Law, Innovation and Technology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rlit20

Can text and data mining exceptions and synthetic data training mitigate copyright-related concerns in generative AI?

Maryna Manteghi

To cite this article: Maryna Manteghi (28 Aug 2024): Can text and data mining exceptions and synthetic data training mitigate copyright-related concerns in generative AI?, Law, Innovation and Technology, DOI: 10.1080/17579961.2024.2392928

To link to this article: https://doi.org/10.1080/17579961.2024.2392928

Published online: 28 Aug 2024.

LAW, INNOVATION AND TECHNOLOGY
https://doi.org/10.1080/17579961.2024.2392928

Routledge
Taylor & Francis Group

Can text and data mining exceptions and synthetic data training mitigate copyright-related concerns in generative AI?

Maryna Manteghi

Faculty of Law, University of Turku, Turku, Finland

ABSTRACT

Rapidly emerging generative artificial intelligence (GenAI) models stand at the epicentre of current public discourse. They demonstrate impressive abilities to generate various types of data promptly and cost-effectively. However, AI developers need to train their systems on massive volumes of data which is usually copyrighted. Therefore, the growth of copyright-related concerns in the field of GenAI comes as no surprise. The study introduces two solutions which could mitigate the tension between copyright holders and AI developers, one legal (text and data mining (TDM) exceptions of the CDSM Directive) and one technical (synthetic data), highlighting the promises and challenges of both. First, the article will discuss the capability of TDM exceptions to facilitate the fundamental right to information and the freedom of research in the context of AI development. Next, the paper will analyse how providers of GenAI models can leverage synthetic data to comply with copyright law while training their systems and what risks might be associated with this approach. The findings of this study will indicate what issues, in both legal and technical spheres, should be addressed to ensure a balance of powers in the digital environment and effective functionality of the EU AI sector.

ARTICLE HISTORY Received 3 November 2023; Accepted 23 April 2024

KEYWORDS generative AI models; copyright; text and data mining; CDSM Directive's TDM exceptions; Artificial Intelligence Act; fundamental rights

1. Introduction

A generative artificial intelligence (GenAI) model refers to an AI technology that can generate new outputs, based on insights derived from existing data.[1]

CONTACT Maryna Manteghi maryna.manteghi@utu.fi

[1]Artha Dermawan, 'Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn from the Japanese 'Nonenjoyment' Purposes?' (2023) 27(1) *The Journal of World Intellectual Property* 44, 45; Van Lindberg, 'Building and Using Generative Models Under US Copyright Law' (2023) 18(2) *Rutgers Business Law Review* 1, 1; Christopher Callison-Burch,

The recent advancement of GenAI demonstrates the shift from 'narrow AI'[2] to more general AI systems, known as Artificial General Intelligence (AGI),[3] distinguished by broad intelligence capabilities at or above the human level.[4] AI experts acknowledge that the rapidly emerging models such as ChatGPT, GitHub Copilot, DALL-E, Midjourney, Stable Diffusion constitute preliminary versions of AGI.[5] The systems possess a remarkable capability to perform a range of challenging tasks, including question answering, images or text generation, coding, sentiment analysis and others.[6] Even though the rapidly emerging GenAI models could advance various aspects of our life, there are some concerns that the cutting-edge applications would affect those working in the creative sector. The concerns are twofold: (1) the use of creative works for the development and exploitation of GenAI systems without prior permission, and (2) the use of AI-powered generators to produce creative works which could replace those created by human actors. The first issue could lead to a situation when human creators become less motivated to produce new content as they would have less control over their works, and the second point may raise questions regarding the future commercial viability of authors as creative workers. The article will specifically discuss fair and unfair exploitation of copyright-protected works during the development of GenAI models to get to the roots of the problem.

The building of GenAI models involves two key phases: training (a so-called 'input' phase) and content generation (a so-called 'output' phase).[7] Text and data mining (TDM) is an essential tool for the training of GenAI systems.[8] The technique facilitates the analysis of mass volumes of existing data by extracting patterns, insights, trends and correlations which can be used to 'feed' AI algorithms.[9] However, a computer has to make copies of obtained content to carry out TDM and thus extract underlying knowledge which

'Understanding Generative Artificial Intelligence and Its Relationship to Copyright' [2023] < https://docs.house.gov/meetings/JU/JU03/20230517/115951/HHRG-118-JU03-Wstate-Callison-BurchC-20230517.pdf > accessed 27 January 2024.

[2]'Narrow AI' refers to AI systems that are designed to perform a specific task or a set of closely related tasks (see Andreas Kaplan and Michael Haenlein, 'Siri, Siri, in my Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence' (2019) 62(1) *Business Horizons* 15, 16).

[3]Peter Henderson et al., 'Foundation Models and Fair Use' [2023] <https://arxiv.org/pdf/2303.15715.pdf> accessed 13 February 2024, 1, 4.

[4]Ibid 4.

[5]Ibid 4, Ehsan Latif, 'AGI: Artificial General Intelligence for Education' [2024] < https://arxiv.org/pdf/2304.12479.pdf > accessed 23 March 2024, 1, 4.

[6]Avanika Narayan et al., 'Can Foundation Models Wrangle Your Data?' (2022) 16(4) PVLDB 738, 738.

[7]Van Lindberg (n 1) 4; Apoorva Verma, 'The Copyright Problem with Emerging Generative AI' [2023] SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4537389 > accessed 12 February 2024, 5.

[8]Ethem Alpaydin, *Introduction to Machine Learning* (MIT Press 2004) 1, 2.

[9]For a general overview of TDM techniques and methods see, for instance, Jiawei Han et al., *Data Mining: Concepts and Techniques* (3rd edn, Elsevier 2012) 1–703; Artha Dermawan (n 1) 45; Van Lindberg (n 1) 4; Christopher Callison-Burch (n 1); Matthew Sag, 'Copyright Safety for Generative AI' (2023) 61 *Houston Law Review* 295, 305; Apoorva Verma (n 7) 5.

would allow AI systems to make predictions and generate desired outputs.[10] Training datasets usually consist of a plethora of creative and original data (e.g. images, music, articles etc.,) which could be copyrighted. In this regard, if input data is mined without authorisation, it may infringe copyright holders' exclusive right to reproduction.[11] To prevent potential copyright infringement, providers of GenAI would have to obtain permission from rightsholders to lawfully reproduce protected works.[12] However, obtaining licences for huge amounts of digital data, the ownership of which is usually difficult to identify, is not an easy task.[13] In this sense, AI developers may consider training their systems on public domain works only, or on data which is freely available online through open-source or Creative Commons licences. However, to effectively train sophisticated GenAI models, plenty of resources should be accessed and mined.[14] If trained data is limited in terms of quantity, scope or quality, the outcome could be insecure and scarce, leading to the development of lower-quality AI applications.[15]

Against this background, providers of GenAI could elucidate if they are eligible for protection under copyright exceptions and limitations. In the EU, TDM is regulated under two specific TDM exceptions introduced under Arts. 3 and 4 of the Directive on Copyright in the Digital Single

[10]Nicola Lucchi 'ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems' [2023] *European Journal of Risk Regulation* < https://www.cambridge.org/core/services/aop-cambridge-core/content/view/CEDCE34DED599CC4EB201289BB161965/S1867299X23000594a.pdf/chatgpt-a-case-study-on-copyright-challenges-for-generative-artificial-intelligence-systems.pdf > accessed 11 February 2024, 1, 11.

[11]Directive (EU) 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 OJ L 167, 22.6.2001, p. 10–19 (the InfoSoc Directive), Art. 2. See for more discussions regarding the question of input, for instance, ibid 10–15.

[12]In this sense, see numerous US GenAI lawsuits (still pending) that allege that using copyrighted data for training AI systems constitutes copyright infringement. For instance, see, *Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135-GBW* (D. Del. Mar. 29, 2023); *Silverman et al. v. OpenAI, Inc. et al., No. 4:23-cv-03416* (N.D. Cal. Jul. 7, 2023); *Tremblay et al. v. OpenAI, Inc. et al., No. 4:2023-cv-03223* (N.D. Cal. Jul. 7, 2023).

[13]Christopher Callison-Burch (n 1) 13; Joao Pedro Quintais, 'Generative AI, Copyright and the AI Act' (*Kluwer Copyright Blog*, 9 May 2023) < https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/ > accessed 12 January 2024; Martin Senftleben, 'Generative AI and Author Remuneration' (2023) 54 IIC 1535, 1544-1545.

[14]Alberto Romero, 'A Complete Overview of GPT-3 – The Largest Neural Network Ever Created' (*Towards Data Science,* 24 May 2021) < https://towardsdatascience.com/gpt-3-a-complete-overview-190232eb25fd > accessed 2 March 2024; Tom Brown et al., 'Language Models Are Few-Shot Learners' (2020) < https://arxiv.org/abs/2005.14165 > accessed 18 February 2024; Carys J. Craig, 'The AI-Copyright Challenge: Tech-Neutrality, Authorship, and the Public Interest' in Ryan Abbott (ed.), *Research Handbook on Intellectual Property and Artificial Intelligence* (Edward Elgar Publishing 2022) 134, 150.

[15]European Copyright Society, 'General Opinion on the EU Copyright Reform Package' (2017) < https://europeancopyrightsocietydotorg.files.wordpress.com/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf > > accessed 20 February 2024, 3; Carys J. Craig (n 14) 150; Christopher Callison-Burch (n 1) 14; Apoorva Verma (n 7) 5; Katherine Lee et al., 'Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain' *Journal of the Copyright Society of the U.S.A.* (forthcoming 2024) SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4523551 > accessed 12 February 2024, 4; Thomas Margoni and Martin Kretschmer, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2022) 71 (8) *GRUR International* 685, 687.

Market (the CDSM Directive).[16] The adoption of the specific TDM exceptions in the course of copyright reform emphasises the importance of protecting the fundamental right to information[17] and the freedom of research[18] in the algorithmic society.[19] As Geiger and Jütte emphasise, the digital rights to TDM have introduced new ways of using fundamental rights to regulate the rapid development of digital technologies, secure creativity and facilitate access to information in the online environment.[20] The specific TDM exceptions could be viewed as an essential constitutional counterweight against the fundamental right to intellectual property (IP).[21] The provisions aim to ease control exercised by copyright holders over protected data and thus facilitate new developments and processes in the digital age.[22] This is an explicit

[16]Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/*EC* [2019] OJ L 130/92, 17.5. 2019, pp. 92–125 (CDSM Directive).

[17]The right to information is protected by art.11 of the EU Charter of Fundamental Rights [2012] OJ C326/391 (EUCFR) which provides the freedom of expression including freedom to hold opinions and to receive and impart information. See also art.10 (1) of the European Convention on Human Rights (ECHR); The European Court of Human Rights (ECtHR) has broadened the scope of the right to freedom of expression and information, also including the right of access to information in *Tarsasag a Szabadsagjogokert v Hungary* (37374/05) 14 April 2009; *Kenedi v Hungary* (31475/05) 26 August 2009; *Youth Initiative for Human Rights v Serbia* (48135/06) 25 September 2013 at [20] and [24]; *Timpul Info-Magazin and Anghel v Moldova* (42864/05) 2 June 2008 at [31]; *WVgrzynowski and Smolczewski v Poland* (33846/07) 16 October 2013 at [57]; *Sdruz˘ení Jihoeské Matky c./ la Rêpublique tchèque* (19101/03) 10 July 2006. In this sense, see also the Court of Justice of the European Union (CJEU) case *Stichting Greenpeace Nederland and Pesticide Action Network Europe (PAN Europe) v European Commission* (T-545/11) EU: T:2013:523.

[18]The freedom of research is based on the right to information (a part of the freedom of expression) (arts. 11 of the EUCFR) and Art. 13 of the EUCFR which requires that the scientific research should be free of constraints. For a more detailed analysis of the scope of the right to information see, for instance, Christophe Geiger and Bernd Justin Jütte, 'Conceptualizing a 'Right to Research' and Its Implications for Copyright Law: An International and European Perspective' (2022) Joint PIJIP/TLS Research Paper Series No.7-2022, https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1079&context=research 29–35; Christophe Geiger and Bernd Justin Jütte, 'The Right to Research as Guarantor for Sustainability, Innovation and Justice in EU Copyright Law' in T. E. Pihlajarinne, J. Mähönen and P. Upreti (eds), *Intellectual Property Rights in the Post Pandemic World: An Integrated Framework of Sustainability, Innovation and Global Justice* (Edward Elgar Publishing, forthcoming, 2023), 138, 161–163; Steve Peers et al. (eds), *The EU Charter of Fundamental Rights: A Commentary* (Hart Publishing, 2021), 336, 405; In this sense see Opinion of AG Szpunar in *Pelham GmbH v Hutter* (C-476/17) EU:C: 2019:624 at [91]; *EIT Nordisk Film & TV A/S v Denmark* (40485/02) 8 December 2005; *Dammann v Switzerland* (77551/01) 25 July 2006 at [52].

[19]Giovanni de Gregorio, 'The rise of digital constitutionalism in the European Union' (2021) 19 (1) *International Journal of Constitutional Law* 41, 57–58.

[20]Christophe Geiger and Bernd Justin Jütte, 'Digital Constitutionalism and Copyright Reform: Securing Access to through Fundamental Rights in the Online World' (*Kluwer Copyright Blog*, 25 January 2022) < https://copyrightblog.kluweriplaw.com/2022/01/25/digital-constitutionalism-and-copyright-reform-securing-access-to-through-fundamental-rights-in-the-online-world/ > accessed 11 February 2024.

[21]The authors' right as a part of IP is guaranteed as a human right under art.1 of Protocol 1 of the ECHR: 'Every natural or legal person is entitled to the peaceful enjoyment of his possessions. No one shall be deprived of his possessions except in the public interest and subject to the conditions provided for by law and by the general principles of international law' and as a fundamental right under art.17(2) of the EUCFR: 'Intellectual property shall be protected.' Copyright as a part of IP is also guaranteed at the international level by art.27 of the Universal Declaration of Human Rights and art.15 of the International Covenant on Economic, Social and Cultural Rights. For more discussions see Ibid.

[22]Christophe Geiger and Bernd Justin Jütte (n 20).

manifestation of 'digital constitutionalism,' a new theoretical trend which advocates the application of constitutional principles to the digital society.[23] Despite many positive nuances,[24] the exceptions consist of various limitations which could impede the capability of these provisions to balance powers in the digital environment.[25] The exception of Art. 3 of the CDSM Directive has been formulated narrowly and applies only to few beneficiaries. In particular, the provision benefits only research organisations and cultural heritage institutions which use TDM tools for the purpose of scientific research.[26] Even though the second so-called 'commercial' TDM exception of Art. 4 allows all types of users to carry out TDM for any purpose, it can be diluted by an 'opt-out' mechanism which allows rightsholders to reserve the use of their works for text and data mining.[27] Since its adoption, the provision has been widely criticised for strengthening the position of authors and other rightsholders in regard to the European AI sector.[28]

Alternatively, providers of GenAI may choose to rely on purely technical solutions to alleviate a conflict of interest in GenAI training. One option could be to train AI algorithms on synthetic data only. Synthetic data generally refers to 'artificially annotated information generated by computer algorithms or simulations.'[29] Synthetic data can be generated from real-world data by capturing the underlying properties of actual datasets (e.g. interactions, correlations etc.,) or it can be synthesised without using actual data, for instance, through statistical models, simulations or background knowledge of the analyst.[30] Once created, AI-synthesised data could be used to train

[23]Digital constitutionalism refers to 'the ideology which aims to establish and to ensure the existence of a normative framework for the protection of fundamental rights and the balancing of powers in the digital environment' (Edoardo Celeste, 'Digital Constitutionalism: A New Systematic Theorisation' (2019) 33 *International Review of Law, Computers & Technology* 88). See also Edoardo Celeste, *Digital Constitutionalism the Role of Internet Bills of Rights* (Routledge 2022) 82; Giovanni de Gregorio, 'The rise of digital constitutionalism in the European Union' (n 19) 57; Christophe Geiger and Bernd Justin Jütte (n 20).

[24]See for more discussions Maryna Manteghi, 'In Search of Balance: Text, Data Mining and Copyright in the Digital Single Market Directive from a Fundamental Rights Perspective' (2023) 48 *E.L. Rev*. 422, 425–427.

[25]Giovanni de Gregorio (n 19) 57-58; Nicolas Suzor, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4 (3) *Social Media + Society* 1, 4. Christophe Geiger and Bernd Justin Jütte (n 20).

[26]CDSM Directive, art. 3 and recitals 12-13.

[27]CDSM Directive, art. 4 (3) and recital 18.

[28]For a critical view on an 'opt-out' mechanism see, for instance, Maryna Manteghi (n 24) 432-433; Maryna Manteghi, 'Overcoming Barriers to Text and Data Mining in the Era of ChatGPT: The Proposed Data Act as a Game-Changer' (2024) 73(1) *GRUR International* 34, 40; Andrew Tyner, 'The EU Copyright Directive: 'Fit for the Digital Age or Finishing It?'' (2020) 26(2) JIPL 275, 281; Bernt Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)' (*Kluwer Copyright Blog*, 24 July 2019) < http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ > accessed 12 February 2024; Martin Senftleben (n 13) 1545.

[29]Yingzhou Lu et al., 'Machine Learning for Synthetic Data Generation: A Review' (2021) 14(8) *Journal of Latex Class Files* 1, 1.

[30]Khaled El Emam, 'Accelerating AI with Synthetic Data Generating Data for AI Projects' (*NVIDIA, O'Reilly Media* 2020) < https://www.nvidia.com/content/dam/en-zz/Solutions/deep-learning/resources/accelerating-ai-with-synthetic-data-ebook/accelerating-ai-with-synthetic-data-nvidia_web.pdf >

GenAI models. The demand for this technological approach to data analysis is growing within different domains, especially in the healthcare and technology sectors.[31] Many AI experts argue that it is cheaper, easier and quicker to generate training datasets based on synthetic data instead of real data.[32] Producing synthetic 'observations' of real-world data could significantly increase the volume and variability of training datasets.[33] Moreover, some authors emphasise that synthetic data is a purely artificial 'product' which has no direct connection to real-world data, but only mimics or simulates it, therefore, the use of this approach may alleviate certain legal risks (e.g. privacy and data protection) in AI training.[34] However, the lack of legal standards and mature tools for the usage of synthetic data, the risk of degradation of synthetic data over generations and the high possibility that AI-synthesised outputs would resemble actual data could limit the efficiency of this solution in the context of GenAI and copyright.[35]

Against this background, the article argues that it is needed to optimise the existing resources to mitigate the tension between copyright holders and providers of GenAI models. The study introduces two solutions, one legal (TDM exceptions of the CDSM Directive) and one technical (synthetic data), highlighting the promises and challenges of both. Therefore, the structure of this paper is as follows. After the introduction (Part I), the article will discuss the capability of the CDSM Directive's TDM exceptions to facilitate

accessed 03 February 2024, 3; Aryan Jadon and Shashank Kumar, 'Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy' (2023) < https://arxiv.org/abs/2305.05247 > accessed 21 January 2024, 2; Tshilidzi Marwala et al., 'The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development' (2023) < https://www.semanticscholar.org/reader/5c593e30366748a764190e4d80533c6e1aaef865 > accessed 12 March 2024, 3-4; Joao Fonseca and Fernando Bacao, 'Tabular and latent space synthetic data generation: a literature review' (2023) 10(115) *Journal of Big Data* 1.

[31] See, for instance, Aryan Jadon and Shashank Kumar (n 30) 3; Yingzhou Lu et al (n 29) 4.

[32] Sina Alemohammad et al., 'Self-Consuming Generative Models Go MAD' (2023) < https://arxiv.org/abs/2307.01850 > accessed 13 February 2024, 2; Abdul Majeed, 'Attribute-Centric and Synthetic Data Based Privacy Preserving Methods: A Systematic Review' (2023) 3 *Journal of Cybersecurity and Privacy* 638, 639; Erroll Wood et al., 'Fake It till You Make It: Face Analysis in the Wild Using Synthetic Data Alone' (2021) < https://arxiv.org/abs/2109.15102 > accessed 10 January 2024, 3682; Joao Fonseca and Fernando Bacao (n 30), 7; Yingzhou Lu et al (n 29) 1; Tshilidzi Marwala et al (n 30), 4.

[33] For instance, we can apply different augmentations to training images e.g., rotations, perspective warps, blurs, the addition of noise and others to increase the amount of training data (Erroll Wood et al. (n 32) 3685, 3688); See also Joao Fonseca and Fernando Bacao (n 30) 13.

[34] Fida K. Dankar and Mahmoud Ibrahim, 'Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation' (2021) 11 *Appl. Sci.* 2158, 2159; Turing, 'Synthetic Data Generation: Definition, Types, Techniques, and Tools' < https://www.turing.com/kb/synthetic-data-generation-techniques > accessed 20 February 2024; Datagen, 'Synthetic Data: The Complete Guide' < https://datagen.tech/guides/synthetic-data/synthetic-data/ > accessed 01 February 2024; Aryan Jadon and Shashank Kumary (n 30) 2; Mostly AI, 'What is Synthetic Data?' < https://mostly.ai/syntheticdata/whatissyntheticdata#:~:text=Synthetic%20data%20is%20generated%20by,create%20statistically%20identical%2C%20synthetic%20data > accessed 14 February 2024; Tshilidzi Marwala et al. (n 30), 3; Mikel Hernandez et al., 'Synthetic Data Generation for Tabular Health Records: A Systematic Review' (2022) 493 *Neurocomputing* 28; Joao Fonseca and Fernando Bacao (n 30); Abdul Majeed (n 32).

[35] Yingzhou Lu et al., (n 29) 14; Fida K. Dankar and Mahmoud Ibrahim (n 34) 2159; Sina Alemohammad et al (n 22) 4. Georgi Ganev (n 32) 2.

the fundamental right to information and the freedom of research in the context of AI development (Part II). In this section, the author emphasises the need for an adequate copyright exception which would balance copyright and other fundamental interests which are essential for the development of GenAI models and AI-related products in general. Further, the paper will analyse how providers of GenAI systems can leverage synthetic data to comply with copyright law while training their systems, and what risks might be associated with this approach (Part III). Finally, Part IV will provide concluding remarks on both solutions and emphasise key points drawing upon the insights provided in prior sections.

2. Legal solution to the conflict: focus on TDM exceptions

2.1. The right to train

AI developers train their systems on large training corpora of data that allows them to create GenAI models suitable for use in multiple applications across various domains.[36] To train AI algorithms, AI developers have to search for available content, then obtain and analyse (mine) it, and finally transmit new information, derived from processed data, to a user. To fully leverage the capabilities of GenAI systems and produce valuable and accurate results on demand (e.g. predictions, solutions etc.,) all types of training data are needed.[37] GenAI models learn how to generate new knowledge by imitating or mimicking such data.[38] However, training datasets often include works protected by copyright the use of which could raise legal concerns in the algorithmic society.[39] In theory, providers of GenAI models could access copyright-protected works to obtain information needed for training their systems without encountering any legal issues since copyright protects the form (expression) of the work and not the substance of the embedded ideas and information.[40] However, in practice, it is impossible to obtain

[36]Avanika Narayan et al (n 6) 738; Erik Brynjolfsson et al., 'Generative AI at Work' (2023) NBER Working Paper No. 31161, 5 < https://www.nber.org/system/files/working_papers/w31161/w31161.pdf > accessed 1 March 2024; Maryna Manteghi (n 28) 34-35; Artha Dermawan (n 1) 5; Van Lindberg (n 1) 2.

[37]Mauritz Kop, 'The Right to Process Data for Machine Learning Purposes in the EU' (2021) 34 *Harvard Journal of Law and Technology* 1, 3.

[38]Martin Senftleben (n 13)1535; Christophe Geiger, 'Elaborating a Human Rights friendly Copyright Framework for Generative AI' (2024) Joint PIJIP/TLS Research Paper Series 123 <https://digitalcommons.wcl.american.edu/research/123 >accessed 12 March 2024, 22; On the technical aspects of machine learning see Josef Drexl, Reto Hilty et al., 'Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective', Max Planck Institute for Innovation and Competition Research Paper No. 19–13 (2019), available on SSRN < https://ssrn.com/abstract=3465577 > accessed 11 January 2024.

[39]Philipp Hacker, 'A Legal Framework for AI Training Data -from First Principles to the Artificial Intelligence Act' (2021) 13(2) *Law, Innovation and Technology* 257, 259.

[40]The idea-expression and fact-expression dichotomy, that is to say in the postulate that copyright protects original expressions, whereas ideas, principles, procedures, facts and data as such are not protected' (Thomas Margoni and Martin Kretschmer, 'A Deeper Look into the EU Text and Data Mining

information about the contents of protected data in the course of TDM without using the form of works; simply put, a machine has to make copies of obtained content in order to analyse (mine) it.[41] In this sense, the protection of a broadly defined right of reproduction extends to the form (expression) of mined works, disrupting the balance between the interests of AI developers and copyright holders.[42] But would it be fair to claim that the training of AI algorithms breaches copyright only because GenAI models cannot learn from existing knowledge in the same way as humans do? The question is not easy to answer as for now no legislation or case law dealing exactly with these issues has been developed in the EU.[43]

The tension between copyright holders and providers of GenAI should not be considered in isolation, the issue needs evaluation in terms of constitutionally protected values.[44] The use of fundamental rights in shaping a balanced application of copyright norms refers to the 'constitutionalisation' of copyright law and its derivative concept, 'digital constitutionalism,' which deals with copyright and fundamental values in a particular context affected by technological advances.[45] It is argued that the framework of fundamental

Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2022) 71(8) *GRUR International* 685, 689; In this sense see, for instance, Christophe Geiger, 'Author's Right, Copyright and the Public's Right to Information: A Complex Relationship (Rethinking Copyright in the Light of Fundamental Rights)'in Fiona Macmillan (ed.), *New Directions in Copyright Law* (Vol. 5, Edward Elgar 2007) 24, 26; Dirk Voorhoof, 'Freedom of expression and the right to information: Implications for copyright' in Christophe Geiger (ed.), *Research Handbook on Human Rights and Intellectual Property* (Edward Elgar Publishing, 2015) 331, 336; Case C-833/18, of 11 June 2020, *Brompton Bicycle*, ECLI:EU:C:2020:461, at 27; Case C-683/17, of 12 September 2019, *Cofemel*, ECLI:EU:C:2019:721, at 29; Case C-393/09, *BSA*, of 22 December 2010, ECLI:EU:C:2010:816, at 49. E.g., Art. 2 WIPO Copyright Treaty and in Art. 9(2) WTO's TRIPs Agreements and Recital 8 CDSMD.

[41] Christophe Geiger (n 40) 26.

[42] Thomas Margoni and Martin Kretschmer (n 40) 695.

[43] As of the time of writing this article, there are some ongoing GenAI cases that could bring some clarity see, for instance, the German case ROBERT KNESCHKE v. LAION e.V (https://cepic.org/news/an-up-date-on-the-robert-kneschke-v-laion-e-v); In the US see for instance *Sarah Andersen et al. v. Stability AI et al* Case No. 3:23-cv-00201-WHO; *The New York Times Company v. Microsoft Corporation et al*., plaintiff's complaint, Case 1:23-cv-11195 (27December 2023); *Richard Kadrey et al. v. Meta Platforms, Case 3:23-cv-03417-VC* at 2. (N.D. Cal. 20 November 2023). P.M. v. GitHub Copilot, Inc., et al. the case was filed in the Northern District of California on June 28, 2023; Paul Tremblay and Mona Awad v. OpenAI LP, et al. (N.D. Cal., filed July 5, 2023); Sarah Silverman v. Meta Platforms, Inc. and OpenAI LP, et al. (N.D. Cal., filed July 7, 2023). Google Bard (*J.L. v. Alphabet Inc, U.S. District Court for the Northern District of California, No. 3:23-cv-03440*; In the UK see for instance *Getty Images (US) Inc. et al. v. Stability AI Ltd*., [2023] EWHC 3090 (Ch).

[44] Christophe Geiger and Bernd Justin Jütte, 'Designing Digital Constitutionalism: Copyright Exceptions and Limitations as a Regulatory Framework for Media Freedom and the Right to Information Online' in Martin Senftleben et al. (eds), *Cambridge Handbook of Media Law and Policy in Europe* (Cambridge University Press, forthcoming), < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4548510 > accessed 13 February 2024, 1-2.

[45] See Christophe Geiger, ''Constitutionalising' Intellectual Property Law? The Influence of Fundamental Rights on Intellectual Property in the European Union' (2006) 37 IIC 371; Christophe Geiger, 'Reconceptualizing the Constitutional Dimension of Intellectual Property – An Update' in Paul Torremans (ed), *Intellectual Property and Human Rights* (4edn, Kluwer Law International 2020); Christophe Geiger and Elena Izyumenko, 'The Constitutionalization of Intellectual Property Law in the EU and the Funke Medien, Pelham and Spiegel Online Decisions of the CJEU: Progress, but Still Some Way to

rights could be used to enable a more 'permissive and enabling' copyright regime needed in the age of AI.[46] Copyright which is recognised as a fundamental right under Article 17(2) of the EUCFR[47] is not an absolute right that could be protected at any cost.[48] The scope of copyright protection must be interpreted in the light of other fundamental rights (e.g. freedom of expression, the right to conduct a business etc.,) to strike a balance of interest between rightsholders and users of protected content.[49] The right to train AI systems may find its justification under the human rights framework.[50] In particular, the right could be derived from the rights to access, receive and impart information,[51] and the freedom of research originated from the right to information[52] and the freedom of the arts and sciences.[53] Access to massive amounts of data, including copyright-protected materials, is an essential precondition for the proper operation and development of AI.[54] Providers of GenAI models analyze digital data, in the course of training,

Go!' (2020) 51 IIC 282; Mylly, Tuomas, 'The New Constitutional Architecture of Intellectual Property', in Jonathan Griffiths, and Tuomas Mylly (eds), *Global Intellectual Property Protection and New Constitutionalism: Hedging Exclusive Rights* (Oxford, 2021; online edn, Oxford Academic, 23 Dec. 2021); Christophe Geiger and Bernd Justin Jütte (n 44) 6.

[46]Christophe Geiger and Bernd Justin Jütte (n 18) 143; Christophe Geiger and Elena Izyumenko, 'From Internal to External Balancing, and Back? Copyright Limitations and Fundamental Rights in the Digital Environment' in Julian Lopez and Conception Saiz Garcia (eds.) *Digitalización, acceso a contenidos y propiedad intelectual* (Madrid, Dykinson, 2022) 103, 105; Christophe Geiger (n 38) 7.

[47]Article 17(2) of the EUCFR, see (n 21).

[48]Maryna Manteghi (n 28) 457; Tito Rendas, 'Fundamental rights in EU copyright law: An overview' in Eleonora Rosati (ed.), *The Routledge Handbook of EU Copyright Law* (1 ed., Taylor and Francis – Balkema) 18, 23; Christophe Geiger (n 40) 32; Case C-469/17 *Funke Medien NRW GmbH v Bundesrepublik Deutschland* para 72; Case C-476/17 *Pelham GmbH and Others v Ralf Hütter and Florian Schneider-Esleben* para 32; Case C-314-12, *UPC Telekabel Wien GmbH v. Constantin Film Verleih GmbH, Wega Filmproduktionsgesellschaft mbH* [2014], Judgment of the Court (Fourth Chamber) of 27 March 2014, para. 61;

[49]CJEU, Case C-70/10, *Scarlet Extended NV v. Belgische Vereniging van Auteurs, Compon- isten en Uitgevers CVBA (SABAM)* [2011], Judgment of the Court (Third Chamber) of 24 November 2011, para. 43, 44, 45 ECR I-11959; CJEU, Case C-314-12, *UPC Telekabel Wien GmbH v. Constantin Film Verleih GmbH, Wega Filmproduktionsgesellschaft mbH* [2014], paras 47 and 55–56. *Promusicae v Telefónica de España SAU* para 70, Case C-275/06 2008, ECLI:EU:C:2008:54.

[50]Christophe Geiger (n 38) 2.

[51]The right to information constitutes the distinct elements of the freedom of expression under Art. 11 of the EUCFR. The provision is silent regarding the rights to seek or search for and access information; however, with the rapid development of digital technologies and AI the existence and scope of such freedoms have been intensively discussed. Moreover, the right has been explicitly recognized under Art. 19 of the International Covenant on Civil and Political Rights (ICCPR)). For a more detailed analysis of the scope of the right to information see, for instance, Steve Peers et al. (n 18) 348-349, 366-367; Christophe Geiger and Bernd Justin Jütte (n 18) 29–35; Christophe Geiger and Bernd Justin Jütte (n 18) 160-162; ECtHR, *Társaság a Szabadságjogokért v. Hungary*, no. 37374/05, 14 April 2009, ECtHR, *Kenedi v. Hungary*, no. 31475/05, 26 May 2009; ECtHR, *Youth Initiative for Human Rights v. Serbia*, no. 48135/06, 25 June 2013, paras 20 and 24; CJEU, Case T-545/11, *Stichting Greenpeace Nederland and PAN Europe* [2013], Judgment of the General Court of 8 October 2013.

[52]Art. 11 of the EUCFR.

[53]Art. 13 of the EUCFR.

[54]Christophe Geiger and Vincenzo Iaia, 'Generative AI, Digital Constitutionalism and Copyright: Towards a Statutory Remuneration Right Grounded in Fundamental Rights' (*MediaLaws,* 19 October 2023) < https://www.medialaws.eu/generative-ai-digitalconstitutionalismandcopyrighttowardsastatutoryremuneration-right-grounded-in-fundamental-rights/ > accessed 11 January 2024.

with the intention to impart hidden knowledge, derived from the underlying ideas and data (e.g. correlations, patterns, new insights and facts etc.,) to the public.[55] Simply put, AI developers make existing information better accessible and understandable that could effectively address public needs in all spheres of human endeavour.[56] This is in line with the right to information which seeks to 'ensure and protect the free flow of information and ideas'.[57] TDM which involves various data analysis techniques (e.g. categorisation, classification, clustering visualisation etc.,)[58] has become a fundamental tool for research.[59] The outputs generated from TDM research often contribute to the advancement of science in either practical or theoretical domains[60] or can be used by users (authors) to explore new forms of cultural (artistic) expression that could benefit the public in general.[61] In this sense, the use of copyright-protected data for training AI algorithms is needed to ensure that novel creative outputs are not based on outdated, inaccurate or incomplete data.[62] In this regard, the value of GenAI systems to support creative activities, access to information and scientific research should not be underestimated.[63]

Against this background, imposing unreasonable limitations on the acts of reproductions, occurring in the course of TDM, could limit access to the very ideas and facts underlying copyright-protected works, thus leading to interference with the rights to information and research.[64] To facilitate these fundamental values in the context of AI training it is needed to ensure that copyright supports and enables TDM and does not create chilling effects through excessive restrictions and legal uncertainty.[65] This does not imply that data must always be accessed and used for free

[55]Christophe Geiger (n 38) 22; Dirk Voorhoof (n 40) 337; In this sense see ECtHR, *Youth Initiative for Human Rights v. Serbia*, no. 48135/06, 25 June 2013, paras 20 and 24; CJEU, Case T-545/11, *Stichting Greenpeace Nederland and PAN Europe* [2013], Judgment of the General Court of 8 October 2013.

[56]Christophe Geiger (n 38) 8; Maryna Manteghi (n 28) 443-444; Maryna Manteghi (n 28) (n 28) 34. It is important to notice that AI output cannot be protected by copyright see, for instance, Christophe Geiger (n 38) 15-17; Christophe Geiger and Bernd Justin Jütte (n 44) 5.

[57]Stijn van Deursen and Thom Snijders, 'The Court of Justice at the Crossroads: Clarifying the Role for Fundamental Rights in the EU Copyright Framework' (2018) 49 IIC 1080, 1081.

[58]Christophe Geiger and Bernd Justin Jütte (n 18) 142; for a general overview of TDM techniques and methods see, for instance, Jiawei Han et al., *Data Mining: Concepts and Techniques* (3rd edn, Elsevier 2012).

[59]Christophe Geiger, 'The Missing Goal-Scorers in the Artificial Intelligence Team: Of Big Data, the Fundamental Right to Research and the failed Text and Data Mining Limitations in the CSDM Directive' (2021) PIJIP/TLS Research Paper Series No.66, 4.

[60]Christophe Geiger (n 38) 9; Steve Peers et al (n 51) 420–421.

[61]Christophe Geiger and Vincenzo Iaia (n 54).

[62]Thomas Margoni and Martin Kretschmer (n 40) 687-689.

[63]Christophe Geiger and Vincenzo Iaia (n 54).

[64]Sean Flynn et al., 'Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action' (2020) Joint PIJIP/TLS Research Paper Series No.48,, <https://digitalcommons.wcl.american.edu/research/48/?utm_source=digitalcommonswcl.american.edu%2Fresearch%2F48&utm_medium=PDF&utm_campaign=PDFCoverPages > accessed 11 January 2024, 5.

[65]Christophe Geiger and Bernd Justin Jütte (n 44) 5-6.

(especially for a commercial purpose) but that the content 'can be accessed without undue hurdles'.[66] In this regard, copyright exceptions and limitations could be seen as 'essential regulatory tools' the application of which may ensure the proper interaction between copyright protection and other fundamental freedoms.[67] Therefore, the following section will discuss the limits placed on copyright protection in the context of TDM and their ability to 'pacify' the conflict at hand.

2.2. TDM exceptions as a solution: a critical analysis

The trend towards constitutionalisation of copyright law is characterised by a more liberal interpretation of exceptions and limitations read in terms of fundamental freedoms protected under the EUCFR and the EUCHR.[68] Exceptions and limitations constitute an integral part of a balanced copyright framework in the EU and could serve as essential safeguards for the users' fundamental rights in the digital environment.[69] The TDM exceptions in Arts. 3 and 4 of the CDSM Directive seek to achieve that balance by providing 'improved' access to copyright-protected materials for the purpose of automatic data analysis.[70] The permitted uses emphasise the need to harness *de facto* control exercised by copyright holders over digital data that could create 'breathing room' for AI developers.[71] This could facilitate technological developments while securing the fundamental right to information, the freedom of research, and the public interests in general.[72] However, the usefulness of these exceptions may be diluted by some limitations that could impede crucial developments within the European AI

[66]Christophe Geiger and Bernd Justin Jütte (n 44) 17.
[67]Christophe Geiger and Bernd Justin Jütte (n 44) 3-4.
[68]Kalpana Tyagi, 'Copyright, text & data mining and the innovation dimension of generative AI' Forthcoming in *Journal of Intellectual Property Law & Practice*, 2024 < https://academic.oup.com/jiplp/advance-article/doi/10.1093/jiplp/jpae028/7624901 > accessed 27 March 2024, 17. See also CJEU, Judgment in *Funke Medien NRW GmbH v. Bundesrepublik Deutschland*, C-469/17, 29 July 2019, EU: C:2019:623; CJEU, Judgment in *Pelham GmbH and Others v. Ralf Hütter and Florian Schneider-Esleben*, C-476/17, 29 July 2019, EU: C:2019:624; and CJEU, Judgment in *Spiegel Online GmbH v. Volker Beck*, C-516/17, 29 July 2019, EU: C:2019:625. For commentary on these decisions, see also Thomas Dreier, 'The CJEU, EU Fundamental Rights and the Limitations of Copyright' (2020) 69 *GRUR International* 223–224; Tito Rendas (n 48) 25.
[69]CJEU, Judgment in *Funke Medien NRW GmbH v. Bundesrepublik Deutschland*, C-469/17, 29 July 2019, EU: C:2019:623 paras. 51, 58, 70; CJEU, Judgment in *Pelham GmbH and Others v. Ralf Hütter and Florian Schneider-Esleben*, C-476/17, 29 July 2019, EU: C:2019:624 para. 60; CJEU, Judgment in *Spiegel Online GmbH v. Volker Beck*, C-516/17, 29 July 2019, EU: C:2019:625 para. 36.; Christophe Geiger and Elena Izyumenko (n 46); Christophe Geiger and Bernd Justin Jütte (n 20); Christophe Geiger and Bernd Justin Jütte (n 44) 2.
[70]Recital 6 of the CDSM Directive provides that TDM exceptions 'seek to achieve a fair balance between the rights and interests of authors and other rightsholders, on the one hand, and of users on the other'. See also Sean Flynn et al., (n 64) 2.
[71]Mauritz Kop (n 37) 15; Christophe Geiger and Bernd Justin Jütte (n 44) 4.
[72]Oreste Pollicino et al., 'Cultural Rights, Cultural Diversity and the EU's Copyright Regime: the Battlefield of Exceptions and Limitations to Protected Content' in Oreste Pollicino (ed.) *Copyright and Fundamental Rights in the Digital Age* (Edward Elgar Publishing Limited, 2020) 124, 141–142.

sector.[73] Both exceptions permit the use of protected works for the purpose of text and data mining only for those actors who have lawful access to content.[74] Moreover, Art. 3 of the CDSM Directive has been formulated narrowly as it applies merely to few beneficiaries such as research organisations and cultural heritage institutions carrying out TDM for the purpose of scientific research.[75] In this sense, the provision would be irrelevant in the case of GenAI as these AI models are normally developed by private actors (e.g. start-ups, SMEs, and other businesses) that often pursue commercial purposes.[76] This limitation would place the EU AI sector at a significant disadvantage as AI developers would be required to negotiate licences over huge masses of data needed for training their systems, which is a complex, time-consuming and costly process.[77] This would specifically affect smaller AI players with low human and financial resources.[78] Against this background, restricting access to information for private actors in the terms of Art. 3 of the CDSM Directive could result in interference with the fundamental rights to information and research, which are not limited to particular types of data, research activity or users.[79]

Some concerns could also be raised regarding a so-called 'commercial' TDM exception of Art. 4, which allows any entity to carry out TDM for any purpose, as the effectiveness of this provision may be diluted by the reservation right which permits rightsholders to opt out of the exception.[80] Recital 18 clarifies that those who have rights over works publicly available online could reserve their rights by using machine-readable means, including metadata and terms and conditions of a website or a service. In other cases, it should be appropriate to reserve the rights by other means, such as contractual agreements or unilateral declarations.[81] The introduction of an 'opt-out' mechanism under the CDSM Directive was meant to 'harness' the specific TDM exception to ensure the balance of rightsholders' interests and those

[73]Christophe Geiger and Bernd Justin Jütte, (n20); Christophe Geiger (n 38) 28.
[74]Arts. 3 and 4 (1) of the CDSM Directive. For more discussions on the 'lawful access' requirement see, for instance, Maryna Manteghi (n 28) 451-453; I*nfopaq International A/S v Danske Dagblades Forening* (C-302/10) EU:C:2012:16 para 58; *Football Association Premier League* (C-403/08 and 429/08) EU:C:2011:631; [2012] 1 C.M.L.R. 29 para 168 provides that 'a use should be considered lawful where it is authorised by the right holder or where it is not restricted by the applicable legislation.'
[75]Art. 3 (1) of the CDSM Directive.
[76]Thomas Margoni and Martin Kretschmer (n 40) 688.
[77]For more critical discussions on the limitations of Art. 3 of the CDSM Directive see, for instance, Maryna Manteghi, (n 28) 448-457; Maryna Manteghi, 'The Insufficiency of the EU's Text and Data Mining Exceptions for Using Artificial Intelligence' (2022) 44 E.I.P.R. 652, 660-663.
[78]Maryna Manteghi (n 28) 449.
[79]Steve Peers et al. (n 18) 344–345; Thomas Margoni and Martin Kretschmer (n 40) 695; Maryna Manteghi (n 28) 448-450.
[80]Art. 4 of the CDSM Directive; Maryna Manteghi (n 28) 453-454; Christophe Geiger (n 38) 26.
[81]Art. 4 (3) and recital 18 of the CDSM Directive.

of TDM users.[82] The intention was to prevent the over-use of copyright-protected works; however, in practice, the reservation of rights has only further strengthened the position of copyright holders who are already allowed to control the use of TDM through licensing terms.[83] This could affect the AI developers' right to training as they would have to pay twice to be able to mine – first to acquire 'lawful access' to works and a second time to read and analyse training data.[84] Moreover, the possibility to opt out may allocate the power and control over the AI market in the hands of very large AI players significantly affecting the capability of small tech companies with limited financial resources to innovate and develop AI-driven products.[85] However, the scope of the exception is not the only issue. It is needed to clarify how an 'opt-out' mechanism would work in practice, particularly since there are still questions regarding the requirements of the reservation right that the use should be reserved 'expressly' and 'in an appropriate manner'.[86] The passive and unclear nature of an 'opt-out' mechanism could potentially limit access to data needed for the development of cutting-edge AI applications by private actors and thus reduce the research power of the EU.[87] As a result, Art. 4 of the CDSM Directive would create a chilling effect on the right to information and the freedom of research through the limitations dictated by the commercial interests of copyright holders and legal uncertainty.[88]

Nevertheless, the recent advancements in machine learning (ML) force us to take prompt actions to find an optimal solution that would alleviate

[82]Recital 6 of the CDSM Directive; Andres Guadamuz, 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs' (2024) 73(2) *GRUR International* 111, 120.

[83]Andrew Tyner, 'The EU Copyright Directive: 'Fit for the Digital Age' or Finishing It?' (2020) 26 J.I.P.L. 275, 281; Maryna Manteghi (n 28) 454; Bernt Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)' (*Kluwer Copyright Blo*, 24 July 2019) < https://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-datamining-articles-3-and4/ > accessed 11 March 2024; Maryna Manteghi,(n 28) 40. The exception of Art. 4 can be overridden by a contract see Art. 7 (1) of the CDSM Directive.

[84]LIBER, 'Myths and Misunderstandings about Text and Data Mining in the Copyright Reform' (2017) <http://eare.eu/assets/uploads/2017/11/Myths-and-misunderstandings-about-TDM.pdf > accessed 11 January 2024.

[85]Gina Maria Ziaja, 'The Text and Data Mining Opt-out in Article 4(3) CDSMD: Adequate Veto Right for Rightholders or a Suffocating Blanket for European Artificial Intelligence Innovations?' (2024) 19(5) *Journal of Intellectual Law and Practice* 453, 454–453. Maryna Manteghi (n 28) 454.

[86]Ibid

[87]Maryna Manteghi (n 28) 40; Maryna Manteghi (n 28) 449; Christophe Geiger and Bernd Justin Jütte (n 18) 150; Sean Flynn et al. (n 64) 9.

[88]Christophe Geiger, 'The Missing Goal-Scorers in the Artificial Intelligence Team: Of Big Data, the Fundamental Right to Research and the Failed Text and Data Mining Limitations in the CSDM Directive' (2021) PIJIP/TLS Research Paper Series No. 66 < https://digitalcommons.wcl.american.edu/research/66 > accessed 12 January 2024, 9; Christophe Geiger et al., 'Crafting a Text and Data Mining Exception for Machine Learning and Big Data in the Digital Single Market' in Xavier Seuba, Christophe Geiger and Julien Pénin (eds.), *Intellectual Property and Digital Trade in the Age of Artificial Intelligence and Big Data* (A CEIPI-ICTSD, Issue 5, 2018) 95; Bernt Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)' (*Kluwer Copyright Blog*, 24 July 2019) < http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ > accessed 20 February 2024; Thomas Margoni and Martin Kretschmer (n 40) 685; Christophe Geiger and Bernd Justin Jütte (n 44) 13.

growing copyright-related concerns in the field of GenAI and prevent the development of AI 'black box' systems.[89] Some academics suggest implementing a remuneration right for creators of works used in AI training as an alternative to the reservation right of Art. 4 (3) of the CDSM Directive.[90] As Geiger and Izyumenko claim, the remunerated right to train AI systems could mitigate the density of the conflict between freedom of expression and copyright holders' fundamental right to IP.[91] However, many scholars argue that it is needed to reconsider Art. 4 of the CDSM Directive in the context of GenAI and adopt generally accepted protocols or standards which would clarify the work of an 'opt-out' mechanism.[92] Recently, AI companies have started to adopt their own 'opt-outs,' which could or could not fit into the statutory concept of rights reservation.[93] One of the possible options, which rightsholders may exercise to restrict the use of their works for training generative AI models, could be the Robots Exclusion Protocol (a so-called robots.txt file).[94] Robots.txt file has served as a simple and low-cost solution for controlling search engines' access to online content for about 30 years, and website owners still widely use this tool to prevent their content from being indexed by web crawlers.[95] However, the exclusion protocol could be less effective when utilised in the context of GenAI. To effectively protect websites from AI crawlers, rightsholders should inform web bots in the robots.txt file which resources may or may not be accessed and scraped for TDM purposes. This means that they should allow or disallow

[89] AI 'black boxes' is the term which describes an 'automated decision-making tool (e.g. AI) which makes decisions in ways that are not intelligible or transparent to humans' (Thomas Margoni and Martin Kretschmer (n 40) 688).

[90] See, for instance, Christophe Geiger (n 38); Martin Senftleben, (n 13); Christophe Geiger and Vncenzo Iaia, 'The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI' (2023) 52 *Computer Law & Security Review* (forthcoming, 2024) < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4594873 > accessed 23 March 2024.

[91] Christophe Geiger and Elena Izyumenko, 'Towards a European 'Fair Use' Grounded in Freedom of Expression' (2019) 35(1) *American University International Law Review* 1, 58; Christophe Geiger (n 38) 1.

[92] Paul Keller, 'Generative AI and Copyright: Convergence of Opt-Outs?' ((*Kluwer Copyright Blog*, 23 November 2023) < https://copyrightblog.kluweriplaw.com/2023/11/23/generative-ai-and-copyright-convergence-of-opt-outs/ > Accessed 23 March 2024; Nicola Lucchi (n 10) 15; Paul Keller and Zuzanna Warso, 'Defining Best Practices for Opting Out of ML Training' *(Open Future* 28 September 2023) < file:///Users/maryna/Desktop/Main%20Article/6opt%20out/defining-best-practices-for-opting-out-of-ml-training.html > accessed 13 February 2024; Péter Mezei, 'A saviour or a dead end? Reservation of rights in the age of generative AI' (2024) SSRN < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4695119 > accessed 20 March 2024, 3.

[93] See for more discussions, for instance, Péter Mezei (n 92) 9-11.

[94] robot.txt is an exclusion mechanism which informs web robots or crawlers about which pages or files of the website should not be processed (OddyTech, 'About /robots.txt' < http://www.robotstxt.org/robotstxt.html > accessed 12 March 2024); In this sense see Martin Senftleben (n 13) 10-11; Chyan Yang and Hsien-Jyh Liao, 'Using the Robots.txt and Robots Meta tags to implement online copyright and a related amendment' (2010) 28(1) *Library Hi Tech* 94, 97; Bernt Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)' (n 88); Google Search Central, 'Introduction to robots.txt' < https://developers.google.com/search/docs/crawling-indexing/robots/intro > accessed 12 March 2024.

[95] M Carl Drott, 'Indexing aids at corporate websites: the use of robots.txt and META tags' (2002) 38 *Information Processing & Management* 209, 212.

access to huge amounts of online content including specific HTML pages, updated or refined web pages, posts, files, CSS documents,[96] images, feeds, email addresses and others. However, a robots.txt file is limited in size which inherently protects this system from 'out-of-memory scenarios'.[97] Therefore, website owners may not be able to include detailed rules for multiple resources in the exclusion protocol to protect their online content from a mass 'invasion' by AI crawlers.[98] Moreover, the protocol is a voluntary measure which could be easily ignored or skipped by 'careless' AI trainers 'without actively forcing any digital fence'.[99] In other words, robots.txt does not lock or protect access to online resources *per se*, it is AI crawlers which should check the rules in a robots.txt file to determine whether they are allowed or disallowed to mine (certain parts of) a website and then choose to either follow the instructions or simply ignore them.[100] However, even if web bots respect the rules embedded in an exclusion protocol, thereby refraining from copying restricted content, they may still be able to access protected recourses from other websites not using robots.txt but embedding links to disallowed content.[101] Against this background, currently, providers of GenAI models are actively exploring alternative and more practical solutions to address the challenges arising from the mining of protected content. For instance, the SPRAWNINGai.txt website[102] allows rightsholders to create an *ai.txt* file, by using Spawning's *ai.txt* generator, to restrict or permit access to their content for AI training. Developers of SPRAWNINGai.txt indicate that AI crawlers would read the rules set in *ai.txt* files when they download content from a website. This contrasts with the work of a robots.txt protocol which is normally read when bots crawl a web resource. In addition, they promise real-time content protection that means that AI trainers would verify permissions established in *ai.txt* files even if they download protected content from websites merely containing a link to original materials.[103] HaveIBeenTrained website allows copyright holders to search for their works in training dataset LAION-5B, the dataset that, for example, Stability AI's Stable

[96]Cascading Style Sheet (CSS) is a style sheet language used for formatting content in HTML webpages < https://techterms.com/definition/css > accessed 12 January 2024.
[97]Martijn Koster et al., 'RFC 9309 Robots Exclusion Protocol' (*RFC*, September 2022) < https://www.rfc-editor.org/rfc/rfc9309.html#KiB > accessed 12 January 2024.
[98]Google Search Central, 'Introduction to robots.txt' (n 94).
[99]Rossana Ducato and Alain M. Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to 'Machine Legibility' (2019) 50(6) IIC 649, 675; Darren Ang, 'The Web Scraper's World of Copyright Exceptions and Contractual Overrides' (2021) 13(22) *Singapore Law Review* 5.
[100]M Carl Drott (n 95) 212; Martijn Koster et al (n 97); Google Search Central (n 94); Darren Ang (n 99) 5.
[101]Cullen Miller, 'Ai.txt: A New Way for Websites to Set Permissions for AI' (*Spawning Blog*, 30 May 2023) < https://spawning.substack.com/p/aitxt-a-new-way-for-websites-to-set > accessed 30 March 2024; Google Search Central (n 94).
[102]See <https://site.spawning.ai/spawning-ai-txt> accessed 10 January 2024.
[103]Ibid; Cullen Miller (n 101).

Diffusion is trained on, and opt them out.[104] Google-Extended, available through robots.txt, allows rightsholders to control access to content on their websites by informing web bots which resources could or could not be accessed and scraped for TDM purposes.[105]

Against this background, devising specific rules for the reservations of rights would eliminate the current fragmentation of 'opt-out' standards offered by AI developers and ensure that rightsholders use standardised ways for excluding their works from being used as training data.[106] Simply put, a standardised 'opt-out' mechanism would apply to all uses falling within the scope of Art. 4 (3) of the CDSM Directive, unlike model-specific 'opt-outs,' which require the use of the form specified by each AI company engaged in training activities.[107] In this sense, the Artificial Intelligence Act (AI Act)[108] could clarify certain aspects of the reservations of rights and thus facilitate the work of an 'opt-out' mechanism.[109] In particular, the AI Act imposes some obligations on providers of GenAI models: first, to comply with 'Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790,'[110] and second, to 'draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model … '.[111] The provisions could strengthen the 'commercial' copyright exception,[112] especially the transparency obligation may on the one hand, protect AI developers from claims of copyright infringement and on the other, allow authors and other rightsholders to make well-informed decisions regarding the reservation of their rights.[113] Creating

[104]See <https://haveibeentrained.com > accessed 10 January 2024.
[105]Danielle Romain, 'An Update on Web Publisher Controls' (Google blog, 28 September 2023) < https://blog.google/technology/ai/an-update-on-web-publisher-controls/ > accessed 28 January 2024; Google Search Central, 'Overview of Google crawlers and fetchers (user agents)' < https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers > accessed 01 March 2024.
[106]Paul Keller (n 92).
[107]Ibid.
[108]Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L 2024/1689, 12.7.2024.
[109]Martin Senftleben (n 13) 14.
[110]Art. 53 (1) (c) of the AI Act (n 108).
[111]Art. 53 (1) (d) of the AI Act (n 108).
[112]Martin Senftleben (n 13) 14.
[113]Communia, 'Policy Paper #15 on using copyrighted works for teaching the machine' (*Communia*, 26 April 2023) < https://communia-association.org/policy-paper/policy-paper-15-on-using-copyrighted-works-for-teaching-the-machine/ > accessed 10 February 2024; Authors and Performers Call for Safeguards Around Generative AI in the European AI Act (InitiativeUrheberrecht, 19 April 2023) < https://urheber.info/media/pages/diskurs/call-for-safeguards-around-generative-ai/c93a5ab197-1681904353/final-version_authors-and-performers-call-for-safeguards-around-generative-ai_19.4.2023_12-50.pdf > accessed 1 March 2024; Shabbir Merali and Ali Merali, 'The Generative AI Revolution Opportunities,

an appropriate reference to data sources is equally important for human writers and GenAI models since it concerns the legitimacy of data usage.[114] Ideally, AI trainers should maintain logs of detailed records of mined works, and provide rightsholders with efficient means for determining the availability of protected works, metadata or links to data, in existing training databases.[115] However, the approach to source citation employed by AI trainers differs from how human writers acknowledge the sources.[116] Many authors have argued that requiring AI developers to itemise and attribute all copyrighted materials used for training purposes might not be practical.[117] It appears that the AI Act has addressed these concerns by clarifying that providers of GenAI would not be required to open an exhaustive list of all training data, but they would have to provide a comprehensive summary, for instance, by 'listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used'.[118] The provisions would definitely play an important role in balancing competing fundamental rights affected by the training process.

The refined 'opt-out' mechanism of Art. 4 (3) of the CDSM Directive would provide more legal certainty for both groups of stakeholders, but it would also strengthen the rightsholders-orientated approach of the exception. Allocating the right to allow or prohibit the use of data for mining to authors and other rightsholders could undermine the AI developers' right to access lawfully obtained information,[119] and conduct research on ideas

Shocks, and Risks' (2023) < https://www.ukonward.com/wp-content/uploads/2023/05/Generative-AI-Revolution-Final.pdf > accessed 1 March 2024, 46. For more discussions regarding the benefits of the transparency requirement see e.g., Leander Nielbock and Teresa Nobre, 'The AI Act and the Quest for Transparency' (*Communia*, 28 June 2023) < https://communia-association.org/2023/06/28/the-ai-act-and-the-quest-for-transparency/ > accessed 2 March 2024; Thomas Margoni and Martin Kretschmer (n 40) 693-697; Martin Senftleben, Thomas Margoni et al., 'Ensuring the Visibility and Accessibility of European Creative Content on the World Market – The Need for Copyright Data Improvement in the Light of New Technologies and the Opportunity Arising from Article 17 of the CDSM Directive' (2022) 13 JIPITEC 67, 74, 82; Martin Senftleben (n 13) 12-19.

[114]Nicola Lucchi (n 10) 16.

[115]Matthew Sag (n 9) 341.

[116]Nicola Lucchi (n 10) 16-17.

[117]Nicola Lucchi (n 10) 16. The rationale behind this, as Quintais fairly argues, can be found in 'the low threshold of originality, the territorial fragmentation of copyright and its ownership, the absence of a registration requirement for works, and in general the poor state of rights ownership metadata' Joao Pedro Quintais, 'Generative AI, Copyright and the AI Act' (*Kluwer Copyright Blog*, 9 May 2023) < https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/ > accessed 12 February 2024; Christophe Geiger (n 38) 24-25.

[118]Recital 107 of the AI Act (n 108); Gina Maria Ziaja (n 85) 458.

[119]Copyright holders could rely on the 'lawful access' requirement of Art. 4 (1) of the CDSM Directive to receive adequate compensation for the use of their works through licensing, subscriptions or other lawful means. In this sense, it seems unreasonable to claim that authors could become less motivated to produce new content since they would have less control over their works without the possibility to opt out. (Maryna Manteghi (n 28) 40.

and facts rooted in such works.[120] These fundamental interests, which are essential for the development of GenAI models and AI-related products in general, have to be protected through adequate copyright exceptions. Any limitations on the enjoyment of such interests could impair the fundamental right to information and the freedom of research.[121] The purpose of GenAI training is not to make exact copies of protected content but to generate new data (output) through analysis of such data.[122] The incidental (indirect) reproductions (both temporary and more permanent) made in the course of TDM would not undermine the primary purpose behind copyright protection which is to prohibit unauthorised copying that could substitute for the author's work.[123] In this regard, it is needed to establish an adequate legal framework for the use of training data for ML purposes[124] in the form of a broader TDM exception not including the reservations of rights. This would increase access to information, and facilitate research and knowledge production in the EU.[125] This would ensure that copyright exceptions align with fundamental rights and facilitate such values in the digital environment.[126] As Margoni and Kretschmer fairly claim, the CDSM Directive's TDM exceptions go far beyond mere copyright exceptions, they should be viewed as 'a property-right approach to the regulation of AI'.[127] In this sense, the explicit reference to Art. 4 of the CDSM Directive in the AI Act[128] and the approach taken by the European Commission on AI and copyright[129] may dispel doubts about the capability of the provision to regulate ML performed by GenAI systems.[130] The refined 'commercial' TDM exception should allow developers of GenAI models, with lawful access to data, to receive the information needed for TDM research without restrictions that would enable the development and optimise the operation of game-changing ML applications.[131] The AI-friendly provision could serve innovation and creativity that would

[120]Maryna Manteghi (n 28) 454; Thomas Margoni and Martin Kretschmer (n 40) 689; Sean Flynn et al (n 64) 5.
[121]Sean Flynn et al (n 64) 5.
[122]Mira T. Sundara Rajan, 'Is Generative AI Fair Use of Copyright Works? NYT v. OpenAI' (*Kluwer Copyright Blog* 29 February 2024) < https://copyrightblog.kluweriplaw.com/2024/02/29/is-generative-ai-fair-use-of-copyright-works-nyt-v-openai/ > accessed 2 March 2024.
[123]Sean Flynn et al (n 64) 4; Christophe Geiger et al., 'Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?' (2018) 49 (7) IIC 814, 817.
[124]Mauritz Kop (n 37) 3; Kalpana Tyagi (n 68) 18.
[125]Maryna Manteghi (n 28) 40.
[126]Thomas Margoni and Martin Kretschmer (n 40) 688, 695.
[127]Ibid 688.
[128]Art. 53 (1) (c) of the AI Act (n 108).
[129]European Commission, 'Answer given by Mr Breton on behalf of the European Commission' (2023) < https://www.europarl.europa.eu/doceo/document/E-9-2023-000479-ASW_EN.html > accessed 4 March 2024.
[130]Christophe Geiger (n 38) 27-28; Péter Mezei (n 92) 8.
[131]Sean Flynn et al (n 64) 11.

put the EU at a significant competitive advantage in regard to AI development and research.[132]

3. Technical solution to the conflict: focus on synthetic data

Nowadays, we are witnessing a shift of focus in the field of AI from real data to AI-synthesised data.[133] Access to large, diverse and high-quality datasets is a prerequisite for successful AI training and the generation of accurate and reliable outputs.[134] AI developers could use synthetic data, regardless of whether it is generated from real-world data or not, to augment training datasets with minimal effort and at reduced costs, especially when data is not available, scarce or biased.[135] The shortage of resources may occur when, for instance, research in a particular field is incomplete, when access to materials is restricted for different reasons (e.g. privacy and data protection), or simply when the reserves of real data needed for training are (almost) depleted.[136] Further, it is argued that synthetic data may solve the issues of privacy and data protection as it does not contain 'personally identifiable information' which could link back to real individuals.[137] Some authors emphasise the effectiveness of the approach compared to an anonymization technique[138] the use of which does not completely remove the possibility of re-identification of a natural person.[139] The use of synthetic data in the healthcare field is increasing in popularity as synthetic data can be generated from actual 'de-identified' data (when personal information is removed or altered) that would protect patients' privacy.[140] For instance, in recent studies on synthetic cancer data, the EU Commission's Joint Research Center has found that synthetic cancer patient records generated from real-world cancer records could be used as effective input data to train GenAI systems since they demonstrate a high level of realism while

[132]In this sense see Andres Guadamuz (n 82) 120-121; Christophe Geiger (n 58) 11.
[133]Sina Alemohammad et al (n 32) 2; Yingzhou Lu et al (n 29) 3.
[134]Yingzhou Lu et al (n 29) 1; Aryan Jadon and Shashank Kumar (n 30) 2-3.
[135]Aryan Jadon and Shashank Kumar (n 30) 2; Abdul Majeed (n 32) 639; Erroll Wood et al (n 32) 3682; Sina Alemohammad et al (n 32) 2; Joao Fonseca and Fernando Bacao (n 30) 6-7; Yingzhou Lu et al. (n 29) 1; Tshilidzi Marwala et al (n 30), 4; James Jordon et al., 'Synthetic Data – What, Why and How?' (2022) < https://arxiv.org/abs/2205.03257 > accessed 23 February 2024, 6, 30.
[136]Sina Alemohammad et al (n 32) 2.
[137]Tshilidzi Marwala et al (n 30), 5; For more discussions see, for example, Fida K. Dankar and Mahmoud Ibrahim (n 34) 2159; Mikel Hernandez et al (34); Joao Fonseca and Fernando Bacao (n 30); Abdul Majeed (n 32); Yingzhou Lu et al (n 29) 4; Aryan Jadon and Shashank Kumar (n 30) 3.
[138]Fida K. Dankar and Mahmoud Ibrahim (n 34) 2159.
[139]Ibid.
[140]Tshilidzi Marwala et al (n 30) 5; Fida K. Dankar and Mahmoud Ibrahim (n 34) 2159; European Medicines Agency, 'Draft reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle' (13 July 2023) < https://www.ema.europa.eu/en/documents/scientific-guideline/draft-reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf > accessed 12 March 2024.

complying with privacy requirements.[141] Moreover, the use of synthetic data may improve the quality of training data which is also critical for the development of GenAI models. If AI developers have to rely only on public domain or open-source works to train their systems, the final output could be scarce and insecure.[142] If the sources are incomplete, training datasets could lack diversity and not be well-aligned with various complex processing tasks that AI-based systems are aimed to perform.[143] Furthermore, it is argued that the application of training methods to synthetic data may even have a better effect on the overall performance of GenAI models.[144] Against this background, synthetic data could serve as an alternative to real-world data, enabling AI companies to develop their systems with fewer manipulations or interactions performed on actual data.[145] Even though synthetic data could be the key enabler for AI,[146] using it for AI training may entail many technical risks (e.g. bias propagation, imbalances, data pollution) as the approach is still underexploited and underdeveloped.[147] Therefore, it is expected that the 'quality, diversity, and fidelity of synthetic data' will improve shortly, enabling a more accurate and versatile application of the approach to GenAI training.[148]

Considering copyright, it could be difficult to assert with certainty that synthetic data generation is the right approach to addressing a conflict between rightsholders and developers of GenAI models. The use of synthetic data which is not generated from actual data (but created by using existing models or background knowledge of data analysts)[149] to train AI systems should not, in theory, pose any risks associated with copyright protection

[141]Jiri Hradec et al., 'Multipurpose Synthetic Population for Policy Applications', EUR 31116 EN, Publications Office of the European Union, Luxembourg, 2022, ISBN 978-92-76-53478-5 < https://publications.jrc.ec.europa.eu/repository/handle/JRC128595 > accessed 23 March 2024, 9.

[142]Rossana Ducato and Alain M. Strowel (n 99) 668; Tshilidzi Marwala et al (n 30) 3.

[143]Khaled El Emam (n 30) 6; Michal S Gal and Orla Lynskey, 'Synthetic Data: Legal Implications of the Data-Generation Revolution' (2024) 109 *Iowa L Rev*. 1087, 1089–1090. More on the work of GenAI systems see, for instance, Rehan Ahmed Khan et al., 'ChatGPT-Reshaping Medical Education and Clinical Management' (2023) 39 *Pak J Med Sci* 605, 605; Brady D Lund and Yi-Shun Wang, 'Chatting about ChatGPT: How May AI and GPT Impact Academia and Libraries?' (2023) 40 *Library Hi Tech News* 26, 27.

[144]Sina Alemohammad et al (n 32) 2; Aryan Jadon and Shashank Kumar (n 30) 3; Erroll Wood et al (n 32) 3687; Adam Zewe, 'In machine learning, synthetic data can offer real performance improvements' (*MIT News Office*, 3 November 2022) < https://news.mit.edu/2022/synthetic-data-ai-improvements-1103 > accessed 3 March 2023.

[145]For more details on the synthetic data generation process see, for instance, Erroll Wood et al (n 32) 3681-3682; Aryan Jadon and Shashank Kumar (n 30) 2.

[146]For more information see, for instance, Jiri Hradec et al (n 141).

[147]Tshilidzi Marwala et al (n 30) 6-7.

[148]Some IT experts propose many concrete technical improvements that could enhance the utilization of synthetic data in the realm of GenAI see, for instance, Aryan Jadon and Shashank Kumar, (n 30) 3; Sina Alemohammad et al (n 32) 11; Tshilidzi Marwala et al (n 30) 8-9.

[149]Existing models could involve e.g., statistical models or simulations and data analyst knowledge of the process can e.g., be based on textbook descriptions. See more discussions on different types of synthetic data generation, for instance, Khaled El Emam (n 30) 3-4, 54; James Jordon et al (n 135) 6.

since there is no direct reliance on real-world data.[150] However, if the aim is to develop highly versatile and functional GenAI models, the application of synthetic data with 'high utility' may be needed.[151] Synthetic data, which is created from real-world data, could demonstrate better quality since it preserves the essential characteristics (statistical properties) of actual data and could be validated on real content.[152] This implies that AI developers who use synthetic data to train their models would get similar outcomes from data analysis as if they performed mining on real-world data.[153] Some argue that even though this type of synthetic data is generated by training AI algorithms on actual data, it should not necessarily be considered 'bad' copying, especially regarding lawfully accessed works.[154] Indeed, the purpose of data synthesis is not to make exact (infringing) copies of protected materials that could substitute for the author's work but to 'read' existing materials in order to learn the underlying patterns, insights, structures, and relationships.[155] Copying that occurs in the process of 'reading' or analyzing protected data is an essential technical element in the proper functioning of modern technologies.[156] Therefore, if the process is carried out for non-expressive purposes, such as the creation of synthetic data samples which would merely 'imitate' real data, it should not raise copyright-related concerns and be prohibited by legal barriers.[157] However, if synthetic data used as input for training closely resembles real-world data protected by copyright, both the generation of the synthetic data and the training of

[150]Gareth Kristensen et al., 'Training AI models on Synthetic Data: No silver bullet for IP infringement risk in the context of training AI systems (Part 1 of 4)' (*Clearly IP and Technology Insights,* 12 December 2023) < https://www.clearyiptechinsights.com/2023/12/training-ai-models-on-synthetic-datanosilverbulletforipinfringement-risk-in-the-context-of-training-ai-systems-part-1-of-4/ > accessed 12 March 2024.

[151]Khaled El Emam (n 30) 4.

[152]Aryan Jadon and Shashank Kumar (n 30) 2-3; Khaled El Emam (n 30) 2; Mostly AI (n 34); Tshilidzi Marwala et al (n 30) 3-4; Joao Fonseca and Fernando Bacao (n 30) 1.

[153]Khaled El Emam (n 30) 2.

[154]In this context see, for instance, Giulio Coraggio, 'How synthetic data can address IP and privacy issues of artificial intelligence' (*GamingTechLaw*, 4 April 2023) < https://www.gamingtechlaw.com/2023/04/how-synthetic-data-can-address-ip-and-privacy-issues-of-artificial-intelligence/ > accessed 12 March 2024. AI developers, intending to generate synthetic training data from real-world data, need to obtain lawful access (through licensing, subscriptions or other lawful means) to protected data to ensure that rightsholders receive adequate compensation for the use of their works. For more on TDM, copyright and lawful access see, for instance, Maryna Manteghi (n 28) 451-453; Maryna Manteghi (n 28) 39.

[155]Peter Lee, 'Synthetic Data and the Future of AI' (2024) 110 *Cornell Law Review* (forthcoming) < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4722162 > accessed 12 March 2024, 15; Aryan Jadon and Shashank Kumar (n 30) 2. In this sense see: Sean Flynn et al (n 64) 4; Christophe Geiger et al (n 123) 817; Mira T. Sundara Rajan (n 122).

[156]Ibid.

[157]Peter Lee (n 155) 15; In this sense, see Matthew Sag, 'Fairness and Fair Use in Generative AI' (2024) 92 (5) *Fordham Law Review* 1887; See also famous US cases which hold that TDM constitute fair use: *Authors Guild v. HathiTrust,* 755 F.3d 87 (2d Cir. 2014) and *Authors Guild v. Google, Inc*., 804 F.3d 202 (2d Cir. 2015).

GenAI models on it could constitute copyright infringement.[158] Therefore, the risk of liability would depend on the quality of the AI-synthesised data, and, in other cases, on compliance with the requirements of Art. 4 of the CDSM Directive (e.g. 'lawful access' and the reservation of the rights).[159]

Further, AI actors could use AI-synthesised data to create more variations of synthetic data, which could be used to train an unlimited number of AI models, probably with no need to obtain additional consent from right-sholders.[160] Any subsequent generation of synthetic data would include fewer (or no) links or other attributions to the original data and accordingly fewer (or no) copyright concerns.[161] This would alleviate (to some extent) legal burdens enabling AI companies to access and use data more smoothly, efficiently and economically compared to generating training data from scratch.[162] This could help AI developers overcome limitations introduced by the CDSM Directive's TDM exceptions, in particular legal uncertainty associated with the reservation right of Art. 4. As it was stated above, with the lack of generally accepted protocols or standards, it is still unclear how AI developers intending to use materials to build their training datasets would 'discover and avoid' data that is subject to an 'opt-out' mechanism.[163] By using 'pure' synthetic training data (when the input data is also synthetic data), providers of GenAI systems could overcome the risk of training AI algorithms on data the use (mining) of which has been reserved by copyright holders. AI-synthesised data is not copyrighted,[164] therefore, the need to seek benefits from the 'commercial' TDM exception becomes irrelevant in this context.[165] Moreover, AI developers who build their training datasets on synthetic data and thus exercise full control over the creation of training data (e.g. generation techniques and methods) would probably find it easier to identify (disclose) data sources used for training under the AI Act's transparency obligation.[166] However, it is not clear how this scenario will work when it becomes necessary to evaluate the performance of AI models, trained on synthetic data,

[158]Peter Lee (n 155) 24; In this sense, see also GenAI pending cases (n 43).
[159]Peter Lee (n 155) 24.
[160]Katherine Lee et al (n 15) 51; In this sense, it is argued that accessing and using different variations of synthetic data for training would require less time, money and effort than obtaining 'lawful access' to real-world data every time when AI developers need to develop new GenAI models, see, for instance, Aryan Jadon and Shashank Kumar, (n 30) 3; James Jordon et al (n 135) 7; Adam Zewe (n 144).
[161]For technical aspects see Daniele Panfilo et al., 'A Deep Learning-Based Pipeline for the Generation of Synthetic Tabular Data' (2016) 11 IEEE < https://www.researchgate.net/publication/371828083_A_Deep_Learning-based_Pipeline_for_the_Generation_of_Synthetic_Tabular_Data > access
[162]Aryan Jadon and Shashank Kumar (n 30) 3; James Jordon et al (n 135) 7.
[163]Paul Keller (n 92); Nicola Lucchi (n 10) 15; Paul Keller and Zuzanna Warso (n 92); Péter Mezei (n 92) 3.
[164]Peter Lee (n 155) 46; Katherine Lee et al (n 15) 58.
[165]Gareth Kristensen et al., 'Training AI models on Synthetic Data: No silver bullet for IP infringement risk in the context of training AI systems (Part 3 of 4)' (*Clearly IP and Technology Insights,* 16 January 2024) < https://www.clearyiptechinsights.com/2024/01/training-ai-models-on-synthetic-datanosilverbulletforipinfringement-risk-in-the-context-of-training-ai-systems-part-3-of-4/ < accessed 12 March 2024.
[166]Ibid.

using real-world data. Without the availability of actual datasets, it would be impossible to assess whether the variations of synthetic training data include essential characteristics of the real data, which is important for producing outputs with a high level of utility and fidelity.[167] Furthermore, it is argued that the quality, reliability and diversity of GenAI models could degrade over generations if training datasets consist only of synthetic data with no fresh real-world data.[168] Some AI experts emphasise that the combination of synthetic data and actual data would be the best solution in this context, but this is not entirely settled in terms of copyright.[169] In this sense, there is still a need for further interdisciplinary research on synthetic data that would help IT specialists and legal experts to better understand this new paradigm in data analysis and, in particular, explore its capability to unlock the full potential of GenAI systems while preserving copyright holders' interests.

4. Concluding notes

Emerging GenAI models come with considerable benefits for users and pressing concerns for authors. Human creators could face a double threat to their interests as they have to compete with AI systems in the markets of creative works and also prevent the unfair use of their works to 'feed' AI algorithms during the training phase. The growing conflict of interest in the context of copyright and GenAI training could be addressed through two promising solutions: the CDSM Directive's 'commercial' TDM exceptions and synthetic data generation. Even though they could improve the current approach to data access to enable (contribute to) successful development and optimal utilisation of AI systems, some challenges persist in their functionality. Legal and technical 'errors and omissions' discussed in this article demonstrate that the remedies at hand may not completely resolve the conflict of interest, however, they could mitigate copyright-related concerns to some extent, if refined properly. To facilitate the fundamental rights to information and research in the context of AI development it is needed to ensure that copyright supports and enables TDM and does not create chilling effects through excessive restrictions and legal uncertainty. The broader 'commercial' TDM exception not including the reservations of rights could increase access to information, and facilitate research and

[167]Training on synthetic dataset and testing on the real dataset (TSTR) see James Jordon et al (n 135) 16; Yingzhou Lu et al (n 29) 13.

[168]Yuxi Ma et al., 'Brain in a Vat: On Missing Pieces Towards Artificial General Intelligence in Large Language Models' (2023) < https://arxiv.org/abs/2307.03762 > accessed 12 February 2024, 4; Thomas Göbel et al., ' Data for Digital Forensics: Why a Discussion on 'How Realistic is Synthetic Data' is Dispensable' (2023) 4(3) *Digit. Threat. Res. Pract.* 1, 14.

[169]Sina Alemohammad et al (n 22) 7; Thomas Göbel et al (n 168) 15. For more discussions on future research related to the development of synthetic data see, for instance, Aryan Jadon and Shashank Kumar (n 30) 3-4.

knowledge production that would enable the development and optimisation of game-changing ML applications in the EU. Further, more extensive research on synthetic data could unlock the full potential of this immature and underexploited approach to provide broader access to large, diverse and high-quality datasets with minimal effort and at reduced costs while complying with data creators' rights and interests. Against this background, active actions, in both legal and technical spheres, are needed to address the current copyright-related concerns in AI training that would be crucial for balancing powers in the digital environment and fostering the functionality of the EU AI sector.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Maryna Manteghi is a doctoral researcher in the Faculty of Law at the University of Turku, Finland.