and employ copyright as a legal tool to make the use of CLSA material in TDM activities (falling outside the scope of the research rule in Article 3 CDSMD) dependent on compliance with conditions that allow the maintenance of SA obligations. This approach need not lead to a categorical exclusion of CLSA material from AI training datasets. By contrast, a tailor-made licence solution can grant AI developers broad freedom to use CLSA resources for training purposes. In exchange for the training permission, however, AI developers would have to accept CLSA obligations. With regard to the AI development phase, this could include the obligation to make the trained model available in accordance with SA conditions. At the AI exploitation stage, AI developers would be obliged to ensure – via a whole chain of contractual obligations – that SA conditions are also attached to AI output generated by AI systems that use CLSA-trained models. As AI output may result from user prompts, this includes an obligation to embed SA conditions in the contractual terms governing the use of the AI system and require users to accept these conditions, for instance by refusing to respond to prompts unless the user agrees to be bound by the SA obligation.

2017), 235; M.R.F. Senftleben, 'Quotations, Parody and Fair Use', in P.B. Hugenholtz, A.A. Quaedvlieg, and D.J.G. Visser (eds), *A Century of Dutch Copyright Law – Auteurswet 1912-2012* (Amstelveen: deLex 2012) 359 365.

97 Cf. P.B. Hugenholtz/J.P. Quintais (n 89), 1212-1213; D. Burk (n 89), 270-321; J.C. Ginsburg and L.A. Budiardjo, 'Authors and Machines', Berkeley Technology Law Journal 34 (2019), 343 (395-396); M.-C. Janssens and F. Gotzen (n 89), 325-327; R. Pearlman (n 89) 4.

As this extension of SA conditions to users would follow from contractual terms accompanying the use of the AI system, it is immaterial whether the AI output displays copyright-protected features of original CLSA material or consists of unprotected ideas, concepts or styles. As the SA obligation follows from a contract, the copyright status of the output is not decisive. However, the copyright status becomes relevant in the case of further downstream use. If the AI output does not contain copyrighted elements, it is unclear how the SA condition can be asserted against downstream users who are not bound by the conditions accompanying the use of the AI system.

73 In addition to these general policy options, the analysis has yielded several more specific insights:

- The SA condition, as expressed in the CC BY-SA licence, is designed with reference to adapted material. For traditional forms of artistic expression that involve investment of time, resources and creativity to adapt pre-existing works, this is a logical design. In the context of ML processes and the generation of AI output, however, the focus on adapted material may be less efficient as it introduces unnecessary complexity to cover activities that for the most part involve technical modifications at the development stage and comparatively few human creative choices in the exploitation phase leading to literary and artistic AI output. It may therefore be preferable to focus on use of original CLSA material in AI training and the potential reappearance of traces of this original material in AI output. In other words: the use and reappearance of CLSA material in these context should be decisive and trigger SA obligations – not the question whether AI processes lead to adapted material.

- A CC licence that includes a ban on TDM activities will remove the applicability of the Article 4(1) CDSMD copyright exception in favour of letting the use be governed by a more specific, tailor-made use permissions. That is, the objective would be to trigger CLSA licence conditions where they otherwise would have been governed by an exception. As follows from the CC Statement on the Opt-Out Exception Regime,[98] the CC BY-NC-SA licence has the potential of effecting an opt-out for non-commercial use. But pursuant to our analysis, for the opt-out to foster CLSA culture more broadly in AI contexts, it may be advisable to abandon the traditional precedence of copyright exceptions in favour of an opt-out protocol that allows a more fine-grained TDM permission that includes SA obligations. As CC has already undertaken initiatives to enable the association of machine-readable licensing metadata with objects offered under CC licences through the CC Rights Expression Language (ccREL), an opt-out declaration of this nature could also be expressed by machine-readable means.

- Interestingly, developers of AI models may experience SA extension difficulties that are comparable to those faced by creators of CLSA material. Copyleft options designed for software may be deemed more or less inadequate for distributing AI models. In this respect, the evolution of AI model licences (ML model licences), for example OPT-175B, CreativeML Open RAIL-M, BigScience OpenRAIL-M, GLM-130B, provides useful insights into trends in the machine-learning sector. These developments in the sector may offer important reference points for adaptations of existing CLSA licence schemes with regard to use of CC resources as training data. For instance, an alternative approach to adapting CLSA licences that is worth exploring is the viability of adapting ML model licences to be compatible with the former by accounting for the training data as a mode of realising responsible AI licensing (RAIL). Such endeavours could additionally align with the proposed obligations imposed on AI model developers to put in place a copyright compliance policy and the making available of detailed summaries about the materials used for training general-purpose AI models pursuant to Article 53(1)(c)-(d) of the AI Act. Arguably, these obligations also apply to developers of AI models released under free and open licences.

- Finally, our analysis has been limited with

98 A. Lazarova and others (n 61).

regard to the spectrum of further technological development that we could cover. We have mostly approached the issues from the perspective of so-called supervised learning. However, advances in self-supervised learning has led to ML processes on unstructured data. Self-supervised learning is likely to involve increasingly less copying, with a comparatively lesser amount of different acts and human interventions. It may ultimately lead to a focus on developing foundational models that have undergone training, diminishing the need for developing them from scratch. You only need to invent the pneumatic tire once and then you concentrate on making it better to achieve the desired shock absorption, traction or manoeuvrability properties. In a similar vein, training datasets might eventually become a thing of the past once AI systems no longer need training but only tweaking. This might not remove the need for supplying new facts or knowledge but it may optimise the entire learning process. Moreover, with advances in generative AI, training may increasingly involve training based on synthetic data generated by AI and lead to systems learning from each other in the same way as AI is used today for finding errors in computer code or optimising it. Perhaps the best way of thinking about AI is as if it were an operating system. In the end, there will be only a few developers because everybody else will be developing or finding applications for it.