Synthetic Data, Data Protection and Copyright in an Era of Generative AI

by **Kalpana Tyagi** *

Abstract: Data protection, privacy and copyright may be closely aligned, yet distinctly respond to the common element called data – that comprises personal as well as non-personal elements. Data can be of many different types, and when extracted from human-authored works, the expressive form of the work is subject to copyright protection. When personal data are included in a given dataset, it may trigger the application of the EU General Data Protection Regulation. Together, all the different sources form training data, which forms a key input for the training of generative AI models. These models have substantially devoured data to reach their current level of sophistication and capabilities. However, generative AI models are advancing at a rapid pace, such that they are no longer a mere consumer of data; they are also a key producer of new data – one that mimics the original data. This data is known as 'synthetic data'. Once the currently available models go a step further than their present level of development, follow-on synthetic data may look like independent works, with remote resemblance, if any, to the original data. While on the one hand, this may be a big promise to meet compliance with the 2016 EU General Data Protection Regulation, it heralds notable challenges for the current IPR (particularly copyright and database rights) framework and the accompanying balancing of authors' and users' rights. This interplay – considering its inter- and intra-disciplinary complexity – remains under-explored in the literature. This contribution, accordingly, explores the interaction between copyright (and other IPRs), database rights and data protection and privacy in the context of synthetic data and generative AI.

Keywords: Synthetic Data, Generative AI (GenAI), Internet of Things (IoT), Copyright, Database Rights, Personal Data, Data Protection, Privacy, Innovation, 2024 EU AI Act, Text and Data Mining, Robert Kneschke v. LAION, Charter of Fundamental Rights (CFR)

Recommended citation: Kalpana Tyagi, Synthetic Data, Data Protection and Copyright in an era of Generative AI , 16 (2025) JIPITEC 176 para 1.

A. Introduction

1 The complexity of the generative AI value chain means that a range of inputs are required to create a high quality general purpose AI model (GPAI), such as Chat GPT. These inputs, together referred to as the AI infrastructure layer, include the following four key elements – 'computing power', 'skilled workforce', large investments for research and development (R&D) and 'data'.[2] Among these elements, this research article concentrates on 'data', a key input that is used to train generative AI models. The OECD defines data as the 'physical representation of information in a manner suitable for communication, interpretation, or processing by human beings or by automatic means'.[3] Data used for training a GPAI is

Dr. Kalpana Tyagi. Maastricht University, The Netherlands. Email: k.tyagi@maastrichtuniversity.nl

* The work is author's research output. The author would like to extend her gratitude to Prof. Henning Gross-Ruse Khan, Prof. David Erdos and all the attendees for their inputs at the CIPIL Seminar at the Faculty of Law, University of Cambridge on 24th October 2024. The author would also like to extend her special thanks to Prof. Miquel Peguera Poch, the editorial team at JIPITEC and the anonymous peer reviewers at JIPITEC for their very insightful inputs. This work covers the legal and technical development in the fast-moving field of generative AI until 24 July 2025.

2 Autorité de la concurrence 'Intelligence artificielle générative: l'Autorité s'autosaisit pour avis et lance une consultation publique jusquáu vendredi 22 mars' (8 February 2024) <https://www.autoritedelaconcurrence.fr/fr/communiques-de-presse/intelligence-artificielle-generative-lautorite-sautosaisit-pour-avis-et-lance> accessed 27 July 2025.

3 OECD, 'Glossary of Statistical Terms' (2008) p. 119.

referred to as 'training data' in the 2024 EU AI Act. Article 3(29) of the Act defines this training data as 'data' that is used to train the 'AI system [by] fitting its learnable parameters'. Data can be proprietary or non-proprietary, and can include personal, factual, real-time flow of information, creative expression of works (subject to copyright protection), organised as a database (subject to copyright and database rights), technical information (such as in the case of patents) or a business secret (protected as a trade secret[4]). In addition, when in large quantities, data must also be organised to enable structured access to its contents. This structured organisation is particularly central to information systems and machine learning wherein terabytes of data are used to train the AI models. This 'data about data' is known as 'metadata'.[5] Structural, descriptive, administrative and markup languages are some of the common ways of organizing metadata, and depending on their type and structure, may facilitate various use cases.[6] Structural metadata helps establish the correlation between different databases and their contents. Descriptive metadata helps identify the source of information. Administrative metadata helps file management and identify various rightholders (such as authors of copyright-protected works). Markup languages facilitate easy navigation and interoperability. This metadata can also be IP-protected. In *Bart v. Anthropic*, for example, the US District Court for the Northern District of California opined that '[accessing] over seven million copies [of works … by downloading] a separate catalog of bibliographic metadata for each collection, with fields like title, author, and ISBN' was unauthorized use and constituted piracy of works.[7]

2 Thus, 'data', the oil that lubricates the digital economy, is not one homogenous element. Instead, it comprises many elements, and depending on the source, nature and form, may be subject to different disciplines of law. This diversification can also be classified on the basis of an access-driven framework, grounded in economic rationales, and a rights-driven framework, grounded in the safeguard for the protection of fundamental rights. In addition, it may also be important to clarify at the outset the difference between 'data' and 'information', which in legal literature has been distinguished on the basis of 'differentiation between the form [the 'digital form' that contains the information] and the meaning contained in that form [that is the information]'.[8] Information at a 'semantic' level may be conceived of as the 'meaning' or 'information *per se*', which is different from the syntactic level, which is the 'form' in which this information is 'expressed'.[9]

3 To regulate the digital economy, in 2020, the Commission proposed a 'European Data Strategy'.[10] To build a European single market for data, or Common European Data Spaces (CEDS), data is the key. The European Commission's vision on developing a data economy envisioned a range of measures to achieve this policy objective and facilitate access to data.[11] The EU Digital Markets Act (DMA) seeks to promote contestability and fairness in the digital markets mandates data portability (under Article 6(9)), and data access obligations (under Article 6(10) and 6(11)) on gatekeepers as regards core platform services.[12] In the Internet of Things (IoT), wherein smart devices must effortlessly communicate with one another for effective functioning, access to data from different service providers and device manufacturers is

x<https://www.oecd.org/en/publications/oecd-glossary-of-statistical-terms_9789264055087-en.html> accessed 27 July 2025.

4 Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, Article 2; Agreement on Trade-Related Aspects of Intellectual Property Rights, 15 April 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, The Legal Texts: The Results of the Uruguay Round of Multilateral Trade Negotiations 320 (1999), 1869 U.N.T.S. 299, 33 I.L.M. 1197 (1994), Article 39(1) & (2).

5 Jenn Riley, 'Understanding Metadata: What is metadata, and what is it for?' (2017) NISO p.5 <https://www.niso.org/publications/understanding-metadata-2017> accessed 27 July 2025.

6 I*bid.*, pp. 6-7.

7 Andrea Bartz, Charles Graeber and Kirk Wallace Johnson v. Anthropic PBC, No. C24-05417 WHA p. 3,5 *United States District Court Northern District of California* (23 June 2025) <https://storage.courtlistener.com/recap/gov.uscourts.cand.434709/gov.uscourts.cand.434709.231.0_2.pdf>

8 Václav Janeček, 'Ownership of personal data in the Internet of Things' (2018) *Computer Law and Security Review* 34(5) p. 1042 <https://doi.org/10.1016/j.clsr.2018.04.007> accessed 27 July 2025.

9 Herbert Zech 'Information as Property' (2015) *JIPITEC* 6(3) pp. 193-194 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2731076> accessed 27 July 2025.

10 Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European Strategy for Data (the European Data Strategy) COM/2020/66 final <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52020DC0066> accessed 27 July 2025.

11 *Ibid.*

12 John Burden, Maurice Chiodo, Henning Grosse Ruse-Khan, Lisa Markschies, Dennis Müller, Seán Ó hÉigeartaigh, Rupprecht Podszun and Herbert Zach, 'Legal Aspects of Access to Human-Generated Data and Other Essential Inputs for AI Training' (December 2024) *University of Cambridge Faculty of Law Research Paper No. 35/2024* p. 30 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5045155> accessed 27 July 2025.

essential to facilitate competition in the sector.[13] This access-driven framework, targeted at the 'Internet of Things', is triggered when data is generated by connected devices. The governing principle here is that users and other firms (particularly start-ups and small and medium enterprises) can have easier, real-time access to the IoT data, which is otherwise within the *de facto* control of the device manufacturers.[14] Together, the legislative measures flowing from the Commission's digital strategy, namely the Data Act, the Data Governance Act, the Open Data Directive and the Regulation on the Free Flow of Non-Personal Data (FFNDPR), collectively form the 'European data laws'.[15] These EU data laws, as well as the Payment Services Directive 2 (PSD2) for the payment sector, the DMA and the general EU competition law framework are driven by principles of contestability, access, competition, economic and market-based rationales.[16] Even in these market-driven data access frameworks, if personal data is included, additional conditions, such as 'valid lawful ground under the GDPR', must be met.[17]

4 Thus, the nature of the data - whether personal or non-personal, even though a somewhat fluid dividing line[18] – influences 'the rhetoric used and the priorities set' and impacts the 'extent of data access'.[19] Even though a simple binary distinction of data as personal/ non-personal may soon become superficial, it is nonetheless a good starting point, as it helps modulate data, and as discussed below, the technological innovation called synthetic data is to be understood within a fundamental rights framework.[20] Data, when derived from works in an expressive form, may be subject to copyright protection.[21] When a given dataset comprises personal elements, it may trigger the application of the 2016 EU General Data Protection Regulation (GDPR), that acts as a safeguard to protect the fundamental right to data protection and privacy of the data subject. Article 8 of the Charter of Fundamental Rights (CFR) and Article 16 of the Treaty on the Functioning of the European Union (TFEU) offer a constitutional safeguard to personal data as a basic fundamental right. While privacy is a long

13 European Commission, 'Final report – Sector inquiry into consumer Internet of Things' (Brussels, 20.01.2022) *SWD* (2022) 10 final <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022SC0010> accessed 27 July 2025.

14 Wolfgang Kerber, 'Governance of IoT Data: Why the EU Data Act Will not Fulfill Its Objectives' (2023) *GRUR International* 72(2) p. 120, 124 <https://doi.org/10.1093/grurint/ikac107> accessed 27 July 2025; Oscar Borgogno and Gieuseppe Colangelo, 'Shaping interoperability for the IoT: the case for ecosystem-tailored standardisation' (March 2024) *European Journal of Risk Regulation* 15(1) p. 150 <https://doi.org/10.1017/err.2023.8> accessed 27 July 2025.

15 Thomas Margoni, Charlottee Ducuing and Luca Shirru, 'Data Property, Data Governance and Common European Data Spaces' (2023) *Computerrecht: Tijdschrift voor Informatica, Telecommunicatie en Recht* p.2 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4428364> accessed 27 July 2025, Thoms Streinz 'The Evolution of European Data Law' in Paul Craig and Gráinne de Búrca (eds) *The Evolution of EU Law* (Oxford University Press, 3rd edn 2021) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3762971> accessed 27 July 2025.

16 Peter Georg Picht, 'Caught in the Acts: Framing Mandatory Data Access Transactions under the Data Act, further EU Digital Regulation Acts, and Competition' (March 2023) *JECLAP* 14(2) <https://doi.org/10.1093/jeclap/lpac059> accessed 27 July 2025; Wolfgang Kerber, 'Data Act and Competition: An Ambivalent Relationship' (2023) *Concurrences* 1/2023 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4342488> accessed 27 July 2025; Inge Graef (2021) 'Why End-User Consent Cannot Keep Markets Contestable: A suggestion for strengthening the limits on personal data combination in the proposed Digital Markets Act' (2 September 2021) *Verfassungsblog* <https://verfassungsblog.de/power-dsa-dma-08/> accessed 27 July 2025; John Burden, Maurice Chiodo, Henning Grosse Ruse-Khan, Lisa Markschies, Dennis Müller, Seán Ó hÉigeartaigh, Rupprecht Podszun and Herbert Zach, 'Legal Aspects of Access to Human-Generated Data and Other Essential Inputs for AI Training' (December 2024) *University of Cambridge Faculty of Law Research Paper No. 35/2024* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5045155> accessed 27 July 2025.

17 Thomas Tombal and Inge Graef 'The Regulation of Access to Personal and Non-Personal Data in the EU: From Bits and Pieces to a System?' in Van der Sloot & van Schendel (eds): *The Boundaries of Data* (Amsterdam University Press, 2024) p. 196.

18 Josef Drexl (2019) 'Data Access and Control in the Era of Connected Devices' p. 124 <https://www.beuc.eu/sites/default/files/publications/beuc-x-2018-121_data_access_and_control_in_the_area_of_connected_devices.pdf> accessed 27 July 2025.

19 Tombal and Graef (2024), *supra* note 17, pp.195-196.

20 Bárbara da Rosa Lazarotto and Gianclaudio Malgieri, 'The Data Act: a (slippery) third way beyond personal/non-personal data dualism?' (4 May 2023) *European Law Blog* <https://www.europeanlawblog.eu/pub/the-data-act-a-slippery-third-way-beyond-personal-non-personal-data-dualism/release/1> accessed 27 July 2025; Ana Beduschi 'Synthetic data protection: Towards a paradigm change in data regulation?' (2024) *Big Data & Society* p. 3 <https://journals.sagepub.com/doi/10.1177/20539517241231277> accessed 27 July 2025.

21 Cf Recital 9, Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (2019 CDSM). The recital states that text and data mining can also be carried out in relation to mere facts that are not protected by copyright, and in such instances no authorization is required under copyright law.

established 'venerable right', with firm foundations in national constitutions and international treaties, data protection has been identified as a more '"third generation fundamental right" or innovation for traditional human rights, one that is now included in the EU Charter of fundamental rights'.[22] Likewise, the need to remunerate the human author and balancing the authors' rights and users' rights is becoming increasingly central to the discussion on copyright and related rights in the digital economy. This has drawn the attention of the policy makers and courts alike, and has given way to a 'new doctrinal stream called "digital constitutionalism" [or constitutionalisation of intellectual property rights]'.[23] A fundamental rights-driven rhetoric is thereby more central to the EU GDPR and copyright and related rights framework.

5 Data generated from this original data is called 'synthetic data'. Synthetic data is artificially generated data that can be generated using techniques such as statistical sampling or more advanced AI learning techniques.[24] The '[synthetic] data generation revolution' is anticipated to significantly influence the 'current balance between utility and competing considerations' as over 60% of training data may be synthetically generated and it carries the 'potential to do to data what synthetic threads did to cotton'.[25] GenAI models are experiencing rapid technological advances and are a key generator of synthetic data. Once the currently available technology goes a step further than its current level of development, synthetic data generated from these GenAI models may not even be closely reminiscent of the original training data. While, on the one hand, this may be a big promise to comply with the GDPR as it may safeguard the identity of the data subject, the rise of synthetic (big) data presents notable challenges for the current IPR framework (particularly copyright), as well as for the balancing of authors' and users' rights – the two key legal frameworks central to this contribution.

6 Different possibilities emerge with the rise of synthetic data. Will synthetically generated data replace all the original human-generated data? Or will original human-generated and machine-generated synthetic data co-exist? These potential future scenarios will also impact (and be impacted by) copyright and data protection laws. They will also influence different fundamental rights, such as the right to property, the right to freedom of expression, the right to data protection and the right to privacy. The interplay thus, involves many composite elements, each of which must be decoded to solve the innovation complex. To follow this discourse, this contribution follows an inter- and intra-disciplinary research methodology and is organised as follows. Section 2, with inputs from the technical literature, offers a working definition of synthetic data alongside a non-technical insight in the technical aspects of synthetic data generation. Considering the innovation potential of synthetic data, section 3 assesses the interplay between synthetic data and IP (especially copyright and database rights). Section 4 develops the discourse from the lens of personal data, the EU General Data Protection Regulation (GDPR) and evaluates its pedigree in the fundamental rights, constitutional and doctrinal legal framework. Section 5 concludes with policy recommendations and directions for further research. It also returns to the central hypothesis of this paper, that is whether synthetic data will emerge as a complete substitute or whether will it remain a partial substitute, in other words, a complement to the human generated data, and whether distinct legal, technical and normative trade-offs may limit substitution of human generated data by synthetically generated data.

22 David Erdos, 'Comparing Consitutional Privacy and Data Protection Rights within the EU'(2021) *University of Cambridge Faculty of Law Research Paper No. 21/2021* pp. 1-2 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3843653> accessed 27 July 2025; *See also* Orla Lynskey, 'Deconstructing Data Protection: The "Added-Value" of a Right to Data Protection in the EU Legal Order' (July 2014) *ICLQ* 63 (3) <https://doi.org/10.1017/S0020589314000244> accessed 27 July 2025.

23 Elena Izyumenko and Christophe Geiger 'Intellectual Property and Human Rights in the Jurisprudence of the CJEU and the ECtHR – An Introduction' (*pre-print,* 2025) in Elena Izyumenko and Christophe Geiger (eds.), *Human Rights and Intellectual Property before the European Courts: A Case Commentary on the Court of Justice of the European Union and the European Court of Human Rights* (Edward Elgar Publishing Forthcoming) p. 1 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5283506> accessed 27 July 2025; Edoardo Celeste, 'Digital constitutionalism: A new systematic theorisation' (2019) *International Review of Law, Computers & Technology* 33(1): British and Irish Law Education and Technology (BILETA Special Edition) p. 88 <https://www.tandfonline.com/doi/full/10.1080/13600869.2019.1562604> accessed 27 July 2025.

24 J Hradec, M Craglia, M Di Leo, S De Nigris, N Ostlaender and N Nicholson 'Multipurpose synthetic population for policy applications' (2022) *Joint Research Center, Digital Economy Unit: Technical Report: Publications Office of the European Union* p.12 <https://publications.jrc.ec.europa.eu/repository/handle/JRC128595> accessed 27 July 2025.

25 Michal S Gal and Orla Lynskey, 'Synthetic Data: Legal Implications of the Data-Generation Revolution' (2024) *Iowa Law Review* 109 p. 1091 <https://ilr.law.uiowa.edu/volume-109-issue-3/2024/03/synthetic-data-legal-implications-data-generation-revolution> accessed 27 July 2025. See also the references therein.

B. Synthetic Data

7 In traditional programming, best suited for constrained and structured environments, pre-defined rules called algorithms instruct machines to perform certain tasks.[26] These are simple decision trees, structured into a pre-defined 'IF-THEN-ELSE' format. Distinct from these are AI systems, that instead of following a structured pre-defined path, 'learn how to solve a problem by examining [the] training data'.[27] There are two key methods for training an AI system, namely, 'machine learning' (ML) and 'deep learning' (DL). In ML, systems are trained on large amount of data, and the quantity and the quality of the training data determines the quality of the AI system. ML is a good technique to train AI systems for weather forecasting and image and speech recognition.[28] DL, a subset of ML, mimics the 'complex processes [known as Artificial Neural Networks] inspired by the human brain' and is deployed in complex, creative and research and development-driven tasks, such as for creating 'new works of art and [for] medical drug discovery'.[29]

8 While AI has been around for a long time, it is the disruption by GenAI applications, such as Open AI's ChatGPT, Google's Bard (since Gemini) and Microsoft's Copilot, that has since 2022, gathered the attention of businesses and policy makers alike. This disruptive rise of GenAI was facilitated by a key innovation from Google's team that introduced 'transformers', a novel form of AI architecture, that relied 'entirely on self-attention to compute representation of its input and output without using sequence aligned RNNs [recurrent neural networks] or convolution'.[30] This technical innovation was disruptive at the time, as it was the first time that an 'encoder-decoder architecture with multi-headed self-attention' was used in place of the traditional recurrent layer architecture.[31] This process was significantly faster, more accurate and more data efficient than the, at the time popular, recurrent and convolutional frameworks used for language translation.[32] It also contributed to substantial improvements in the Google translate feature. The framework offered in the said paper, however, was limited to text-based inputs. Following the introduction of transformers, rapid developments took place in the field of deep learning. Follow-on works led to newer innovations and efficiency in image, audio and video generation using deep learning techniques.

9 Common to all these models is the need for the input 'data'. GenAI and large language models (LLMs) follow the 'neural scaling laws', wherein data is a key input. Neural scaling means that the efficient performance of the model enjoys a positive correlation with the size of the training datasets.[33] The larger the number of datasets used to train a GenAI model, the higher the quality and throughput of the model. Quantity alone is insufficient to ensure a high quality GenAI model; the quality of the datasets matters as well. Data is qualified by its five characteristics – volume, velocity, variety, veracity and value, also known as the 'Vs of Big Data'.[34] Higher quality datasets contribute to robust and well-functioning models. This hunger for data is endemic to large-scale deep learning models, and not to the traditional machine learning methods, such as decision trees, SVM, KNN models and discriminant analysis.[35] Large scale deep learning models that are trained on vast amounts of datasets, are referred to as foundation models (FM), and offer potentially diverse capabilities across a range of applications such as text generation, creative applications and even coding.[36]

26 Microsoft, 'Introduction to Artificial Intelligence (AI) Technology' (2024) pp. 8-9 <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/2024-wttc-introduction-to-ai.pdf> accessed 27 July 2025.

27 *Ibid.*, p. 9.

28 *bid.*, p. 10.

29 *Ibid.*, p. 10.

30 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkorelt, Llion Jonesm, Aidan N. Gomez, Łukasz Kalser, Illia Polosukhin, 'Attention Is All You Need'(2017) p. 2 <https://arxiv.org/abs/1706.03762> accessed 27 July 2025.

31 *bid.*, pp. 9-10.

32 *Ibid.*, p. 10.

33 *See* reference to J Kaplan, S McCandlish, T. Henighan, TB Brown, B Chess, R Child, S Gray, A Radford, J Wu and D Amodei (2020) 'Scaling laws for neural language models' and to J Hoffmann, S Borgeud, A Mensch, E Buchatskaya. T Cai, E Rutherford, D. d.L. Casas, LA Hendricks, J Welbl, A Clark, T Henngan, E Noland, K Millican, G. v.d. Drissche, B Damic, A Guy, S Osindero, K Simonyan, E Elsen, JW Rae, O Vinyals and L Sifre 'Training compute-optimal large language models' (2022), in Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim and Marius Hobbhahn, 'Will we run out of data? Limits of LLM scaling based on human-generated data' (4 June 2024) p. 1 <https://arxiv.org/abs/2211.04325> accessed 27 July 2025.

34 Annie Badman and Matthew Kosinski, 'What is big data?' (18 November 2024) *IBM* https://www.ibm.com/think/topics/big-data#:~:text=Subscribe%20today-,The%20V's%20of%20big%20data,needed%20to%20manage%20it%20effectively.

35 Zehui Zhao, Laith Alzubaidi, Jinglan Zhang, Ye Duan and Yuantong Gu, 'A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations' (15 May 2024) *Expert Systems with Applications* 242 p. 23 <https://www.sciencedirect.com/science/article/pii/S0957417423033092> accessed 27 July 2025.

36 Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Syndeny von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, and Emma Brunskill and other 'On the Opportunities and Risks of

10 Thus, a digital firm that aspires to develop a competitive large-deep scale learning model needs data, which is a key input required to train the GenAI model. Between digital incumbents and start-up firms, the latter more frequently confront barriers to accessing quality datasets.[37] The real-world data must also be cleaned and labeled before it can be used for training purposes.[38] Digital gatekeepers, such as Google, Apple, Meta, Microsoft and Amazon (GAMMA) control large volumes of data that serve as training inputs to the LLMs. To add to the complexity, human-generated data has limited availability. Can it be that one may soon confront a paucity of quality data available to train these models? In other words, in addition to the barriers to accessing currently available human-generated data (both personal as well as non-personal), which by default is under the *de facto* control of the digital gatekeepers, will one soon confront another additional challenge – for instance, that this human-generated data becomes scarce and is eventually exhausted? Considering the current rate of data consumption used for training the models, this seems like a reasonable possibility. Villalobos *et al.* predict that if the pace of LLM training continues at the current rate, we may run out of 'public human text data between 2026 and 2032'.[39] Villalobos *et al.* make this observation in the context of LLMs, even though they also estimate the available text and non-text data in their empirical analysis. In addition, it also seems that the terms GenAI, LLMs, and Foundation Models (FMs), even though distinct, are sometimes used interchangeably.[40] Thus, before going further, it may be useful to offer a working definition of GenAI models, LLMs and Foundation Models (FMs), and organize their classification in the AI landscape, including the EU AI Act 2024/1689 (2024 EU AIA).

11 GenAI are AI models trained on large datasets, to generate new content – such as audiovisual, text, code, music or any other content that can be perceived by the senses – upon a mere prompt.[41] LLMs are a sub-category of GenAI as they are used to generate text-based data.[42] Thus, GenAI is a broader term that also covers the LLMs. The general purpose GenAI models, such as Bidirectional Encoder Representations from Transformers (BERT), the first-ever GenAI model, the Generative Pre-trained Transformer (GPT) by OpenAI, Stable Diffusion by Stability AI and Titan FM by Amazon are all widely-trained foundation models (FMs).[43] They are, in the language of the EU AI Act, known as the GPAI models, and require large amounts of training data. These models can then be fine-tuned to perform certain niche and personalized tasks.[44] Broadly speaking, AI models may be trained using one of the following three techniques: supervised learning (wherein models are trained on correctly labeled data), unsupervised learning (wherein data has not been labelled), and reinforcement learning (which involves learning by doing or the trial and error method).[45]

12 The 2024 EU AIA is a lengthy product-safety regulation comprising 108 recitals, 113 articles, and 13 annexes, and is divided into 13 chapters.[46] Even though principally a product-safety regulation, the Act has notable implications for copyright and data protection laws as well. The AI Act, in recitals 104-

Foundation Models' (16 August 2021) < https://arxiv.org/abs/2108.07258> accessed 27 July 2025.

37 Competition and Markets Authority (18 September 2023) 'AI Foundation Models: Initial Report' pp. 28-32

38 Peter Lee, 'Synthetic Data and the Future of AI' (*Cornell Law Review 110 forthcoming, pre-print 2024*) pp. 9-10 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4722162> accessed 27 July 2025.

39 Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim and Marius Hobbhahn, 'Will we run out of data? Limits of LLM scaling based on human-generated data' (4 June 2024) p. 6 <https:// arxiv.org/abs/2211.04325> accessed 27 July 2025.

40 This article will use the term GenAI to offer consistency to the discussion. As synthetic data remains central to the discussion, use of the term GenAI is also more representative of the technical field.

41 TechMobius, 'Generative AI vs. LLM, What is the difference?' <https://www.techmobius.com/blogs/generative-ai-vs-llm-what-is-the-big-difference/> accessed 27 July 2025.

42 *Ibid.*

43 Amazon Web Services, 'What are foundational models?' <https://aws.amazon.com/what-is/foundation-models/> accessed 27 July 2025.

44 Kalpana Tyagi, 'Mapping competition concerns along the generative AI value chain' (Forthcoming 2025) *SSRN* p. 18 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5282596> accessed 27 July 2025. See also the references, and various examples of targeted, niche FM models therein. Astro LLaMa for example is a niche, vertical FM model trained on limited data vis-à-vis, its parent FM model Llama by Meta.

45 Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever, 'Improving language understanding by generative pre-training' (2018) <https://www.mikecaptain.com/resources/pdf/GPT-1.pdf> accessed 27 July 2025; Microsoft, 'Introduction to Artificial Intelligence (AI) Technology' (2024) pp. 11-12 <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/2024-wttc-introduction-to-ai.pdf> accessed 27 July 2025; DeepSeek-AI 'DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning'(2025) <https://arxiv.org/pdf/2501.12948> accessed 27 July 2025.

46 João Pedro Quintais 'Generative AI, Copyright and the AI Act' (April 2025) *Computer Law & Security Review* 56 p. 6 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4912701> accessed 27 July 2025.

109 and Chapter V, dealing with 'General-purpose AI Models', provide copyright-related obligations for GPAI model providers. Notably, Article 53(1)(d) requires the GPAI model providers to offer a 'publicly available' and 'sufficiently detailed summary' of the datasets used for training the model. This summary must be provided in accordance with the explanatory notice and annex template contained in the 'Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models'.[47]

13 The GPAI models are a part of the larger subset comprising GP(AI) systems.[48] As an example, whereas GPT (Generative Pre-trained Transformer) is a model, ChatGPT and Midjourney are systems.[49] The copyright-related (and the data protection-related) obligations under the AI Act principally concern GPAI model providers with regard to the GPAI models, and not GPAI system providers with regard to the AI system.[50] This distinction becomes relevant to determine the application of the AI Act, as also noted by the Hamburg Regional Court in its recent *Kneschke v. LAION* decision, the EU's first GenAI and text and data mining (TDM) decision discussed in section 3, *infra*.

14 As available and clean human-generated data is limited in quantity, how do digital firms resolve the challenge of limited access to superior quality datasets? There are three inter-related technical possibilities that may help overcome this data bottleneck: enhancing the efficiency of data consumption by the GenAI models (1), developing new techniques such as transfer learning (2), and synthetic data generation (3).[51] The current training of GenAI, including BERT, Stable Diffusion and ChatGPT, is inefficient and resource-intensive, and optimizing the efficiency of the training process is a key research agenda in the field.[52] Transfer learning (TL) and self-supervised learning (SSL) can help overcome the data bottleneck associated with the incumbent inefficient learning methods.[53]

15 Transfer learning (TL) is a popular learning method in computer vision and natural language processing (NLP) tasks such as sentiment analysis.[54] TL or knowledge transfer uses pre-trained parameters from earlier trained models to develop a robust foundation model.[55] For example, if a model has been trained to identify deep-faked political news, this pre-trained model can be used to train a new model that can identify political satire. In TL, the relevant parameters from the earlier trained models are pre-selected to further fine-tune and adapt to develop a new FM that is suitable for the task under consideration. Simply put, TL may be suitable when the developer uses 'pre-trained parameters from earlier trained models' to develop vertically-specialised applications in certain domains.

16 Self-supervised learning (SSL), like transfer learning, is another commonly used learning approach to overcome limited data availability. SSL, an unsupervised learning technique, extracts 'reusable features from source data' and recycles them to make new models.[56] In other words, this recycling of the data helps develop personalized models that can leverage the capabilities of earlier trained models.

17 Interestingly, these different approaches are also interconnected. To train the GenAI models, data is an input as well as an output. Though not always, human-generated data may serve as a good and robust input to generate synthetic data. Synthetic data, especially when real datasets are unavailable, may serve as a good input for TL and SSL models and thereby, help optimize the overall learning process. Interestingly, synthetic data is also the output from the GenAI models, which in turn also serves as an input for further training and fine-tuning these models. There is thus a circular element in the GenAI value chain – the input generates an output which is further recycled to generate more output. These efficiency-enhancing steps also save the expert resources required for the creation of large, labeled datasets and considerable time that is otherwise required to train deep neural networks for complex tasks.[57] Interlinkages between different steps – such as how synthetic data is the output and may also

47 European Commission, 'Annex to the Communication to the Commission – Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models required by Article 53(1)(d) of Regulation (EU) 2024/1689 (AI Act)' (24 July 2025) <https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models> accessed 27 July 2025.

48 Quintais (2025), *supra* note 46, p. 6.

49 *Ibid.*

50 *Ibid.*, p. 7.

51 Villalobos et al (2024), *supra* note 39, p. 9.

52 Sashank J. Reddi, Sobhan Miryoosefi, Stefani Karp, Shankar Krishnan, Satyen Kale, Seungyeon Kim and Sanjiv Kumar, 'Efficient Training of Language Models using Few-Shot Learning' (2023) *Proceedings of the 40th International Conference on Machine Learning* p. 1, 7 <https://proceedings.mlr.press/v202/j-reddi23a/j-reddi23a.pdf> accessed 27 July 2025.

53 Zhao et al (15 May 2024), *supra* note 35, pp. 2,22.

54 Niklas Donges, Matthew Urwin and Parul Pandey, 'What Is transfer Learning? Explore the Popular Deep Learning Approach' *Builtin* (15 August 2024) <https://builtin.com/data-science/transfer-learning#:~:text=Transfer%20learning%20is%20a%20machine,model%20despite%20having%20limited%20data> accessed 27 July 2025.

55 *Supra* note 53, p. 2.

56 *Ibid.*

57 *Supra* note 54.

serve as an input in the training process – create economies of scale and scope across the GenAI value creation process.

I. Synthetic Data: Use Cases and Methods of Generation

18 Synthetic data is artificially generated data. Synthetically generated data can be visual, written, tabular, audiovisual, graphic or, as technological advances may permit, data that can be perceived by the senses. An important feature of synthetic data is that it shares the same statistical properties as the original data.[58] Synthetic data are superior to traditional anonymization techniques. Traditional anonymization techniques cover only certain aspects of the data, and are therefore unsuitable in the case of big data, in which there are 'no non-sensitive attributes'.[59] In other words, big data and advanced algorithms makes it possible to deanonymize datasets that may also include non-sensitive attributes. Synthetically generated data that can be reverse-engineered to reconstruct the original dataset cannot qualify as synthetic data.[60] Synthetically generated data, thus, retain the statistical properties of the original dataset without revealing any personal attributes of the original dataset, and in this respect, are a viable option to facilitate compliance with data protection laws, as compared to traditional anonymization techniques. There are, however, certain subtle legal aspects that qualify the conditions under which synthetic datasets may indeed be exempt from the scope of data protection laws – these different use cases are further discussed in section 4 *infra*, which deals with the interface between synthetic data and data protection laws.

19 There are various methods of synthetic data generation. Data synthesis, using statistical techniques such as 'Synthetic Reconstruction' (SR) and 'Combinatorial Optimization' (CO), dates back to the 1980s, and has been regularly used by statisticians to fill-in the missing data and construct artificial populations.[61]. The rise of big data and computing power, along with the use of probabilistic models such as Deep Generative Models (DGM), created 'a new generation of models that exploit deep learning for creating synthetic data'.[62] Herein, Ian Goodfellow's work on deep learning and Generative Adversarial Networks (GANs), contributed significantly to the uptake of synthetic data.[63] Synthetic data can be generated in a fully autonomous environment, such as through a large language model, whereby output from the model is re-fed and used as input to train the model. Alternatively, the training cycle can be mixed with initial training by human-generated data, followed by the next round of training with a mix of synthetic and human-generated data.

20 Synthetic data have a range of applications. They can be used to enhance privacy (1); to de-bias datasets to homogeneously represent under-represented populations in datasets (2); to test products, such as IoT-enabled products for safety and accuracy prior to a formal product launch (3); to train GenAI models (4); and to create safe data spaces for data-driven innovation in digital markets (5). The following paragraphs illustrate some practical use cases of synthetic data to establish its value and policy significance.

1. De-Biasing Datasets for Homogeneous Population Representation

21 A key limitation of current GenAI models is that they tend to exhibit bias.[64] This bias is attributed to the limitations of input data, which itself may exhibit a range of biases, such as bias against certain gender or socio-economic backgrounds. Synthetic data generation techniques are a sustainable approach to de-biasing the datasets. One such approach, Synthetic Minority Over-sampling Technique (SMOTE) involves deliberately adding more artificially generated data about under-represented populations.[65] In large data sets, the majority or over-represented population groups are represented as the normal class, whereas the interesting or under-represented examples are the abnormal class.[66] SMOTE involves over-sampling the minority and under-sampling the majority class for a more uniform and egalitarian representation of the population.[67] This technique, dating back to decades-old statistical sampling techniques, has a wide range of applications to de-bias datasets. Consider, for

58 *Supra* note 24, pp. 14-15.
59 *Ibid.,* p. 44.
60 *Ibid.*
61 *Ibid.,* p.12.
62 *Ibid.,* p.14.
63 Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep learning* (2016) The MIT Press <http://alvarestech.com/temp/deep/Deep%20Learning%20by%20Ian%20Goodfellow,%20Yoshua%20Bengio,%20Aaron%20Courville%20(z-lib.org).pdf>
64 *Supra* note 24, p. 18, 45
65 N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer 'SMOTE: Synthetic Minority Over-sampling Technique' (2002) *Journal of Artificial Intelligence Research* 16 pp. 321-357 <https://www.jair.org/index.php/jair/article/view/10302> accessed 27 July 2025.
66 *Ibid.*, pp. 329-330.
67 *Ibid.*, p. 331, 352.

example, the datasets of scientists working in the STEM field. An original dataset may over-represent those with a certain gender and socio-cultural and ethnic background. With SMOTE, one can de-bias such a dataset and create a more homogenous output that can then be used for better policy decisions. However, this oversampling method needs to be fine-tuned for deep learning architecture, wherein traditional SMOTE techniques may have limited effectiveness. Herein, computer scientists have proposed over-sampling approaches, such as 'Deep SMOTE', that deploys a modified SMOTE architecture in a deep neural network and 'Deep Adversarial SMOTE', that is a further fine-tuned Deep SMOTE model for unsupervised networks.[68]

2. Testing IoT-Enabled Products for Safety prior to Formal Product Launch

22 Internet of Things (IoT) is a world of inter-connected devices, whereby there is smooth machine-to-machine, and human-to-machine communication. As different devices talk to each other to offer novel goods and services, quality data may oftentimes be unavailable to train these IoT models prior to a product launch. Amazon, the world's largest online retailer, with rich data on the products and services sold on its e-commerce platform, too suffered from this data limitation. Even though Amazon has a rich database about user behaviour on its platform, it had limited insight about how users might respond to its yet-to-be-launched voice assistant, Alexa, at the time. To address the challenge of limited access to quality data to 'bootstrap the machine learning models that interpret customer requests', Amazon used large volumes of synthetic data to train Alexa.[69] Amazon required training data to anticipate what its potential customers might want Alexa to do while using it. This challenge is not unique to Amazon alone. Other large digital firms, such as Microsoft, Apple, and Google, confronted similar issues in the early stage of developing their respective virtual personal assistants, that could interact with human users 'via spoken interactions' and 'gesture recognition'.[70] To train Alexa's *Natural Language Understanding* (NLU) systems, the Alexa AI team used the available customer data as a basic template and used it to identify general syntactic and semantic patterns. These patterns were then used as a basis to construct a large number of 'new, similar sentences'.[71] NLU models trained on such data sets could more easily identify complex patterns and deliver higher performance.[72] Following Alexa's success, Amazon has continued to integrate synthetic data and GenAI techniques to make its product lines more robust. Amazon, as a leading IoT player, has detailed user profiles, including personal data, about user interaction with Alexa, an aspect that is subject to data protection laws.[73] This factual information, including elements of personal data, can be simulated and mixed with synthetically-generated data to train the models.[74] Amazon also used this synthetic data generation technique to develop voice recognition systems across different languages, wherein 'it faced a shortage of collected data'.[75] Once developed and successfully launched, these GenAI-enabled, IoT products further leverage on a mix of human generated data to continue optimising their performance. GenAI-enabled IoT, such as Amazon's DialFRED, facilitate 'user-device interaction and foster problem-solving capabilities'.[76] Successive iterations of synthetic and human-generated data, help voice, image and other non-text-based

68 Hadi Mansourifar and Weidong Shi, 'Deep Synthetic Minority Over-Sampling Technique' (2020) <https://arxiv.org/pdf/2003.09788> accessed 27 July 2025.

69 Janet Slifka, 'Tools for generating synthetic data helped bootstrap Alexa's new-language releases' (11 October 2019) *Amazon Science: Conversational AI* <https://www.amazon.science/blog/tools-for-generating-synthetic-data-helped-bootstrap-alexas-new-language-releases> accessed 27 July 2025.

70 Veton Z Këpuska and Gamal Bohouta, 'Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home' (University of Nevada, Las Vegas, 2018) *IEEE 8th Annual Computing and Communication Workshop and Conference* <https://ieeexplore.ieee.org/abstract/document/8301638, open access version available here: https://www.researchgate.net/publication/322418348_Next-Generation_of_Virtual_Personal_Assistants_Microsoft_Cortana_Apple_Siri_Amazon_Alexa_and_Google_Home> accessed 27 July 2025.

71 Slifka (11 October 2019), *supra* note 69.

72 *Ibid.*

73 Guido Noto La Diega and Christiana Sappa, 'The Internet of Things at the intersection of data protection and trade secrets. Non-conventional paths to counter data appropriation and empower consumers' (2020) *Reveue européene de droit de la consummation/ European Journal of Consumer Law*, p. 4,5,11 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3772700> accessed 27 July 2025. See also the references therein.

74 Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, Jasha Droppo, 'SynthASR: Unlocking Synthetic Data for Speech Recognition' (2021) <https://arxiv.org/pdf/2106.07803> accessed 27 July 2025.

75 Gal and Lynskey (2024), *supra* note 25, p. 15. See also the references therein.

76 Mazlan Abbas, 'Generative AI Applications for IoT: Exploring the Future of Smart Devices' (3 April 2023) *IoT World* p. 3 <https://iotworld.co/2023/04/3-generative-ai-applications-for-iot-exploring-the-future-of-smart-devices/> accessed 27 July 2025; Maria Kanwal, '3 Generative AI Applications for IoT you must know about' (12 December 2023) *SSRN* p. 5 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4667577> accessed 27 July 2025.

models continuously strengthen their capabilities, and offer better user experience.[77] Simply put, a mix of human and synthetically generated data is continuously fed into the IoT system to enhance and optimise their performance and capabilities. Co-existence of human and synthetically generated data boosts model performance, and is a particularly appealing approach to training 'end-to-end (E2E) *Automatic Speech Recognition* (ASR) models' for new applications, whereby human-generated data may be sparsely available.[78]

3. Training GenAI Models

23 Synthetic data is a commonly deployed tool for the 'development, test[-ing] and validation' of machine learning systems, where data may either be unavailable or inaccessible.[79] ChatGPT was trained using a large corpus of data – both original human-generated and synthetic data. It was trained using 'unsupervised, Reinforcement Learning coupled with Human Feedback (RLHF) and semi-supervised' learning techniques.[80] The original human-generated data may have elements of personal data, as data can be classified as personal once there is a possibility of identification.[81]

24 Synthetic data can also be generated with sequential modelling, simulated data and decision trees.[82] Another important technique for synthetic data generation is *Fully Visible Belief Networks* (FVBNs) that follow a probability-driven approach to generating synthetic data.[83] FVBN is the basic model that has been further fine-tuned to create and train advanced models such as WaveNet by DeepMind.[84] As FVBN was relatively slow in generating output, it limited the successful commercial application of the technology. DeepMind fine-tuned this technology, and used a neural network driven-approach to accelerate the pace of output generation.[85] This transition to neural network-driven learning and use of 'transformers' by Google, as discussed above, and the addition of 'bidirectionality' to the learning process was disprutive.[86] Bidirectionality meant that GenAI tools, starting with BERT, could read the text in both directions – from the left to right, as well as the right to left. Pre-BERT, GenAI models could read only in one direction, either left to right or right to left. The ability to read bidirectionally offered GenAI to *Contextualize, Iterate and Improvise* (CII) with every iteration.[87] This enabled the GenAI tools to offer meaningful and contextual outputs, such as synthetic data, at a faster pace, and in larger quantities that could be further used as input to train and fine-tune these AI models.

25 General Adversarial Networks (GANs) and Variational Auto Encoders (VAE) are two current and state-of-the-art, commonly deployed deep learning algorithms, that are also the more frequently used models to generate synthetic data.[88]

77 Nishant Prateek, Mateusz Łajszczak, Roberto Barra-Chicote, Thomas Drugman, Jaime Lorenzo-Trueba, Thomas Merritt, Srikanth Ronanki, and Trevor Wood 'In other news: a bi-style text-to-speech model for synthesizing newscaster voice with limited data' in A. Loukina, M. Morales and R. Kumar (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2019) pp. 205-213 <https://arxiv.org/abs/1904.02790> accessed 27 July 2025.

78 Fazel et al (2021), *supra* note 74, p. 1, 4.

79 Agencia Española Protección Datos, 'Synthetic data and data protection' (2 November 2023) AEPD Innovation and Technology Division <https://www.aepd.es/en/prensa-y-comunicacion/blog/synthetic-data-and-data-protection> accessed 27 July 2025.

80 Kanwal (12 December 2023), *supra* note 76, p. 3.

81 Michèle Fink and Frank Pallas, 'They who must not be identified – distinguishing personal from non-personal data under the GDPR' (2020) *International Data Privacy Law* 10(1) p. 29 <https://academic.oup.com/idpl/article/10/1/11/5802594> accessed 27 July 2025.

82 Agencia Española Protección Datos (2 November 2023), *supra* note 79.

83 Brendan J. Frey, Geoffrey E. Hinton and Peter Dayan, 'Does the wake-sleep algorithm learn good density estimators?' in D. Touretzky, M. Mozer and M. Hasselmo (eds) *Advances in Neural Information Processing Systems 8* (NIPS, 1996) (MIT Press, Cambridge, MA) pp. 661-666 <https://proceedings.neurips.cc/paper_files/paper/1995/file/55b1927fdafef39c48e5b73b5d61ea60-Paper.pdf> accessed 27 July 2025.

84 Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior and Koray Kavukcuoglu 'Wavenet: A generative model for raw audio' (2016) <https://arxiv.org/abs/1609.03499> accessed 27 July 2025.

85 *Ibid*.

86 OECD Digital Economy Papers, 'AI Language Models: Technological, Socio-Economic and Policy Considerations' (2023) pp. 14-22 <https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/04/ai-language-models_46d9d9b4/13d38f92-en.pdf> accessed 27 July 2025.

87 Kalpana Tyagi, 'Copyright, text & data mining and the innovation dimension of generative AI' (2024) *Journal of Intellectual Property Law & Practice* 19(7) pp. 559-562 <https://academic.oup.com/jiplp/article/19/7/557/7624901> accessed 27 July 2025.

88 Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherijl Ozair, Aaron Courville and Yoshua Bengio, 'Generative adversarial networks' Advances in neural information processing systems' (2014) <https://arxiv.org/abs/1406.2661> accessed 27 July 2025; For a discussion on the interface between deep fakes, synthetic data and personality rights, see Kalpana Tyagi 'Deepfakes, Copyright and Personality Rights: An Inter-disciplinary Perspective' in Klaus Mathias and Avilasham Tor (eds)

26 GAN as an approach to machine learning offered an important thrust to the GenAI revolution that we are currently witnessing. In GAN, a training dataset is used to train models to generate samples that are varying representations of the original data inputs.[89] GAN was quickly adopted as a standard tool in machine learning, as it substantially multiplied the input and output points, and thus accelerated the rate of output generation.[90] The basic GAN model used supervised learning to approximate actual functions and posited a sustainable technological model for 'image generation and manipulation systems'.[91] Learning may be supervised, unsupervised or reinforced. The resilience of GAN is that it has been updated over time, and there are many variations of GAN that can be adapted and used across all these different approaches to machine learning. Interactive GANs, for instance, are applications used for creating images, whereby input data is used to train models to create similar realistic images.[92] The unsupervised GAN model includes a 'generative' neural network and a 'discriminative' neural network. The generative neural network creates the noise, such as by generating correct and incorrect outputs. The discriminative neural network assesses the factual correctness to ascertain which of the given outputs are factually correct. To visualize how GANs function in practice, consider the generative neural network as the teacher that offers an exam with multiple choice questions to the students, the discriminative neural network. The student has over the course of the year learnt from books, class notes and lectures, which is the equivalent of training data. Based on this learning, the student learns to discriminate amongst correct and incorrect options. The process is iterated until the discriminative neural network learns to correctly distinguish the 'noise' from the data.

27 VAE is another scientifically robust approach for variational learning in deep generative models. Stable Diffusion, a GenAI model, offers realistic images, videos and animations, with a mere text and image prompt. Stable Diffusion uses 'variational autoencoder, forward and reverse diffusion, a noise predictor and text conditioning' to perform its function.[93] VAE translates human expressions into mathematical representations and distribute them over a range of functions. As an example, the human expression of smile may be attached with a higher probability to parodied works and a thoughtful expression be assigned with a higher probability to quotations and news reporting. The mathematical expressions are then coded and decoded to generate images. Stable Diffusion V1 was trained on LAION's datasets using Common Crawl. This process involved scrapping billions of images and text that are protected by copyright and related rights. The following section 3 further develops this discussion in the context of copyright, and other intellectual property rights.

C. Copyright: Synthetic Data, Text and Data Mining and Follow-on Works

28 Intellectual property rights (IPRs) and notably copyright and related rights have an important interplay with GenAI and synthetic data. The key question as regards synthetic data and copyright is whether the process of generation of synthetic data, and the output of GenAI models, namely the synthetic data itself, infringe copyright. To answer this, it is important to look at how GenAI models text and data mine to make inferences, draw correlation and generate new works (Section 3.1). Sub-section 3.1.1 addresses the scope of TDM as discussed by the German regional court in *Robert Kneschke v. LAION*. Subsection 3.1.2 develops the scope and meaning of opt-outs under Article 4(3), 2019 CDSM, and its interpretation thereof by the German court in light of the 2024 EU AI Act (2024 EU AIA). Section 3.2 discusses whether synthetic data infringes copyright, and whether the 2024 EU AIA can safeguard rightholders of the original works from synthetically generated data, that is at some point in the data value chain, based on the original human-generated data.

I. GenAI, Text and Data Mining and Synthetic Data in the EU

29 GenAI involves two key phases – namely, the input/training phase and the output phase.[94] The training

Law and Economics of the Digital Transformation (2023) (Economic Analysis of Law in European Legal Scholarship 15, Springer Switzerland) <https://link.springer.com/chapter/10.1007/978-3-031-25059-0_9> accessed 27 July 2025.

89 Ian Goodfellow (3 April 2017) 'NIPS 2016 Tutorial: General Adversarial Networks' Open AI pp. 2-3 <https://arxiv.org/abs/1701.00160> accessed 27 July 2025.

90 *Ibid.*

91 *Ibid.*, p. 51.

92 J.Y. Zhu, P. Krähenbühl, E. Shechtman and A.A. Efros (2016) 'Generative visual manipulation on the natural image manifold' *European Conference on Computer Vision* pp. 597-613 (Springer) <https://arxiv.org/abs/1609.03552>

93 Amazon Web Services 'What is Stable Diffusion?' <https://aws.amazon.com/what-is/stable-diffusion/> accessed 27 July 2025.

94 Eleonora Rosati, 'Infringing AI: Liability for AI-generated outputs under international, EU, and UK copyright law' (June 2025) *European Journal of Risk Regulation* 16(2) p. 611 <https://doi.org/10.1017/err.2024.72> accessed 27 July

phase involves text and data mining (TDM). TDM may be defined as 'any automated analytical technique aimed at analysing text and data in digital form which includes but is not limited to patterns, trends and correlations'.[95] Data is the input required to train these GenAI (including LLM) models. ChatGPT, one of the fastest adopted and most popular GenAI models, for example, was trained on '300 billion words systematically scraped from the internet'.[96] This included both copyright-protected content, such as books, works, and poems, as well as personal data, such as posts by users on social and professional networking sites, such as Facebook, Instagram, and LinkedIn.[97] Whereas works are subject to copyright, personal data is subject to GDPR. The alleged infringing use of data in the input/training phase is a key complaint in the GenAI-related cases, currently pending before the US courts, and the case of Stability AI, pending before the UK courts. In the EU, text and data mining (TDM) is covered by Articles 3 and 4, 2019 Copyright in the Digital Single Market Directive (2019 CDSM).[98] Article 3 offers an exception for the right of reproduction and extraction made under the Database Directive 96/9/EC, 2001 InfoSoc Directive, and the press publishers right under Article 15 of the 2019 CDSM. Article 4, 2019 CDSM, in addition to these rights, also offers an exemption from the right of reproduction and translation, adaptions and alterations of a computer programme under Article 4(1)(a) and (b) of the 2009 Computer Programmes Directive (CPD). As per Article 4 (1)(a) and (b), 2009 CPD, the rightholder of the computer programme has the right to authorize any permanent or temporary reproduction, translation, adaptation, arrangement, or any other alteration of a computer program. Article 4 of the 2009 CPD is, however, subject to Articles 5 and 6, and offers a set of restricted rights available for computer programmes. Article 5(3) in particular authorizes the user who has a right to use a copy of a computer program, the possibility to 'observe, study or test the functioning of the programme' without seeking an explicit permission from the rightholder. One therefore sees an interplay between Article 4, 2019 CDSM and Article 5 of the 2009, CPD. Article 5(3) of the CPD offers a 'black box exception' to study and evaluate the basic ideas and principles of the program.[99] It must, however, be added, that this research exception under Article 5(3), 2009 CPD is only limited to study the underlying principles, ideas and designs that underlie the programme. In other words, TDM that falls outside the scope of research purposes is not covered under Article 5(3), 2009 CPD. In that respect, Article 4, 2019 CDSM clarifies that TDM beyond those specified purposes are permitted provided that the rightholder has not limited this possibility, such as through the imposition of 'machine-readable means' or any other such reservation[100] an issue further developed in sub-section 3.1.2 *infra.*

1. *Robert Kneschke v. LAION*: EU's First Decision on TDM

30 Article 25 of the 2019 CDSM emphasizes the minimum harmonizing nature of exceptions and limitations (E&Ls), meaning that Member States may adopt a broader TDM exception than the one prescribed in the 2019 CDSM. Further, when the matter reaches the CJEU regarding the scope and interpretation of the TDM exception, it may be an opportunity to offer a wider meaning to the exception, as, for example, happened with exceptions under the 2001 Information Society Directive.[101] Meanwhile, the German national implementation of Articles 3 and 4, 2019 CDSM, recently reached the Hamburg regional court for interpretation and the Court's interpretation of Article 3, along with *obiter dicta* regarding Article 4, seem to indicate the direction of one such flexible interpretation. Articles 3 and 4, 2019 CDSM, alongside other provisions of the said directive, were to be transposed by the EU Member States in their national legislation by 7 June 2021. Germany also transposed the said provisions into its Copyright Act, Urheberrechtsgesetz (UrhG). Section 60d UrhG is the German equivalent of Article 3,

2025.

95 Article 2(2), 2019 CDSM.

96 Uri Gal, 'ChatGPT is a data privacy nightmare. If you've posted online, you ought to be concerned' *The Conversation* (Online 10 February 2023) <https://theconversation.com/chatgpt-is-a-data-privacy-nightmare-if-youve-ever-posted-online-you-ought-to-be-concerned-199283> accessed 27 July 2025.

97 *Ibid.*

98 See Tyagi (2024), *supra* note 87, pp.9-11; Rosati (2024), *supra* note 94, pp. 2,6-9 on the scope of Articles 3 and 4, 2019 CDSM.

99 Rossana Ducato and Alain Strowel, 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out' (2021) *European Intellectual Property Review* 43 (5) p.11 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3278901> accessed 27 July 2025.

100 *Ibid.*, at pp. 14-15.

101 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 27 July 2025. It may be interesting to follow the developments in Like Company v Google, C-250/25, a request for preliminary ruling on the scope and interpretation of Articles 4 and 15 of the 2019, CDSM, currently pending before the CJEU.

2019 CDSM and permits TDM for non-commercial scientific research purposes undertaken by research organizations such as universities and cultural heritage institutions. Section 44b UrhG implements Article 4, 2019 CDSM, which allows commercial TDM by private enterprises.

31 In its decision in *Robert Kneschke v. LAION*, dated 27 September 2024, the German Regional Court of Hamburg (Landgericht) offered an interpretation of the scope of Section 60d UrhG. The case offers clarity on an important issue, namely whether 'AI data scrapping' can qualify as TDM.[102] It may be useful to add that the choice of Section 60d, and not 44b, made a significant impact in determining the outcome of the case. The Court also discussed the scope of Section 44a UrhG, which transposes Article 5(1) of the 2001 Information Society Directive, which covers only temporary and transient acts of reproduction, and is the only mandatory E&L in the 2001 Information Society Directive. However, the said section was found inapplicable in light of the scope of the right covered in the case at hand. Common to all these copyright provisions covering the act of transient copyright (Article 5(1), 2001 InfoSoc Directive) and text and data mining (Articles 3 and 4, 2019 CDSM), is the prerequisite of lawful access, meaning that the user must have lawful access to the copyright-protected works. This is an important fundamental pre-requisite that also impacts whether the process of generating synthetic data, and synthetic data itself, may infringe copyright, an issue developed in Section 5.2 *infra.*

32 The facts in *Robert Kneschke v. LAION* may be briefly described as follows. LAION (Large-scale Artificial Intelligence Open Network), a German-registered not-for-profit firm, created a LAION-5B training dataset, comprising '5.85 billion database-filtered image-text pairs' that it then made available without any access restrictions on its website.[103] The database, available for free on the website, did not contain any images. It only included image descriptions and hyperlinks to the image sources at the time of the creation of the database.

33 Robert Kneschke, the Plaintiff, claimed that LAION had infringed his copyright by scraping his photographs from a stock photo website, bigstock.com. Photographs can be protected as works if they are original, that is, have the 'author's personal touch' [104] or otherwise, under Article 6 of the Term of Protection Directive[105]. The 'image was freely available without a paywall' on the said website.[106] LAION copied only the watermarked versions of Kneschke's photographs that were freely accessible. [107]

34 The Court dismissed Kneschke's cease-and-desist request because, in the Court's opinion, LAION, being a not-for-profit firm, could benefit from the scientific research exception for TDM (Section 60d UrhG, equivalent EU provision Article 4, 2019 CDSM). Kneschke also alleged that LAION received funding from for-profit firms, and that the output of the TDM process, namely the dataset, was used by commercial for-profit firms. The Court, however, was of the opinion that the fact that two LAION members also worked for commercial firms, such as Stability AI, or that Stability AI contributed to financing the LAION-5B dataset, did not automatically translate into 'preferential access [by private enterprises, such as Stability AI] to the findings of LAION's scientific research'.[108] The Court found that non-profit entities such as LAION and Common Crawl contribute to decoding the black box of the data on which GenAI models are trained, how they work, and develop datasets that can then be used by commercial developers.[109] The Court dismissed the profit-driven

102 Ronak Kalhor-Witzel, 'German Court Says Non-Commercial AI Training Data Meets Scientific Research Exception to Copyright Infringement' *IP WatchDog* (Online 10 October 2024) <https://ipwatchdog.com/2024/10/10/german-court-non-commercial-ai-training-data-meets-scientific-research-exception-copyright-infringement/id=182008/#> accessed 27 July 2025.

103 Kristina Ehle and Yeşim Tüzün, 'To Scrape or Not to Scrape? First Court Decision on the EU Copyright Exception for Text and Data Mining in Germany' *Morrison Foerster: Client Alert* (Online 4 October 2024) <https://www.mofo.com/resources/insights/241004-to-scrape-or-not-to-scrape-first-court-decision> accessed 27 July 2025.

104 Case C-145/10 *Eva Maria Painer v. Standard VerlagsGmbH and Others*, paras 87-88.

105 Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32006L0116
Article 6, Protection of photographs 'Photographs which are original in the sense that they are the author's own intellectual creation shall be protected in accordance with Article 1. No other criteria shall be applied to determine their eligibility for protection. Member States may provide for the protection of other photographs.'

106 Paul Keller, 'Machine readable or not? – notes on the hearing in LAION e.v. vs Kneschke' *Kluwer Copyright* (Online 22 July 2024) <https://copyrightblog.kluweriplaw.com/2024/07/22/machine-readable-or-not-notes-on-the-hearing-in-laion-e-v-vs-kneschke/> accessed 27 July 2025.

107 Ehle and Tüzün (Online 4 October 2024), *supra* note 103.

108 Mirko Brüß, 'German court finds LAION's copying of images non-infringing' *IPKat* (Online 28 September 2024) <https://ipkitten.blogspot.com/2024/09/guest-post-german-court-finds-laions.html> accessed 27 July 2025.

109 Paul Keller, 'LAION vs Kneschke: Building public datasets is covered by the TDM exception' *COMMUNIA* (Online 11 October 2024) <https://communia-association.org/2024/10/11/laion-vs-kneschke-building-public-

nature of LAION on the grounds that the dataset was made 'freely available to the public' and that subsequent commercialization by another for-profit firm 'is irrelevant for [the] assessment under Section 60d(2) UrhG'.[110] Moreover, the Plaintiff, in the Court's opinion, could not establish that these commercial firms exercised 'a decisive influence' or had 'preferential access to the findings of [LAION's] scientific research'.[111]

35 The decision of the Hamburg court was soon subject to criticism for choosing the wrong legal basis. As the dataset offered only hyperlinks made available following the completion of the TDM activity, Rosati is of the opinion that the issue should have been properly dealt with under Articles 2 and 3, 2001 Information Society Directive.[112] Article 3, 2019 CDSM covers only TDM and not the acts 'following the completion of TDM activities'.[113] In the case at hand, this follow-on act was that LAION made the dataset freely available on its website, without any access or usage restrictions. Prof. Rosati explains how TDM is limited to certain economic rights as described within the scope of Articles 3 and 4, 2019 CDSM; however, the subsequent acts, such as the one in the present case, are not covered by these articles.[114] With the new dataset available on its website, LAION performed an act of communication and making available to the public. In a string of case law, starting with *Svensson*, and later *GS Media* and *VG Bild-Kunst*, the CJEU has stated that the 'link provider's own knowledge… even when the link in question [has been offered] for non-profit purposes' is relevant for a finding of infringement.[115] While Prof. Rosati's remarks merit detailed academic discussion, in light of their insightful awareness of the choice of the correct legal basis, the Hamburg court's decision, at least until the pending appeal is heard[116], remains a good legal precedent on web crawling for access to data to train GenAI models. Once trained using crawled data, these GenAI models also generate synthetic data based on the inputs from human-generated data. If such training from web-scraped data is exempted early on, then follow-on synthetic works may have minimal realistic chances of being caught by copyright infringement rules.

36 Interestingly, the Hamburg Court did not stop there. It went a step further and also elaborated on the scope of the opt-out for machine learning in light of the provisions of the AI Act.

2. Article 53(1)(c), EU AI Act and Article 4(3), 2019 CDSM

37 Article 4(3) of the 2019 CDSM permits rightholders to opt out their copyright-protected works and prevent AI model developers from using their works for training purposes. In *Robert Kneschke v. LAION*, as bigstock.com's terms of service restricted use of automated programs, the Hamburg court found this to be a clear communication of an opt-out for the purposes of Article 4(3), 2019 CDSM. The difference in the opinion of the Plaintiff and the Defendant can be summarized as follows. Kneschke argued that 'digital plain text [being] sufficiently readable' sufficed to express an opt-out, whereas LAION argued that 'to be considered machine readable, an opt-out should be provided in a specific standardized format (in this case, robots.txt) that can be easily understood by crawlers and other bots'.[117] The Hamburg court assessed the facts of the case in light of the provisions of Article 53(1)(c), EU AI Act, which suggests that opt-outs may be exercised in light of the available state-of-the-art technology, and opined that a liberal approach should be taken as regards the machine readability of an opt-out by the rightholder. Furthermore, the Court was of the opinion that as natural language processing tools advance, an opt-out in 'words,' such as the one made by Mr Kneschke may suffice to meet the requirements of Section 44b UrhG. This observation resonates with the discussion in section 2 *supra,* which illustrates how GenAI models, starting with BERT, can bidirectionally read text and therefore contextualise linguistic content. The approach suggested by the Court to the reading of opt-outs during crawling is a rightholder-friendly one. Considering the rapid pace of technological advancements, the question is: what should be the relevant time frame that serves as a benchmark to

datasets-is-covered-by-the-tdm-exception/> accessed 27 July 2025.

110 Simon Hembt, Niels Lutzhöft and Toby Bond, 'Long-awaited German judgment by the District Court of Hamburg (Kneschke v. LAION) on the text and data mining exception(s)' *Bird & Bird* (Online 1 October 2024) <https://www.twobirds.com/en/insights/2024/germany/long-awaited-german-judgment-by-the-district-court-of-hamburg-kneschke-v-laion> accessed 27 July 2025.

111 Ehle and Tüzün (4 October 2024), *supra* note 103.

112 Eleonora Rosati, 'The German LAION decision: A problematic understanding of the scope of the TDM copyright exceptions and the transition from TDM to AI training' *IP Kat* (Online 7 October 2024) <https://ipkitten.blogspot.com/2024/10/the-german-laion-decision-problematic.html> accessed 27 July 2025.

113 *Ibid.*

114 Rosati (June 2025), *supra* note 94, p. 612.

115 Rosati (2024), *supra* note 112.

116 The decision has been appealed before a higher court. CEPIC 'CEPIC supports Robert Kneschke in his copyright lawsuit against LAION and welcomes the appeal' <https://www.cepic.org/post/cepic-supports-robert-kneschke-in-his-copyright-lawsuit-against-laion-and-welcomes-the-appeal> accessed 27 July 2025.

117 Keller (22 July 2024), *supra* note 106.

assess infringement? In *LAION*, the Hamburg Court took into account the state of the technology at the time of the decision (by which point ChatGPT had already been released) and not the technology in use around 2022, when LAION crawled the website to scrape data.[118] However, as the case was decided under Article 3, and not Article 4, the opinion of the Court as regards the format of the opt-out under Article 4 is *obiter dicta,* and the issue remains unresolved. The Hamburg court's *obiter dicta* may serve as a useful reference point for the AI Office to clarify the provisions of Article 53 of the 2024 EU AIA. Another related issue is whether there should be a harmonized and recognized legal standard, such as in the form of 'robots.txt', to standardize web instructions for crawlers.[119] Traditionally, robots.txt has been a *de facto* accepted standard to prevent crawling of a given website. Robots.txt files, usually located at 'websiteaddress.com/robots.txt' are like notices at the door entrance, indicating who is permitted, or restrained, from entering the room. Robots.txt has been a *de facto* web standard for decades, whereby robots (also known as crawlers, worms, or web crawlers) of search engines took it as a signal as to whether they were permitted to crawl a given website.[120] Until the advent of GenAI, there was a *quid pro quo* between websites and Google, whereby Google could crawl and index these websites and display them in the search results.[121] Websites benefitted from appearing in the search results, and crawling was thus seen to bring benefits for both the website and the search engines.[122] With the advent of GenAI, however, it turned against even the interests of formerly robot.txt compliant firms like Google to conform to the instructions on the entrance door, namely the 'robots.txt', of the website. Presently, not only do Google, and GenAI firms, crawl through these webpages, but they also fail to offer any credit to the website owners for content extracted from their pages.[123] In this respect, the Hamburg Court's *obiter dicta* on opt-outs under Article 4(3), 2019 CDSM, in *Robert Kneschke v. LAION* become relevant.[124] Notably, the Court's opinion that advanced GenAI can understand text in plain human language, and therefore, such a communication of an opt-out should suffice for the purposes of Article 4(3) 2019 CDSM and AI Act is a positive development in line with the balancing of authors' rights vis-à-vis users' rights.

II. GenAI, Text and Data Mining and Synthetic Data: Infringing or non-Infringing?

38 Following the completion of text and data mining, GenAI tools generate data. GenAI tools, such as ChatGPT, accelerate the rate and quality of output generation.

39 To generate synthetic data using Deep Generative Models (DGM), data is a key input. This data, in turn, may have different components. Data, as discussed in Section 1 *supra*, can be of many different types, and may include 'unprotected data' such as raw data, or 'protected works of authorship and other protected subject matter'.[125] Copyright only protects expressions of works that are original. Article 2 of the Berne Convention offers an open-ended definition of works, and includes, 'literary and artistic works' in any mode or form of expression. Article 9(2) of the TRIPS agreement clearly brings out this copyright protection for expression in works that are original. This implies that 'unstructured raw data', such as 'mere facts and data "as such"' does not benefit from copyright protection.[126] To qualify for protection, these works must be original, and they must be an expression, and not a mere idea. Copyright 'subsists not in ideas, but in the form in which the ideas are expressed'.[127]

40 While there is no explicit reference to protection

118 Kalhor-Witzel (10 October 2024)), *supra* note 102.

119 Paul Keller and Zuzanna Warso, 'Defining Best Practices for Opting out of ML Training' *Open Future Policy Brief* #5 (Online 29 September 2023) <https://openfuture.eu/wp-content/uploads/2023/09/Best-_practices_for_optout_ML_training.pdf> accessed 27 July 2025.

120 David Pierce 'The text file that runs the internet' *The Verge* (Online 14 February 2024) <https://www.theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders> accessed 27 July 2025.

121 *Ibid.*

122 This symbiotic relationship between websites and search engines, such as Google also draws a parallel with the relationship between press publishers and Google. In the press publishing industry as well, the emergence of GenAI seems to have altered the dynamics of the relationship. Cf Kalpana Tyagi, 'Generative AI, EU press publishers' rights & the Australian News Bargaining approach: Copyright & Competition law as enablers of media plurality & diversity of opinion' (Forthcoming, 2024, pre-print) *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4933421> accessed 27 July 2025.

123 Pierce (14 February 2024), *supra* note 120.

124 Stepanka Havlikova, 'Technical Challenges of Rightholders' Opt-Out from Gen AI Training after Robert Kneschke v. LAION' (2025) *JIPITEC* 16(1) <https://www.jipitec.eu/jipitec/article/view/422> accessed 27 July 2025.

125 Thomas Margoni 'TDM and Generative AI: Lawful Access and opt-outs' *Auteurs & Media* (forthcoming, pre-print 2024) p. 3 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5036164> accessed 27 July 2025.

126 *Ibid.* p.7.

127 *Designers Guild Ltd. V. Russell William (Textiles) Ltd.* (2000) UKHL 58, (2011) 1 WLR 2416.

of 'collections of data' in the Berne Convention, however, Article 2(5) affords protection to 'collections of literary or artistic works', which offers the possibility for protection to the extent the collection is an 'intellectual creation'.[128] Different jurisdictions have different thresholds for originality. In the EU, the work is deemed original, if it is the author's own intellectual creation[129], in other words, if it carries the author's personal touch. Databases, when curated or arranged in such a manner that constitute 'the author's own intellectual creation' can also benefit from copyright protection.[130] In addition, Article 10(2) of the TRIPS offers Contracting Parties the possibility to offer protection to compilations of data by virtue of the investment made in the creation of databases or other similar material. It reads thus:

> *Compilations of data or other material, whether in machine readable or other form, which by reason of the selection of arrangement of their contents constitute intellectual creations shall be protected as such. Such protection, which shall not extend to the data or material itself, shall be without prejudice to any copyright subsisting in the data or material itself.*

41 This is the basis on which the EU offers a *sui generis* database right, wherein database are protected for the investment made in the creation and compilation of these database.[131] The Database Directive thus, offers copyright protection for compilations of data, and also offers a 'non-copyright, *sui generis* right in databases to protect the investment of the database maker'.[132] Substantial qualitative or quantitative extraction or reutilization of the database, can therefore, lead to the infringement of the *sui generis* database right.[133]

42 From the lens of copyright and related rights and *sui generis database rights*, unauthorized use of protected works, may mean that both the synthetic data and the system 'training on it' are infringing uses.[134] As the value chain of the data elongates, for example, when synthetic data is generated from human generated data, and further synthetic data is generated using varying proportions of human generated and synthetically-generated data, how does one know whether human-generated data has been used in the GenAI value chain for the generation of this synthetic output? If the synthetically generated data is similar, or bears resemblance to the human generated-data, then it may be possible to assess this through infringement tests, and by showing similarity between the human-generated data that may have been used to train the GenAI model to produce this synthetic output, infringement can be ascertained. This is also a key issue in the ongoing GenAI-related cases. In the Authors Guild/OpenAI case for instance, the Plaintiff, Authors Guild offers examples of how ChatGPT can offer precise summaries and excerpts from copyright-protected works.[135] In the case at hand, ChatGPT could not only accurately summarize the works of authors such as John Grisham, it could also create follow-on works. For example, ChatGPT offered a realistic follow-on novel to the Grisham's famous work, 'The King of Torts' and offered it an equally plausible title, namely 'The Kingdom of Consequences'. [136] While these outputs, in this case, 'The Kingdom of Consequences', may be closer to the original human-generated works, and hence, easier to be identified as infringing, how does one determine infringement, when the GenAI tools offer follow-on works, that are generated based on the synthetically-generated works, and that the human author may have likely created himself? In other words, what happens when synthetic data is used to generate successive generations of synthetic data?

128 Sam Ricketson and Jane C. Ginsburg, *International Copyright and Neighbouring Rights: The Berne Convention and Beyond* (3rd edition, Oxford University Press 2022) pp. 489-490.

129 Case C-5/08 *Infopaq International A/S v Dansk Dagblads Forening*, paras 37, 45, 47.

130 Article 3(1), Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (Database Directive) OJ 1996 L 77.

131 Case C 203/02 *British Horseracing Board v. William Hill Organisation Ltd*, para 31 usw; Case C-338/02 *Fixtures Marketing Ltd. V. Svenska Spel AB*, paras 24-29.

132 Daniel J. Gervais, 'The Protection of Databases' (2007) 82 *Chicago-Kent Law Review* p. 1120 <https://scholarship.law.vanderbilt.edu/faculty-publications/839/> accessed 27 July 2025; Estelle Derclaye, (2002) 'What is a Database? A Critical Analysis of the Definition of a Database in the European Database Directive and Suggestions for an International Definition' 5 *Journal of World Intellectual Property* p. 981 <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1747-1796.2002.tb00189.x> accessed 27 July 2025.

133 Gervais (2007), *supra* note 132, p.1123.

134 Lee (2024), *supra* note 38, p. 24.

135 For a discussion on the future of work in an age of generative AI in creative industries, see discussion on Authors Guild v. OpenAI (Complaint filed on 19 September 2023) No. 1:23-cv-8292, Kalpana Tyagi 'Redefining a Normative Framework for Meritocracy in the Era of Generative AI: An Inter-Disciplinary Perspective' in Klaus Mathis and Avishalam Tor (eds), *Law and Economics of Justice: Efficiency, Reciprocity and Meritocracy* (Springer 2024) <https://www.springerprofessional.de/en/redefining-a-normative-framework-for-meritocracy-in-the-era-of-g/27041164> accessed 27 July 2025.

136 See discussion on Authors Guild v. OpenAI (Complaint filed on 19 September 2023) No. 1:23-cv-8292 in Kalpana Tyagi 'Redefining a Normative Framework for Meritocracy in the Era of Generative AI: An Inter-Disciplinary Perspective' in Klaus Mathis and Avishalam Tor (eds), *Law and Economics of Justice: Efficiency, Reciprocity and Meritocracy* (Springer 2024) <https://www.springerprofessional.de/en/redefining-a-normative-framework-for-meritocracy-in-the-era-of-g/27041164> accessed 27 July 2025.

This may be particularly true for fiction and art-based works, whereby 'hallucination'[137], abstraction and 'creativity' are closely intertwined. This is distinct from scientific facts and assertions, which may at least, relatively speaking, be easier to identify and are grounded in research. The original human-generated data becomes sequentially distanced from the GenAI value chain, making it increasingly difficult to establish infringement. It may be even more difficult to impose liability in the US, where the GenAI developers are likely to benefit from the 'transformativeness' test, included in the four-factor fair use test.[138] Even if the models simply extend and correlate 'specific memorized examples within the training data, the output will only infringe if enough original expression of any particular example were evident in the final product'.[139] Article 53(1)(d) 2024 EU AI Act offers an important safeguard as it suggests that the providers of general-purpose AI models shall

> *draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office.*

43 Article 53(1)(d) of the Act thereby requires transparency regarding the data used to train the General-Purpose AI (GPAI) models. This includes datasets and data sources that are protected not only by copyright, but also by other legal frameworks, such as data protection laws, discussed below. Notably, as regards synthetic data, Warso, Gahntz and Keller offer a valuable proposition for the implementation of Article 53(1)(d). In the AI transparency blueprint, the authors suggest a detailed plan as regards datasets that must be mentioned, and how they can be sufficiently described. The authors also suggest that Article 53(1)(d) can be used to request information on synthetically generated data by the model provider, including the time of generation and methods used to create the synthetic output.[140] In its recent guidance, the EU Commission does seem to follow this approach, as the template also offers a section to include details about the use of synthetic data to train the model.[141]

44 Reference to not only human-generated works and datasets, but also to synthetic datasets, within the possibilities offered by the 2024 EU AIA, will, in the author's opinion, make the Act even more human-centric. Thus, the relationship between the AI Act, GenAI and copyright is a very special one[142], which can be positively leveraged to bring to the surface the human element within the 2024 EU AI Act. This can be attributed to the inherent nature of these two fields of law: whereas the AI Act is public law; copyright, like other IPRs, is private law. While obligations under the AI Act are monitored by the AI Office, a part of the European Commission, breach of copyright is subject to private enforcement by copyright holders.[143] Interestingly, as the above-suggested blueprint indicates, provisions of the 2024 EU AIA can be effectively deployed to remunerate the human author, even when the output is derived from synthetically-generated data. Remuneration of the human author, alongside text and data mining, are, generally speaking, two of the key concerns as regards training of generative AI models.[144] Article 17 of the Charter of Fundamental Rights (CFR) refers to the right to property. Notably, Article 17(2) explicitly refers to protection of intellectual property. The other relevant right, particularly in the context of copyright and related rights, is the right to freedom of expression. The discussion on fundamental rights is gaining prominence in the intellectual property discourse, and has led to the constitutionalisation of intellectual property rights. In *Poland v. Parliament*, Poland challenged the new liability regime against online content sharing service providers under Article 17, 2019 CDSM on the grounds it violated the principles enshrined in the CFR.[145] The Court was of

137 Hallucination is a frequently cited limitation of GenAI. Cf Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chenjian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song and Bo Li, 'Decoding Trust: A Comprehensive Assessment of Trustworthiness in GPT Models' (20 June 2023) *NeurIPS 2023 Outstanding Paper* pp. 6,11 <https://arxiv.org/abs/2306.11698> accessed 27 July 2025.

138 Cf Andrea Bartz v. Antropic (2025), *supra* note 7.

139 Matthew Sag, 'Copyright Safety for Generative AI' (2023) *Houston Law Review* 61(2) p. 312, 322 <https://houstonlawreview.org/article/92126-copyright-safety-for-generative-ai> accessed 27 July 2025.

140 Zuzanna Warso, Maximilian Gahntz and Paul Keller, 'Blueprint of the template for the summary of content used to train general-purpose AI models (Article 53(1)d AIA) – v.2.0' (2024) *Open Future Foundation* p. 2 <https://openfuture.eu/wp-content/uploads/2024/09/240919AIAtransparency_template_requirements-blueprint_v.2.0.pdf> accessed 27 July 2025.

141 European Commission (24 July 2025), *supra* note 47.

142 Quintais (2025), *supra* note 46, pp. 7-8.

143 *Ibid.*

144 Kalpana Tyagi, 'Generative AI: Remunerating the human author & the limits of a narrow TDM Exception' (13 December 2023) *Kluwer Copyright Blog* <https://copyrightblog.kluweriplaw.com/2023/12/13/generative-ai-remunerating-the-human-author-the-limits-of-a-narrow-tdm-exception/> accessed 27 July 2025.

145 Maria Alexandra Mărginean, *Republic of Poland v. European Parliament and Council of the European Union: Balancing Freedom of Expression with the Filtering Obligations of Article 17 of the DSM Directive* (Bachelor Thesis, Maastricht University 2022); Adrien Dubois, *A comparative analysis of Article 17 CDSM national implementation and Poland v Commission's ECJ Framework* (Master Thesis, Maastricht University 2022).

the opinion that when there are multiple possible interpretations, the one that best complies with fundamental rights should be preferred.[146]

45 Considering the use of human-generated works in the training of GenAI models, copyright scholars have consistently called for a framework to adequately remunerate the human author.[147] Prof. Senftleben, for instance, presents an interesting proposal that a levy, 'an AI levy' can be imposed on GenAI systems that produce literary and artistic outputs.[148] Article 53(1)(d) of the EU AIA 2024, can serve as a useful complement to author remuneration, and to make this effort coherent and effective, Prof. Senftleben's suggestion for an 'AI levy' could provide the required revenues, which may then be equitably and proportionately distributed amongst rightholders.

46 Thus, transparency as regards datasets, under Article 53(1)(d) the EU AIA 2024, will also go a long way in safeguarding the authors' rights. With a requirement that datasets not only refer to human-generated works and other data (including personal data), but also to synthetically-generated data, as mentioned above, identifying the human author, even when high up in the data value chain, can serve as a useful complement for legal enforcement and timely, adequate, and proportionate remuneration of the human author. As a balancing and proportional framework, the rules may prescribe a proportional remuneration for the human author, based on how closely they can be linked in the value chain to the output generated.

47 In addition to copyright and the human author, the 2024 EU AI Act and the EU CFR also connect with personal data protection, that is palpably grounded in the right to privacy and data protection, an issue that we turn to next.

D. Data Protection and Privacy: Promises and Concerns

48 The 2016 GDPR is triggered when the processing of personal data is involved. The EU Data laws clearly specify that unless otherwise specified, the 'provisions are "without prejudice" to ... personal data protection and intellectual property rights [except for *sui generis* database rights]'.[149] Sub-section 4.1 highlights the fundamental rights-driven nature of the GDPR. Sub-section 4.2 assesses how and when the GDPR is triggered while developing the GenAI models, and what safeguards in the 2024 EU AI Act are relevant thereto. Sub-section 4.3 assesses whether synthetic data can help effectively comply with the principles of data protection. From an innovation perspective, synthetic data can play a pertinent role as training GenAI models requires deep learning, and once learnt, it may be subsequently difficult (or perhaps even impossible) for data subjects to exercise their individual rights, such as the right of erasure and portability.

I. The Fundamental Rights-Driven Nature of the GDPR

49 The GDPR concerns personal data and is grounded in respect for fundamental rights.[150] This is clear from the

Copy available with the author upon request.

146 Case C-401/19 *Poland v European Parliament and Council*, Judgment of the Court (Grand Chamber) 26 April 2-22, EU:C:2022:297.

147 Cf Christophe Geiger and Vincenzo Iaia, 'Generative AI, Digital Constitutionalism and Copyright: Towards a Statutory Remuneration Right grounded in Fundamental Rights – Part 1' (17 October 2023) *Kluwer Copyright Blog* <https://legalblogs.wolterskluwer.com/copyright-blog/generative-ai-digital-constitutionalism-and-copyright-towards-a-statutory-remuneration-right-grounded-in-fundamental-rights-part-1/> accessed 27 July 2025; Christophe Geiger and Vincenzo Iaia, 'Generative AI, Digital Constitutionalism and Copyright: Towards a Statutory Remuneration Right grounded in Fundamental Rights – Part 2' (19 October 2023) *Kluwer Copyright Blog* <https://legalblogs.wolterskluwer.com/copyright-blog/generative-ai-digital-constitutionalism-and-copyright-towards-a-statutory-remuneration-right-grounded-in-fundamental-rights-part-2/> accessed 27 July 2025 and Martin Senftleben (2023) 'Generative AI and Author Remuneration' *International Review of Intellectual Property and Competition Law* Vol. 54 <https://link.springer.com/article/10.1007/s40319-023-01399-4> accessed 27 July 2025.

148 Senftleben (2023), *supra* note 147; Martin Senftleben (2022) 'A Tax on Machines for the Purpose of Giving Bounty to the Dethroned Human Author – Towards an AI Levy for the Substitution of Literary and Artistic Works' <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4123309> accessed 27 July 2025.

149 Margoni, Ducuing and Shirru (2023), *supra* note 24, p. 9.

150 David Erdos, 'Comparing Constitutional Privacy and Data Protection Rights within the EU' (2021) *University of Cambridge Faculty of Law Research Paper No. 21/2021* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3843653> accessed 27 July 2025; Maximilian von Grafenstein 'Redfining the Concept of the Right to Data Protection in Article 8 ECFR – Part II: Controlling Risk through (Not to) Article 8 ECFR against Other Fundamental Rights' (2021) *European Data Protection Law Review*6(4) p. 516 <https://doi.org/10.21552/edpl/2020/4/7> accessed 27 July 2025; Max van Grafenstein, *The Principle of Purpose Limitation: The Risk-Based Approach, Legal Principles and Private Standards as Elements for Regulating Innovation* (Nomos 2018) <https://www.nomos-elibrary.de/de/10.5771/9783845290843/the-principle-of-purpose-

recitals to the GDPR that underline the fundamental rights as the foundation of the Regulation.[151] Personal data concerns any data 'relating to an identified or identifiable' natural person, also referred to as the data subject.[152] The CJEU case law has clearly established the fundamental rights-driven approach of the GDPR on several occasions. The respect for the protection of personal data is referred to in Article 16(1) of the TFEU and Article 8(1) of the EU Charter of Fundamental Rights (CFR).[153] Article 7 of the EU CFR also refers to the right to respect for personal life. Even though neither the European Convention on Human Rights nor the national constitutions of many EU Member States explicitly refer to a right to data protection, they do refer to the freedom of expression (Article 10) and right to respect for private and family life (Article 8) of the European Convention on Human Rights, and that any derogations thereto, must meet the principle of proportionality.[154] The EU courts have on several occasions underlined the need for personal data processing to be compliant with the foundational principles in Article 8 of the EU CFR, as the use of personal data evokes data protection laws.[155] From the lens of 'human rights law', GDPR 'functions as a justificatory regime' to facilitate 'proportionate' data processing, by offering 'safeguards to ensure' that processing does not transgress 'beyond what is necessary'.[156]

50 From a fundamental rights perspective, a distinction is drawn between the right to privacy and the right to data protection.[157] Following an empirical assessment of the evolution of privacy and data protection laws across EU Member States, Prof. Erdos suggests how the parallel emergence of these rights 'confirms the close and even symbiotic relationship' between data protection and privacy rights.[158] It emerges that data protection, while 'a novel and mercurial phenomenon'[159] has 'roots' in privacy, a thread that becomes clearer while looking at 'EU States which have recognised privacy but *not* data protection as a constitutional fundamental' right.[160] This interplay is also evident as one looks at the ePrivacy rules. For example, the rules on the protection of personal data – such as content, traffic and location data – in the ePrivacy Directive are far more granular than in the GDPR.[161] However, the EDPS and the Article 29 Working Party (WP29) have a different opinion, namely that the GDPR concerns personal data, whereas the ePrivacy rule also 'additionally protect[-s] the confidentiality of electronic communications, as well as the integrity of one's device'.[162] What remains clear though is the distinct, but inter-related nature of data protection and privacy. From the lens of synthetic data, the concerns are two-fold – first, whether the training of GenAI models triggers GDPR, as it involves processing of personal data (section 4.2); and second, if so, whether synthetic datasets can help comply with the GDPR, considering that it may be practically infeasible for GenAI models to unlearn following deep learning from the datasets (sub-section 4.3).

II. Does GenAI Trigger GDPR, as it Crawls and Processes Data for Training Purposes?

51 Big data driven processes, such as GenAI, process personal data and are therefore subject to the GDPR. In the digital economy, when every physical aspect of our activity can be mapped on a digital device, scholars, such as Prof. Purtova, have referred to data protection as the 'law of everything'.[163] Consider,

limitation-in-data-protection-laws> accessed 27 July 2025.

151 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural person with regard to processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation, GDPR), recitals 1, 4, and 51.

152 Article 4(1), GDPR.

153 Charter of Fundamental Rights of the European Union OJ 2010/C83/389.

154 David Erdos, 'European Union Data Protection Law and Media Expression: Fundamentally Off-balance' (January 2016) *International & Comparative Law Quarterly* 65(1) p. 145 <https://doi.org/10.1017/S0020589315000512> accessed 27 July 2025.

155 Joined Cases C-293/12 and 594/12 *Digital Rights Ireland Ltd. and Seitlinger and others* EU:C:2014:238, paras 36, 37.

156 Orla Lynskey (2023) 'Complete and Effective Data Protection' *Current Legal Problems* 76 (1) p. 301 <https://doi.org/10.1093/clp/cuad009> accessed 27 July 2025.

157 Erdos (May 2021), *supra* note 150, pp. 1-2.

158 *See ibid* p. 3, pp. 16-19 23 for the National provisions on the general right to privacy (or broad equivalent) across EU Member States and pp. 22-23 for the National provisions on the right to data protection across EU Member States.

159 *Ibid.*, p. 31.

160 *Ibid.*, p. 27, 31.

161 Rosa Barcelo 'The ePrivacy Directive: then and now' in Brendan Van Alsenoy, Julia Hodder, Fenneke Buskermolen, Miriam Čakurdová, Ilektra Makraki and Estelle Burgot (eds) *Two decades of personal data protection. What next? EDPS 20th Anniversary* (European Data Protection Supervisor: Publication Office of the European Union 2024) pp. 49-50 <https://www.edps.europa.eu/data-protection/our-work/publications/book/2024-06-20-two-decades-personal-data-protection-whats-next_en> accessed 27 July 2025.

162 For a discussion on the EDPS and Article 29 Working Party and the European Commission's opinion on the interplay between right to privacy and data protection *see ibid*, pp. 50-51.

163 Nadezhda Purtova 'The Law of Everything. Broad concept

for example, a daily activity, such as shopping for groceries in the supermarket. If one may not be familiar with the local area, the shopper will first look on the internet for nearby markets, followed by the use of a navigation app to go to the nearby supermarket. The physical footprint of the user thus will also be present in the form of a digital map.

52 As GenAI uses publicly available data, from the lens of GDPR, the key concern is whether they have a valid legal basis to undertake such processing activities.[164] The Italian Data Protection Authority (DPA) was amongst the first authorities to initiate an action against ChatGPT for non-compliance with the principles of the GDPR. In 2023, ChatGPT was seen as non-compliant with the requirement for a valid legal basis, required for processing personal data, by the Italian DPA. To be GDPR compliant, the processor must have a valid legal basis to comply with the 'lawfulness principles' prescribed in Article 5(1)(a) GDPR. The Italian DPA initially banned ChatGPT, requiring amongst others 'to change the legal basis of the processing of users' personal data'.[165] To comply with the Italian DPA, OpenAI changed its privacy policy across the EU and the EFTA, and expressed its legitimate interest in 'developing, improving, or promoting' its services, including the training of its models.[166] This change of legal basis only addresses the concern with the users of ChatGPT. The question is much larger – on what legal basis do the GenAI models, and ChatGPT in particular, process the personal data of the internet users at large? The EDPS addressed this aspect in its recent opinion, and its suggestion seems aligned with the Italian DPA's decision. The EDPS opined that GenAI model providers may rely on 'legitimate interest... [especially] with regard to the collection of data', as well as for 'training and validation purposes'.[167]

53 GenAI tools process data to offer meaningful outputs. In an earlier article, I describe how GenAI models contextualize, iterate, and improvise information to generate new works. In addition to the data derived from IP (copyright) protected works, this processing may also involve personal data. Considering that Article 4(2), GDPR offers a very broad meaning to processing, even when GenAI does not copy, store or retain data (whether obtained by web crawling or through other sources) in its database, in the initial training phase at least, GenAI tools do tend to process personal data within the meaning of GDPR. Such an interpretation is aligned with landmark decisions such as *Google Spain*, as also the more recent opinion by the EDPS. In *Google Spain*, the data subject appeared in Google's search results, which offered links to webpages providing data on how the data subject had formerly been involved in a bankruptcy auction. As this step was identified as processing, the CJEU was of the opinion that Google, as the data controller (and also the processor in the case at hand), must ensure compliance with the data subject's rights, including but not limited to the right to removal of the bankruptcy related data from the search results.[168] In the said case, compliance with the data subject's request for erasure was much simpler. However, how does one facilitate compliance with such personal rights under the GDPR in the context of GenAI models that are based on deep learning? To comply with the request, the model may be required to unlearn, a cost that may be disproportionate, and may practically speaking even impossible, particularly in light of the black box nature of these GenAI models.

54 Another important and related question is the duration for which the personal data collected may be stored. In *Schrems*, the Court opined that an unlimited duration may seem intrusive as it may offer an impression of continuous monitoring of the personal life of the data subject.[169]

55 What happens when GenAI systems return incorrect

of personal data and future of EU data protection law' (2018) *Law, Information and Technology* 10 (1) <https://www.tandfonline.com/doi/full/10.1080/17579961.2018.1452176> accessed 27 July 2025.

164 Taner Kuru (2024) 'Lawfulness of the mass producing of publicly accessible online data to train large language models' *International Data Privacy Law* 14(4) p. 330 <https://academic.oup.com/idpl/article/14/4/326/7816718> accessed 27 July 2025.

165 Garante per la protezione dei dati personali (GDDP), Provvedimento dell'11 aprile 2023 <https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9874702> accessed 27 July 2025.

166 Kuru (2024), *supra* note 164, p. 332. See also the references therein.

167 European Data Protection Supervisor, 'Generative AI and the EUGDPR: First EDPS Orientations for ensuring data protection compliance when using Generative AI systems' (3 June 2024) p. 13 <https://www.edps.europa.eu/system/files/2024-05/24-05-29_genai_orientations_en_0.pdf> accessed 27 July 2025.

168 David Erdos, 'Generative AI, Search Engines and GDPR' (15 January 2024) slide 2 <https://www.slideshare.net/slideshow/generative-ai-search-engines-and-gdpr/265438456#1> accessed 27 July 2025. Prof. Erdos maps the search indexing and European Data Protection (EDP) timeline and identifies the following four phases. In phase 1, between mid-1980s and mid-1990s, the EDP identified certain concerns and mildly regulated the news archive searches; phase 2, lasting between late-1990s and 2000s, whereby search engines were seen as 'out of reach' and focus remained on limiting exposure; third phase starting 2007-08, whereby the Spanish DPA identified search engines as ex-post controllers and the fourth phase starting 2014-present, with CJEU's Google Spain as a precedent.

169 C-446/21 *Maximillian Schrems v. Meta Platforms Ireland Ltd.*, ECLI:EU:C:2024:834, paras 58, 60, 62.

results when presented with questions about the data subjects? GenAI output is synthetic. In addition, this synthetic output often gives incorrect information. When this information, correct or incorrect, identifies or can identify a person, it involves personal data and is covered by the GDPR. Can factual incorrectness of information cause prejudice, and can it be covered by the GDPR? As referred to in section 3 *supra,* GenAI models tend to hallucinate which may lead to offering false information, including personal data by these models. In case of inaccurate or false results, GenAI models may be found in breach of the principle of data accuracy as required under Article 4(1)(d), 2016 EU GDPR. To avert such an inaccurate outcome, data accuracy must be assured 'throughout the whole lifecycle of the generative AI systems'.[170] Simply put, accuracy should not be ascertained at the output stage, rather, in light of the black box nature of these deep learning models, the GenAI model developer should be able to ensure accuracy from the input to the output stage of the model.

56 The scope of the term 'personal data' gains significance, as GenAI models, a sub-field of AI, deal with 'multi-layered AI models where each layer performs a specific task of input data analysis or manipulation' and the process goes on in a loop for the GenAI model to improvise itself with each successive iteration.[171] As data goes from one layer to the next in this multi-layered processing, in practice, 'it may be burdensome, or even impossible to trace exactly what behaviour was learned based on' the data subject's personal data.[172] This leads to practical challenges as regards the exercise of individual rights of the data subjects, such as the right of access, rectification, erasure and objection to the processing of personal data prescribed in Chapter III of the EU GDPR.[173] Consider for example, the right to erasure under Article 17 of the GDPR, particularly when the data subject may have initially offered their consent for processing, but withdrew it subsequently. In such a scenario, it may be difficult to exercise the right of erasure, 'as the actual using of that data persists throughout the operation' of the GenAI model.[174] Considering that the exercise of these individual rights may disproportionately 'impact the effectiveness of the model', the CJEU's recent opinion in *Meta v. Bundeskartellamt*, discussed below, may offer a useful benchmark to proportionately balance the distinct fundamental rights (the right of the data controller vis-à-vis that of the data subject). In addition, the 2024 EU AI Act's requirement for detailed information on datasets used to train the GenAI model, may enhance transparency and help decode the GenAI black box.

57 In *Meta v. Bundeskartellamt*, Meta, the processor enjoyed a dominant position, and the question was whether the data subject could, in such a setting, offer free consent under Articles 6(1)(a) and 9(2)(a) GDPR for the processing of personal data.[175] The CJEU was of the opinion that in order to facilitate positive compliance with the requirements therein, three cumulative conditions must be met – first, the pursuit of a legitimate interest; second, the processing of personal data for this legitimate interest; and third, the interests or fundamental freedoms of the data subject do not overshadow the legitimate interests of the controller or a third party.[176]

58 Would the case of processing of sensitive personal data, manifestly made available by the data subject, for training GenAI models be different? To understand the scope and interpretation of Article 9(2)(e) GDPR, the recently decided Schrems case is insightful, whereby the CJEU was of the opinion that

> *'the fact that a person has made a statement about his or her sexual orientation on the occasion of a panel discussion open to the public does not authorize the operator of an online social network platform to process other data relating to that person's sexual orientation, obtained, as the case may be, outside that platform using partner third-party websites and apps, with a view to aggregating and analysing those data, in order to offer that person personalized advertising'.*[177]

59 Processing of personal data that can be deemed 'sensitive' is prohibited by Article 9(1) GDPR, unless the processor can benefit from the exceptions under Article 9(2) GDPR. This could be a better ground to make GenAI firms accountable, as the grounds therein are stricter, and the CJEU has offered 'a broad interpretation of what constitutes sensitive data'.[178]

170 European Data Protection Supervisor (3 June 2024), *supra* note 167, p. 15.

171 Aleksandr Kesa and Tanel Kerikmäe, 'Artificial Intelligence and the GDPR: Inevitable Nemeses?' (2020) *TalTech Journal of European Studies* 10(3) p. 70 <https://sciendo.com/article/10.1515/bjes-2020-0022> accessed 27 July 2025.

172 *Ibid.*

173 European Data Protection Supervisor (3 June 2024), *supra* note 167, p. 22.

174 Kesa and Kerikmäe (2020), *supra* note 171, p. 75.

175 For a discussion of the case from a competition law perspective, see Anne C. Witt (2024) 'Meta v Bundeskartellamt – data-based conduct between antitrust law and regulation' *Journal of Antitrust Enforcement* 12(2) <https://academic.oup.com/antitrust/article/12/2/345/7642048> accessed 27 July 2025.

176 Case C-252/21 *Meta v. Bundeskartellamt*, ECLI:EU:2023:537, para 106; C-597/19 *Microm International Content Management & Consulting (M.I.C.M.) Limited v. Telenet BVBA and others*, ECLI:EU:C:2021:492, para 106.

177 C-446/21 *Maximillian Schrems v. Meta Platforms Ireland Ltd.*, ECLI:EU:C:2024:834, paras 83-84

178 Kuru (2024), *supra* note 164, p. 335.

60 Crawling by GenAI FM developers should be identified as GDPR compliant only after they meaningfully filter sensitive personal data. Making a 'sensitive personal' announcement in a public forum, such as was the case in *Schrems*, does not necessarily mean that the data subject has made the personal data manifestly available under Article 9(2)(e) GDPR.[179]

III. Synthetic Data to Facilitate Compliance with the GDPR

61 The EU GDPR requires that data is processed in a 'lawful, fair and transparent' manner[180] and that the users have given clear consent for the collection and processing of personal data[181]. There are, however, clear gaps between the ideals of data protection and the practice of digital firms. An empirical survey identified how over 92 per cent of the most popular websites tracked users without notification and even when the users clearly opted-out, over 85 per cent of the websites continued to track its users.[182] These limits imposed by the GDPR are breached with even greater impunity by the GenAI model developers as they indiscriminately crawl the web for data. In this regard, can synthetic data, provided it complies with the safeguards of the A29 WP on anonymization techniques – such as singling out, linkability and inferences – facilitate compliance with data protection laws?

62 It may be trite to clarify that privacy is not the same as data protection. However, the fact that synthetic data successfully anonymizes data[183] ensures compliance with the GDPR. GDPR is driven by openness and control over one's data, as distinct from privacy that may require secrecy.[184]

63 With synthetic data as input, datasets may be fully synthetic (1); a part of it may be synthetic and a part human-generated (2) or it may be a combination of human generated, and synthetic data (3).[185] For synthetic data to be exempt from the provisions of Articles 6(1)(b) to 6(1)(f) of the GDPR, it must be sufficiently anonymized. Anonymization is not a shield from the requirement for compliance with the data protection laws. Anonymization is not the same as data protection. Complete anonymization may help evade the requirements under the GDPR, as successful anonymization means that data cannot be linked to a pre-identified individual.[186] In its most recent opinion, the EDPS opines thus[187]:

> *When a developer or a provider of a generative AI system claims that their system does not process personal data (for reasons such as the alleged use of anonymised datasets or synthetic data during its design, development and testing), it is crucial to ask about the specific controls that have been put in place to guarantee this.*

64 The data must be 'evaluated as anonymous, through a process of proven quality [whereby there is] reasonable evidence of impossibility of reidentification'.[188] When a given dataset is fully synthetic, in other words, it consist of fully anonymized data, such a dataset may qualify as pseudonymous or anonymized data under Article 4(5) of the GDPR, which defines pseudonymisation as follows:

> *'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.*

65 Fully synthetic datasets (case 1) can help comply with data privacy and facilitate innovation through more collaborative data sharing.[189] It can help overcome constraints such as trade secrets and privacy regulations while doing an inter or intra-firm data

179 *Ibid.*, p. 345.

180 Article 5, GDPR.

181 Articles 6-7, GDPR.

182 See reference to the study on Europe's 2000 most visited websites by Iskander Sanchez-Rola et al 'Can I Opt Out Yet? GDPR and the Global Illusion of Cookie Control' (2019) Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security as discussed in Filippo Lancieri, 'Narrowing Data Protection's Enforcement Gap'(January 2022) *Maine Law Review* 74(1) pp. 17-18, 65 <https://digitalcommons.mainelaw.maine.edu/mlr/vol74/iss1/3/> accessed 27 July 2025.

183 On the techniques of synthetic data generation, and how it successfully anonymizes data, see *supra,* Section 2.

184 La Diega and Sappa (2020), *supra* note 73, pp. 7-8.

185 Nicolas Ruiz, Krishnamurthy Muralidhar and Josep Domingo-Ferrer 'On the privacy guarantee of synthetic data: a reassessment of the maximum-knowledge attacker perspective' in Josep Domingo-Ferrer and Francisco Montes (eds) *Privacy in Statistical Databases* (Cham: Springer Nature 2018) pp. 59-74 <https://link.springer.com/chapter/10.1007/978-3-319-99771-1_5> accessed 27 July 2025.

186 Gal and Lynskey (2024), *supra* note 25.

187 European Data Protection Supervisor (3 June 2024), *supra* note 167, p. 7.

188 Agencia Española Protección Datos, 'Anonymization III: The risk of re-identification' (Online 23 February 2023) *AEPD Innovation and Technology Division* <https://www.aepd.es/en/prensa-y-comunicacion/blog/anonymization-iii-risk-re-identification> accessed 27 July 2025.

189 Lee (2024), *supra* note 38, p. 22.

transfer.[190] Synthetic data can help comply with privacy, but does this also mean compliance with the principles of data protection?[191] The criteria in recital 26 of the GDPR is the appropriate benchmark to ascertain 'whether the final identification [of the data subject] is sufficiently probable to assume a *specific* risk to fundamental rights' and whether the processing must comply with the safeguards under the GDPR.[192]

66 Possibility of re-identification, such as in case of the dataset being partly synthetic, and partly human-generated or alternatively, a combination of synthetic and human-generated data can evoke the applicability of the EU GDPR. Moreover, as many GPAI model providers are based outside the EU, it is important that data transfers must have sufficient safeguards for the protection of personal data, in compliance with the EU GDPR.[193] Even when the datasets can distantly identify a data subject, they may involve processing of personal data, as per Article 4(1), GDPR.[194] In such a scenario, compliance with the following requirements under Article 5 GDPR, namely 'lawfulness, fairness, transparency' (Article 5(1)(a)); 'purpose limitation' (Article 5(1)(b)); 'data minimization'((Article 5(1)(c)), 'storage limitation'(Article 5(1)(e)); 'accountability' (Article 5(2)); accuracy (Article 5(1)(d)), integrity and confidentiality ((Article 5(1)(f))), must be met.

E. Conclusion, Policy Recommendation and Further Research

67 Generative AI models are a key source for synthetic data generation. Synthetic data is in turn used as an input for further training these GenAI models, and thereby 'leverage the diversity and scale of artificial datasets' to make these (generative) models more robust.[195] In an earlier research output, I develop the 'contextualise, iterate and improvise' (CII) model to explain how generative AI models such as ChatGPT 'think and improvise with every successive iteration'.[196] In this paper, I go a step further to add the element of synthetic data. This holistic model (as visually represented in Figure 1) helps understand the situation whereby synthetic data strengthens the capability of Generative AI models, and scenarios wherein human generated data inputs may continue to be a requirement to prevent the 'collapse of [these] models'[197]. This aspect is vital to appreciate why notwithstanding the emergence of synthetic data, high quality human-generated data will continue to be in demand and coexist alongside synthetically-generated data. The additional layer, represented by the blue arrows refers to the additional layer of synthetic data. Synthetic data complements and adds to the quality, variety, and volume of human-generated data available to train the GenAI models.

190 See reference to Steven M. Bellovin, Preetam K. Dutta and Nathan Reitinger, 'Privacy and Synthetic Datasets' (2019) *Stanford Technology Law Review* as discussed in Gal and Lynskey (2024), *supra* note 25, p. 1109.

191 Juliana Kokott and Christoph Sobotta, 'The distinction between privacy and data protection in the jurisprudence of the CJEU and the ECtHR' (2013) *International Data Protection Law* 3(4) <https://academic.oup.com/idpl/article/3/4/222/727206> accessed 27 July 2025.

192 Valentin Rupp and Max von Grafenstein, 'Clarifying "personal data" and the role of anonymization in data protection law: Including and excluding data from the scope of the GDPR (more clearly) through refining the concept of data protection' (2024) *Computer Law & Security Review: The International Journal of Technology Law and Practice Vol. 52,* p.18 <https://www.sciencedirect.com/science/article/pii/S0267364923001425> accessed 27 July 2025.

193 The Court did not hesitate to invalidate the then EU-US Safe Harbour and Privacy Shield on the grounds the required legal obligations for transatlantic data transfers, and protection of the European citizens' data was not met. Case C-362/14 *Maximilian Schrems v. Data Protection Commissioner* ('Schrems I', Safe Harbour case) ECLI:EU:2015:650, paras 88-90 and Case C-311/18 *Data Protection Commissioner v. Facebook Ireland Ltd. & Others* ('Schrems II'), ECLI:EU:C:2020:559, paras 184, 185, 191.

194 Beduschi (2024), *supra* note 20, pp. 1, 2.

195 Cem Dilegani 'Synthetic Data Generation: Techniques & Best Practices' (1 October 2024) AI *Multiple Research* <https://research.aimultiple.com/synthetic-data-generation/> accessed 27 July 2025.

196 Tyagi (2024), *supra* note 87.

197 Aatish Bhatia 'When A.I.'s Output Is a Threat to A.I. Itself' (Online 25 August 2024) *New York Times* <https://www.nytimes.com/interactive/2024/08/26/upshot/ai-synthetic-data.html?unlocked_article_code=1.Uk4.lUcJ.NURrC0S1B4oq&smid=em-share> accessed 27 July 2025.

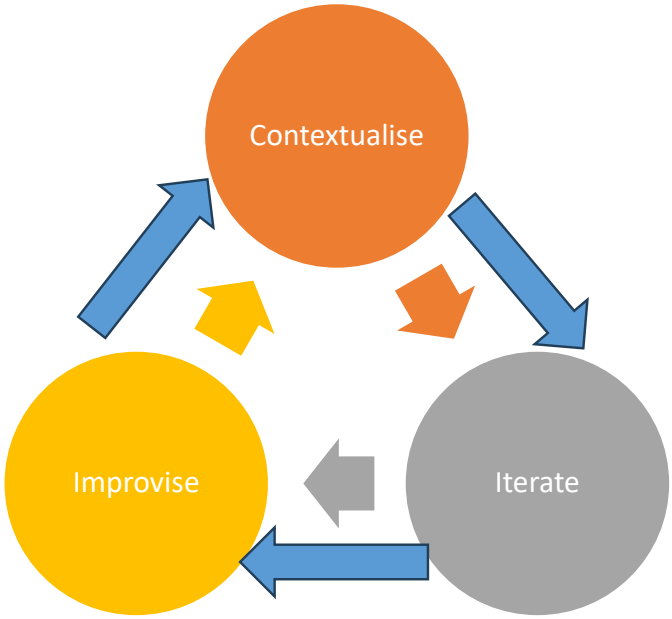

Figure 1: Co-creation by man and machine: Co-existence of human-generated and synthetic data. The outer blue colored arrows indicate a mix of human-generated and synthetically generated data. The inner multi-coloured arrows indicate human-generated data.

68 Synthetic data generation, as section 2 *supra* elaborates, is a complex technical task with substantial sunk costs. The quality of the output is dependent on the complexity of the model used to generate synthetic data.[198] Even though commercial and open-source models for synthetic data generation are currently available in the market, the quality of proprietary commercial models is far superior to the small scale or open-source models.[199] This is not difficult to comprehend. Synthetic data generation requires technical capabilities. There also exist synergies between GenAI and other key technologies, such as the internet of things (IoT), whereby synthetic data is the input as well as the output. Amazon's training of Alexa is a case in point. Another related consideration is that whereas first movers, and early winners in the digital economy and the GenAI race, such as OpenAI and Google, scraped the web to train their GenAI models, they 'have [since] updated their terms of service to prohibit [others from using] their data to train AI models'.[200] This implies that the availability of synthetic data is determined by the degree of competition and contestability in the digital markets. Innovation in high quality synthetic data generation requires substantial sunk costs for research and development, and like the platform economy, and the market for Generative AI, exhibits sectors-specific features such as barriers to entry, and economies of scale and scope. Herein, the scope of the EU Data Laws, EU competition law and the Digital Markets Act becomes relevant to facilitate the continued availability of high quality data.[201] This area of law is driven by economic rationale, and the driving principle is to facilitate access and contestability, a focus area of the follow-on research article. Copyright and data protection, on the other hand are respectively grounded in innovation and respect for fundamental rights, an issue addressed at length in this contribution.

69 Moreover, even when some of the algorithms, such as those provided by Google, IBM and Microsoft, are open source; the datasets used to train these algorithms is a black box.[202] The 2024 EU AI Act calls for transparency and disclosure of the training data. The Commission's Template offers a 'common minimal baseline' for GPAI model providers to publicly share information about the list of data sources, including synthetic data used for training the model.[203]

70 Synthetic data can notably help overcome the limitation of quality datasets – a key input and barrier to entry in the digital markets – and can thereby contribute to the innovation dimension of the economy. As regards copyright, the follow-on works may be deemed infringing if the initially synthetically generated data is derived from copyright-protected works. Synthetic datasets only when fully anonymised can facilitate compliance with data protection rules.[204] Even when a part of dataset may remotely identify personal data of the data subject, the processing of such a dataset must meet the requirements under Article 5, GDPR.

71 While it may be true, provided that the analysis is correct, that over 60 per cent of data currently available on the internet is synthetic, 'recursive training of the GenAI models [with synthetic data] can lead to model collapse'[205]. Training

198 *Supra* note 24, p. 60.

199 *Ibid.*, pp. 59-60, 62.

200 Alistair Barr 'AI Hypocrisy: OpenAI, Google and Anthropic Won't Let Their Data Be Used to Train Other AI Models, But They Use Everyone Else's Content' (Online 2 June 2023) *Business Insider* as referred in Lee (2024), *supra* note 38, p. 8.

201 *See supra* note 12.

202 Amanda Levendowski 'How Copyright Law Can Fix Artifical Intelligence's Implicit Bias Problem' (2018) Washington Law Review 93(2) p. 583 <https://digitalcommons.law.uw.edu/cgi/viewcontent.cgi?article=5042&context=wlr> accessed 27 July 2025.

203 European Commission (24 July 2025), *supra* note 47.

204 *See* Gal and Lynskey (2024), *supra* note 25.

205 *See* Illia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot and Ross Anderson, 'The Curse of

GenAI models repetitively on synthetic data can overtime deteriorate the quality and 'negatively affect downstream performance' of the model.[206] To prevent this, there will always be a demand for high quality human-generated works. For a sustainable digital future and growth of GenAI applications, the models will require a mix of synthetic and human-generated data.[207]

72 Simultaneous availability of human-generated and synthetic data can be useful to address the copyright and data protection-related concerns, and the need to balance distinct fundamental rights – such as the rights to author remuneration (safeguarded under Article 17(2), CFR), the right to privacy (Article 7, CFR) and the right to data protection (Article 8, CFR)). Thus, an innovation-driven synthetic data paradigm can also be an enabler of different rights and competing interests at stake. From a technical lens, constraints, such as model collapse, may help avert complete substitution of human-generated data by synthetic data, and a sound fundamental rights-driven legal framework may ensure a balanced sharing of profits among the co-creators of data (human generated as well as synthetic) in the generative AI value chain. To meaningfully facilitate this, effective enforcement should not succumb to corporate lobbying or the US calls for relaxing compliance requirements by the big tech.[208]

Recursion: Training on Generated Data Makes Models Forget' (31 May 2023) <https://www.cl.cam.ac.uk/~is410/Papers/dementia_arxiv.pdfin> accessed 27 July 2025, also discussed in Lee (2024), *supra* note 38, p. 26.

206 Ryuichiro Hataya, Han Bao and Hiromi Arai 'Will Large-scale Generative Models Corrupt Future Datasets?' (15 November 2021) *Cornell University Computer Science: Artificial Intelligence* <https://arxiv.org/abs/2211.08095> accessed 27 July 2025.

207 Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez and Rik Sarkar, 'Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet' (8 June 2023) *Cornell University Computer Science: Artificial Intelligence* <https://arxiv.org/abs/2306.06130> accessed 27 July 2025.

208 Mathieu Pollet and Pieter Haeck 'EU could postpone flagship AI rules, tech chief says' (Online 6 June 2025) *Politico* <https://www.politico.eu/article/eu-could-postpone-parts-of-ai-rulebook-tech-chief-says/> accessed 27 July 2025.