Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn from the Japanese 'Non-Enjoyment' Purposes?

By Artha Dermawan[1]

ABSTRACT

The European Union (EU) text and data mining (TDM) provisions are a progressive move, but the horizon is still uncertain for both generative artificial intelligence (GenAI) models researchers and developers. This article suggests that to drive innovation and further the commitment to the digital single market, during the national implementation, EU Member States could consider taking the Japanese broad, all-encompassing and 'non-enjoyment-based' TDM as an example. The Japanese 'non-enjoyment' purposes, however, are not foreign to the European continental view of copyright. A similar concept can be found under the German concept of *"Freier Werkgenuss"* or enjoyment of the work. A flexible TDM exception built upon the German notion of non-enjoyment purposes could become an opening clause to foster innovation and creativity in the age of GenAI. Moreover, the article argues that an opening clause allowing TDM with 'non-enjoyment' purposes could be permissible under the so-called three-step test.

This article further suggests, if there is no political will to safeguard "the right to read should be the right to mine" and to provide a welcoming environment for GenAI researchers and developers, when shaping the legal interpretation through national case law, the EU Member States could consider the following: (1) advocate for 72 hours of response if technological protection measures (TPMs) are preventing TDM, and (2) Robot Exclusion Standard (robot.txt) as a warning when TDM is not allowed on a website.

It is now in the hands of the EU Member States, whether to protect the interests of rightholders or to create a balance between safeguarding 'the right to read should be the right to mine', protecting rightholders exclusivity, and creating a supportive environment for the GenAI models researcher and developers.

Keywords: Text and Data Mining, Copyright and Related Rights, Exceptions and Limitations, Generative AI Models, 'Non-Enjoyment' Purposes, Freier Werkgenuss, Three-Step Test, Innovation.

[1] Indonesia-qualified lawyer and a doctoral (LL.D) student supported by the Max Planck Institute for Innovation and Competition (Munich, Germany) and a member of the Law, Technology and Design Thinking (LTDT) Research Group at the Faculty of Law, University of Lapland (Rovaniemi, Finland). E-mail: artha.dermawan@ip.mpg.de and ORCID: 0000-0003-4357-4656.

I. Introduction

'L'homme est un néant à l'égard de l'infini,
un tout à l'égard du néant entre rien et tout.'

- B PASCAL IN "LES DEUX INFINIS."[2]

Pascal through his poetry illuminates the disparity that people foster while comparing themselves to the world, which is transforming quickly all around them. Humans have no choice but to seek refuge in their beliefs. Although those words date from a time when technology was not even a phantasmagoria, they might nevertheless capture the mindset of a copyright owner who is both captivated and horrified by the current progress of artificial intelligence ("AI"). It is difficult to avoid an adaptation, but at the same time, lawmakers find it challenging to create a normative vision that can capture and keep up with this technological advancement.

Nowadays, AI systems are capable of producing human-level creative output, such as poetry, stories, jokes, music, paintings, etc., as well as, the growing automation of tasks typically performed by human artists. In this article, these AI systems are referred to as 'generative AI (GenAI)' models.[3] These GenAI systems have been fuelled in particular by new data-driven technologies.[4] The development of GenAI models or AI in general, cannot be separated from data (data in this article refers to non-personal data which includes any literary and artistic works such as text, music, pictures etc.).[5] The value produced by data is a key factor in

---

[2] Unofficial translation in English: "Man is a nothingness in relation to the infinite, a whole with respect to the nothingness between nothing and everything." GP Clermont-Ferrand, 'Pensées de Blaise Pascal' (Pensées de Blaise Pascal) <http://www.penseesdepascal.fr/Transition/Transition4-moderne.php> accessed January 13, 2023.

[3] OpenAI, 'Generative Models.' Available at: <https://openai.com/research/generative-models> accessed January 13, 2023. See, McKinsey & Company, 'What is Generative AI?' (2023) Available at: <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai> accessed January 13, 2023. See also, D Brady, 'What Developers Need to Know About Generative AI.' (Github, 2023) Available at: <https://github.blog/2023-04-07-what-developers-need-to-know-about-generative-ai/> accessed January 13, 2023.

[4] A Elgammal, L Bingchen, M Elhoseiny and M Mazzone. 'CAN: Creative Adversarial Networks, Generating "Art" by Learning about Styles and Deviating from Style Norms.' at 1. Every day, quintillions of bytes of data is being generated, and it is estimated that by 2023, the world will be populated by 29 billion smart connected devices capable of collecting and sharing data in real time and making autonomous decisions. Y Ménière, 'Patents and the Fourth Industrial Revolution: The Global Technology Trends Enabling the Data Economy', (2022), Progress in IS, in: A Bounfour (ed), *Platforms and Artificial Intelligence*, at 103-109. *See also,* European Union Agency for Fundamental Rights, 2020. Getting the Future Right – Artificial Intelligence and Fundamental Rights. Available at: https://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights#publication-tab-0. accessed December 10, 2022.

[5] In general, data is divided into two parts, namely personal data and non-personal data. Non-personal data is data that does not refer to an identified or identifiable person. T Pihlajarine and R Ballardini, 'Owning Data via Intellectual Property Rights: Reality or Chimera?' in R Ballardini, P Kuoppamäki and O Pitkänen (eds), *Regulating Industrial Internet Through IPR, Data Protection and Competition Law* (Wolters Kluwer, 2019) at 116. *See also*, the importance of non-personal data to the economy, J Drexl, 'The Future EU Legal Framework for the Digital Economy: A Competition-Based Response to the 'Ownership and Access' Debate' in S Lohsse, R Schulze and D Staudenmayer (eds), *Trading Data in the Digital Economy: Legal Concepts and Tools* (2017).

determining the present and future of GenAI.[6] The value of data as such generally lies in the extraction of value rather than in the data or text considered independently.[7] Enabling the discovery of new patterns and relations of creative outputs requires GenAI to conduct an analysis of the substantial amounts of data. The analysis of the data, which is practically impossible to accomplish manually, is efficiently done using an automated computational analysis known as 'Text and Data Mining' ("TDM").[8]

TDM (*stricto sensu*) can be described as "the selection and application of complex algorithms to the transformed alphanumerical dataset to gather hidden information."[9] From the copyright and related rights microscope, TDM plays an important role in analysing large amounts of information in digital form including images, text and sound contained in "a large amount of diversified time series data generated at a high speed by industrial equipment"[10] or well-known as 'Big Data'. The purpose is to gain new knowledge and uncover new patterns, for the development of GenAI.[11] In essence, the process of creating outputs with GenAI models involved TDM through (i) access to content, (ii) extraction and/or copying of content, and (iii) mining of text and/or data and knowledge discovery, TDM creates rich and diverse data sets that are then utilized to train and feed AI for creative purposes.[12] The data used in the 'extraction and/or copying of content' stage may require authorisation from the relevant rightholders. To create a balance between rightholders exclusivity and TDM, the EU passed Directive 2019/790 ("EU CDSM Directive"),[13] which includes two necessary TDM exceptions. This was done to

---

[6] "AI will exploit the digital data from people and things to automate and assist in what we do today, as well as find new ways of doing things that we've not imagined before." A Popescu, 'EconPapers: The Value of Data From an Artificial Intelligence Perspective', (Econpapers: The Value of Data From An Artificial Intelligence Perspective, 2019). Available at: <https://econpapers.repec.org/article/edtaucjcm/v_3a5_3ay_3a2019_3ai_3a1_3ap172-194.htm> accessed December 10, 2022 at 176.

[7] E Rosati, 'An EU text and data mining exception for the few: would it make sense?' (Journal of Intellectual Property Law & Practice, Vol 13, Issue 6, 2019) at 429.

[8] E Rosati, 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market: Technical Aspects', Briefing requested by the JURI committee, Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament, at 2.

[9] GS Muto, 'A New Text and Data Mining Exception: Waiting for Godot?' (Munich Intellectual Property Law Center, 2017) at 11, citing M Mariscal and Others, 'A Survey of Data Mining and Knowledge Discovery process Models and Methodologies' (Vol 25(2), The Knowledge Engineering Review, 2010) at 137-166.

[10] The concept of 'Big Data' first appeared in 2012, along with the notion of "Industry 4.0," and refers to data generated by industrial equipment that may have significant economic value. T Pihlajarine and R Ballardini, *supra* note 6, at 117.

[11] K Christensen, 'A European Solution for Text and Data Mining in the Development of Creative Artificial Intelligence: With a Specific Focus on Articles 3 and 4 of the Digital Single Market Directive' (A European solution for Text and Data Mining in the development of creative Artificial Intelligence : With a specific focus on articles 3 and 4 of the Digital Single Market Directive, 2021). Available at: <http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1584979&dswid=-462> accessed July 7, 2022, at 19. *See also*, E Rosati, 'Copyright as an obstacle or an enabler? A European perspective on text and data mining and its role in the development of AI creativity', (Asia Pacific Law Review, Vol 27, Issue 2, 2019) at 198-199.

[12] E Rosati, *supra* note 8 at 4. *See also*, K Christensen, *supra* note 11 at 1.

[13] Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, OJ L 130, 17.5.2019.

eliminate legal uncertainties and to compete with legal systems that offer a more conducive environment for TDM, for example, Japan which provides the broadest TDM exception in the world. However, the question remains, will these TDM exceptions be able to encourage innovation? Unfortunately, academics and legal experts have the opposite opinion.

This article aims to answer the following question: How could the legal framework in the EU best accommodate research and innovation in the development of GenAI models made possible by TDM? Should the EU Member States, during the national implementation of the CDSM Directive or when shaping the legal interpretation into national case law, take the Japanese TDM exception as an example? This article will try to answer the following questions by assessing the EU and Japanese TDM exceptions and related law cases, and analysing whether the Japanese TDM exceptions suit the European Continental copyright system and are compatible with the so-called 'three-step test'. The article is structured as follows: Section 2 examines the importance of TDM in the development of GenAI models and copyright issues that might arise. Section 3 analyses the newly introduced TDM exception in the EU. Section 4 presents the Japanese TDM exceptions and the rationale behind the 'non-enjoyment' purposes. Section 5 discusses the possible implications of the Japanese 'non-enjoyment' purposes doctrine to the EU Member States and its similarity to the German doctrine *'Freier Werkgenuss'*, the three-step test and several recommendations to the EU Member States who do not wish to implement a broader TDM exception.

II. TDM, Copyright and the Development of GenAI Models

1. Definition of TDM

The definition of TDM must be made crystal clear if the rights, exceptions, and current legal discourse concerning TDM and AI are to be addressed. TDM generally refers to the process of obtaining valuable information from massive amounts of data. It is generally acknowledged that TDM plays an important role in the knowledge discovery process.[14] The EU CDSM Directive describes TDM as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations."[15] According to the Japanese Copyright Act, TDM is a "data analysis (meaning the extraction, comparison, classification, or other statistical analysis of language, sound, or image data, or another element of which a large number of works or a large number of data is composed."[16] TDM is a technique for processing large amounts of text or data that are beyond the capacity of human minds and is recognized as such by both the EU

[14] Y Li and T Beaubouef. 'Data Mining: Concepts, Background and Methods of Integrating Uncertainty in Data Mining.' (2010); For more technical details of TDM, *see also,* Fayyad, Usama, G Piatetsky-Shapiro, P Smyth, 'From Data Mining to Knowledge Discovery in Databases', (1996); *See,* for further reference regarding the mining concepts, J Han, M Kamber, 'Data Mining: Concepts and Techniques', London: Academic Press, 5, 2001; M Kantardzic, 'Data Mining: Concepts, Models, Methods, and Algorithms', (New York: John Wiley & Sons Inc publishes, 2003).

[15] Article 2(2) of the EU CDSM Directive.

[16] Article 30-4(ii) of the Japanese Copyright Act or 著作権法. *See,* Copyright Research and Information Center (CRIC), Copyright Law of Japan. 2020.

CDSM and the Japanese Copyright Act. This allows for the discovery of new, useful information among enormous amounts of potentially irrelevant information.[17]

2. The Procedure of TDM

Large amounts of text and data can be processed, extracted, and recombined using the TDM technique to disclose new insights into the existing information or even produce new knowledge.[18] As stipulated in Illustration 1 below, the AI systems must have access to the content to accomplish this, and they might even need to copy or extract the content. This section attempts to describe the TDM process in a simple manner and to learn more about the legal issues involved. In general, TDM activities can be carried out in various ways and for a myriad of purposes and often fall into one of the following categories:

2.1 Step 1: Access to Content

The first and most important phase in TDM activities is content accessibility.[19] Access to content might be in the form of text or data, depending on the type of mining that will be done. As shown in Illustration 1 below, in general, raw data, target data, and pre-processed data are all related to one another and are all indispensable for this first step of TDM.[20]

2.2 Step 2: Extraction and/or Copying of Content

In this stage, as shown in Illustration 1 below, to transform raw data, target data and/or pre-processed data into patterns, one requires to do the extraction and/or copying of content during the TDM process.

2.3 Step 3: Mining of Text and/or Data and Knowledge Discovery

---

[17] MA Hearst, 'Untangling Text and Data Mining' (1999) Proceedings of the 37th Annual meeting of the Association for Computational Linguistics 3, at 3. *See also*, P Kollár, 'Mind if I Mine? A Study on the Justification and Sufficiency of Text and Data Mining Exceptions in the European Union,' (2021). Available at: <https://ssrn.com/abstract=3960570.> accessed December 10, 2022, at 3.

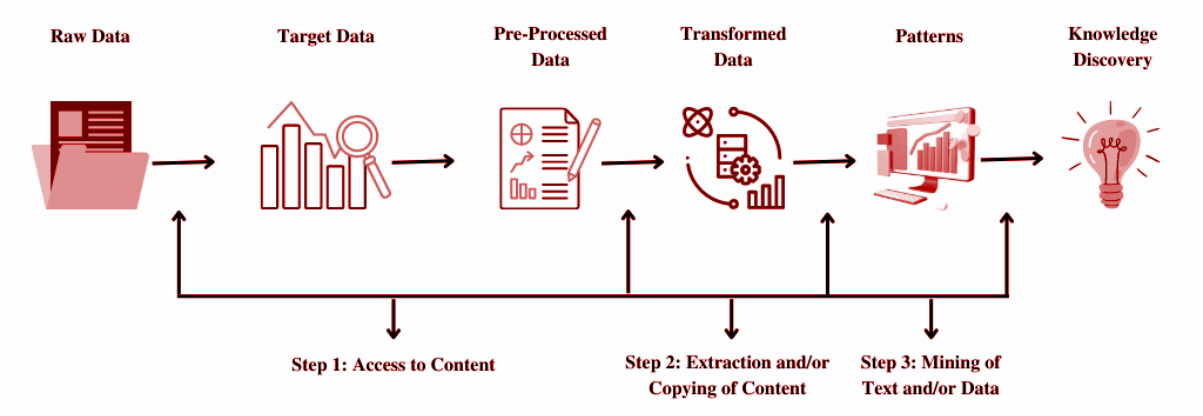
[18] E Rosati, 'the Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspects' (2018), at 1. Regarding applications of TDM, *see*, OpenMinted, 'TDM Stories'. Available at http://openminted.eu/tdm-stories/, accessed December 10, 2022.

[19] E Rosati. *Copyright in the Digital Single Market : Article-by-Article Commentary to the Provisions of Directive 2019/790* (Oxford, 2021) at 68.

[20] It is important to distinguish the definition of raw data, target data and pre-processed data. First, Raw data may be recorded on magnetic media, computer printouts, microfilm or microfiche copies, dictated observations, and automated instrumentation, as well as on photographs, microfilm or microfiche copies, and computer printouts. RD McDowall, 'Focus on Quality: What Exactly Are Raw Data?', (2016), Available at: <https://rx-360.org/wp-content/uploads/2018/08/What-Exactly-Are-Raw-Data-by-R.D.-McDowall-2016.pdf.> accessed July 10, 2022, at 18. Second, target data is a component of a dataset that we are interested in learning more about. A supervised machine learning method makes use of previous data to discover links between the target and other features of our dataset. *See,* 'Target Variable and DataRobot Artificial Intelligence Wiki', (2022). Available at: <https://www.datarobot.com/wiki/target/> accessed July 10, 2022. Third, Data pre-processing, which converts the data into a format that is more readily and efficiently processed in data mining, machine learning, and other data science tasks, produces pre-processed data. To ensure reliable findings, the techniques are typically applied at the very beginning of the machine learning and AI development pipeline. *See*, G Lawton, 'Data Pre-processing', (2022). Available at: <https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing.> accessed December 10, 2022.

The final method in most GenAI models occurs in step 3, as shown in Illustration 1 below.[21] In most cases, mining of text and/or data and knowledge discovery includes data cleaning and pre-processing, data transformation, and pattern evaluation. First, to increase the dependability of the data and its effectiveness, data cleaning and preprocessing will look for missing data and delete noisy, redundant, and low-quality data from the data collection.

Illustration 1. Three Common Steps in TDM.



Based on application-specific criteria, specialized algorithms are utilized to search for and remove undesirable data.[22] Second, data transformation prepares data for use by data mining algorithms. As a result, the data must be consolidated and aggregated. The data is consolidated based on functions, attributes, features, and so on.[23] Third, pattern evaluation requires the trend and patterns obtained from various data mining methods and iterations to be represented in discrete forms such as bar graphs, pie charts, histograms, and so on to study the impact of data collected and transformed during previous steps.[24]

3. TDM, GenAI Models and Copyright

3.1 TDM and GenAI: As Close As Two Coats of Paint

'There is no reason why the simple shapes of stories can't be fed into computers'

\- K VONNEGUT.[25]

In 1995, Vonnegut presented his theory about the shapes of stories. The theory holds that emotional arcs can take a variety of forms and that stories often follow them. In his lecture,

---

[21] E Rosati, *supra* note 19, at 71. For more details regarding the objective of predictive TDM, *see also*, for further reference, UM Fayyad, G Piatetsky-Shapiro and P Smyth, 'Knowledge Discovery and Data Mining: Towards a Unifying Framework.'(KDD, 1996). *See also*, for further reference, RA Sarker, *et al*, 'Introducing Data Mining and Knowledge Discovery.' (2002).

[22] R Sharma, 'KDD Process in Data Mining: What You Need to Know', (2020), Available at: <https://www.upgrad.com/blog/kdd-process-data-mining/> accessed December 10, 2022.

[23] *Ibid.*

[24] *Ibid.*

[25] Short lecture by K Vonnegut on the 'simple shapes of stories.' Available at: <https://www.youtube.com/watch?v=oP3c1h8v2ZQ> accessed December 10, 2022.

Vonnegut sketched up a number of storylines, such as "Man falls into a hole, Man climbs out of a hole" and the more complicated "Boy meets Girl, Boy loses Girl, Boy gets Girl." However, there is no consensus about the number of various emotional arcs that appear in stories or how long it takes a story to reach its climax.[26] A couple of decades later, we are finally witnessing a major shift in the process of mapping 'emotional arcs'. Researchers at the University of Vermont in Burlington used sentiment analysis to map the emotional arcs of over 1,700 stories and then used TDM techniques to reveal the most common arcs.[27] This research eventually inspired GenAI models researchers and developers and proved that TDM may be used to train machine learning, which is one of the most fundamental parts of AI, for the aim of AI-driven creativity.[28]

There are myriad examples of GenAI models producing artistic and literary content,[29] ChatGPT-4, DALL-E 2 and Stability AI are some of the GenAI models that have caught the attention of many people worldwide. This section will focus on analysing the use of DALL-E 2 and Stability AI systems. In 2022, OpenAI and Stability AI introduced a revolutionary deep neural network that can create original, realistic images and art from a text description, inspired by Vonnegut's theory, for example, "an astronaut chilling on Mars," or "a teddy bear playing a basketball."[30] In its operation process, both DALL-E 2 and Stability AI use the TDM technique to obtain realistic images and art from a text description.[31] It employs a technique known as "stable diffusion,"[32] which begins with a pattern of random dots and progressively changes that pattern to resemble a picture when it identifies certain characteristics of that

---

[26] "Data Mining Reveals the Six Basic Emotional Arcs of Storytelling | MIT Technology Review" (MIT Technology Review, July 6, 2016) <https://www.technologyreview.com/2016/07/06/158961/data-mining-reveals-the-six-basic-emotional-arcs-of-storytelling/> accessed December 10, 2022.

[27] AJ Reagan, L Mitchell, D Kiley, *et al*. 'The emotional arcs of stories are dominated by six basic shapes.' (EPJ Data Sci. 5, 31, 2016).

[28] *S*N Kühl, M Goutier, R Hirt and G Satzger, 'Machine Learning in Artificial Intelligence: Towards a Common Understanding.' (2019).

[29] e.g., 'DALL-E 2', Available at: <https://openai.com/dall-e-2/,> accessed 10 July 2022. 'Ai-Da' the world's first ultra-realistic artist robot, Available at: <https://www.ai-darobot.com/,> accessed 10 July 2022. 'MuseNet' (OpenAI, 25 April 2019), Available at: <https://openai.com/blog/musenet/> accessed 10 July 2022 (music generation); 'InferKit Demo'. Available at: <https://app.inferkit.com/demo> accessed 10 February 2022 (text generation); 'Image GPT' (OpenAI, 17 June 2020). Available at: <https://openai.com/blog/image-gpt/ > accessed 10 February 2022; A Newitz, 'An AI Wrote All of David Hasselhoff's Lines in This Bizarre Short Film' (Ars Technica, 25 April 2017). Available at: <https://arstechnica.com/gaming/2017/04/an-ai-wrote-all-of-david-hasselhoffs-lines-in-this-demented-short-film/ > accessed December 10, 2022.

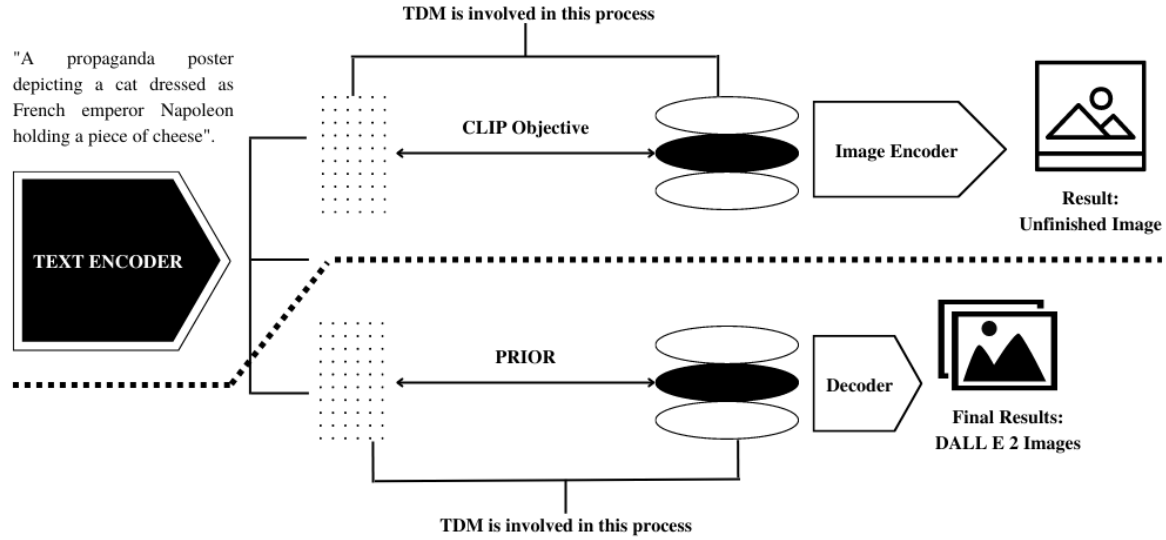
[30] DALL E 2 official website, *ibid*. Stability AI, Available at: <https://stability.ai/> accessed August 1, 2022. *See also*, for further reference, A Elgammal, B Liu, M Elhoseiny and M Mazzone, 'CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms.' (2017) Available at: <https://arxiv.org/abs/1706.07068.> accessed July 20, 2022. at 1.

[31] Raw and pre-processed data used in the TDM process are texts and hundreds of images which are then analysed based on text input from the user to get the final result. *Ibid.*

[32] Stable Diffusion is a text-to-image latent diffusion model created by the researchers and engineers from CompVis, Stability AI and LAION. It is trained on 512x512 images from a subset of the LAION-5B database. LAION-5B is the largest, freely accessible multi-modal dataset that currently exists. S Patil, *et al*, 'Stable Difussion with Difuser' (Stability AI, 2022) Available at: <https://github.com/huggingface/blog/blob/main/stable_diffusion.md.> accessed August 22, 2022.

image.[33] Both DALL-E 2 and Stability AI are operated by a contrastive model called CLIP or 'Contrastive Language-Image Pre-training' which has been shown to learn robust representations of images that capture both semantics and style.[34] Stability AI, however, has obtained its training data from the world's best multi-modal datasets called "LAION-5B".[35] This dataset is "a CLIP-filtered dataset of 5.85 billion high-quality image-text pairs, their CLIP ViT-L/14 embeddings, kNN-indices, a web interface for exploration & subset-creation and NSFW- and watermark-detection scores and tools." The datasets used by LAION-5B are licensed under the Creative Common CC-BY 4.0 license.[36] In the example of DALL-E 2, as shown in Illustration 2 below, the system involves four iterative stages to produce an image namely (1) CLIP, (2) Prior Model, (3) Decoder Diffusion Model or unCLIP and (4) DALL-E 2 as the final output.[37]

Illustration 2. A Simplified unCLIP Training Process.[38]



Without TDM, the DALL-E 2 system cannot perform steps 1, 2 and 3 for the following reasons:[39] First, from the stage (1) until (3), the DALL-E 2 system analyzes hundreds of texts and images with the TDM method. In these three stages, DALL-E 2 does not copy the copyrighted works being fed to the system, instead, the system uses the data to find a new

[33] DALL E 2 official website, *ibid.*

[34] A Ramesh, P Dhariwal, A Nichol, C Chu and M Chen, 'Hierarchical Text-Conditional Image Generation with CLIP Latents'. (ArXiv, 2014). Available at: <https://arxiv.org/pdf/2204.06125.pdf.> accessed August 20, 2022.

[35] C Schuhmann, R Vencu, R Beaumont, T Coombes, C Gordon, A Katta, R Kaczmarczyk, and J Jitsev, 'LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets.' (2022) Available at: https://laion.ai/blog/laion-5b/> accessed August 2022.

[36] *Ibid.*

[37] A Romero, 'DALL·E 2, Explained: The Promise and Limitations of a Revolutionary AI', (2022). Available at: <https://towardsdatascience.com/dall-e-2-explained-the-promise-and-limitations-of-a-revolutionary-ai-3faf691be220.> accessed August 8, 2022.

[38] Based on a similar figure from A Ramesh, *et al*, *ibid*, at 3.

[39] *Ibid.*

pattern. DALL-E 2 is an example of a two-part model consisting of a previous model and a decoder or unCLIP.[40] Second, the decoder is termed unCLIP because it reverses the original CLIP model's (step 1) process and TDM makes it possible by assisting the system to construct a 'mental' representation (embedding) from an image and make an original picture from a generic mental representation. Last, with the help of TDM, the mental representation encodes the main features that are semantically meaningful during the process such as pictures of people, animals, objects, style, colours, background, etc., so that the DALL-E 2 system can generate a novel image that retains these characteristics while varying the non-essential features.

To conclude, the TDM processes employed in DALL-E 2 generate robust and diverse data sets that are then used to feed and train the DALL-E 2 system or any other GenAI models for creative purposes. However, DALL-E 2 does not publicly announce where they obtain the training data. If they used a dataset available online, there is a possible conflict between TDM techniques and copyright protection, because works or subject matter used in the TDM process, such as pictures and text, may be protected under copyright law.[41] In the EU, under Directive 2001/29/EC (InfoSoc Directive),[42] Directive 2009/24/EC (Software Directive)[43] or Directive 96/9/EC (Database Directive),[44] one is required to ask the relevant rightholder's permission before copying a work.

3.2 Can Big Data be Protected by Copyright and Related Rights?

The emergence of AI-driven creativity is predominantly driven by the rising availability of data.[45] It is nearly impossible for any GenAI models to analyse large amounts of digital text and/or data to discover new patterns without the help of TDM.[46] Because the value of data does not lie in the data or text itself, but in the extraction of value,[47] and since the main function of data during the TDM process is to find new patterns, should GenAI models researchers and developers worry about the copyright protection of the data being used in the extraction phase? As being said, "one of the basic and fundamental principles of copyright law is that data is as such not protected; copyright only protects the creative form, not the information incorporated

[40] This AI system converts a sentence to a picture by concatenating both models. DALL E 2 inserts a text into the 'black box,' through TDM which generates a well-defined image. *Ibid.*

[41] K Christensen, *supra* note 11, at 18.

[42] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society, OJ L 167, 22.6.2001, at 10-19 (InfoSoc Directive).

[43] Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (codified version) OJ L 111, 5.5.2009, at 16-22 (Software Directive).

[44] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77, 27.3.1996, at 20-28 (Database Directive).

[45] European Commission, 'White Paper on Artificial Intelligence: a European Approach to Excellence and Trust', Brussels (2020), at 1. *See also,* Christensen K, *supra* note 11, at 18.

[46] E Rosati, *supra* note 8, at 2.

[47] E Rosati, *supra* note 7, at 429.

in the protected work".[48] Given that certain uses in TDM may not be subject to copyright laws, GenAI models researchers and developers may not need to worry about any copyright and related rights issues.[49] As it has been argued by Geiger, Frosio and Bulayenko:[50]

> this activity [TDM] is outside the scope of exclusive rights and that any restriction would amount to undermine the underlying rationales of copyright protection and result in an inadmissible restriction of freedom of expression and information as protected by e.g. the European Court of Human Rights (ECHR) and the Charter of Fundamental Rights of the European Union.

The potential of copyright infringement in this circumstance does not pose a concern because data as such is not protected by copyright.[51] However, given the three Vs (volume, velocity, and variety) that apply to big data, ordinary "data" must be separated from big data. As a result, copyright may exist in the text, images, sounds, and other artistic works, which are eventually susceptible to TDM activities.[52] Moreover, big data may apply to the right of reproduction as well as sui generis database rights in some instances. As shown in Illustration 1, not all TDM activities include data copying and/or extraction throughout the mining process, which occurs in step 2. The material used, technological instruments used, and the scope of the mining technique mostly determine copying.[53] Not all copying activities require prior consent, such as those that come outside the purview of EU Acquis' exceptions and limitations.[54] However, there may be legal restrictions in place when TDM techniques include copying and/or extracting the relevant data for AI projects.[55] By way of example, the CJEU confirmed in the landmark case of Infopaq I, C-5/08 when it was ruled that copying of text excerpts containing at least eleven words of copyrightable materials may trigger copyright protection (and the risk

---

[48] C Geiger, G Frosio and O Bulayenko, 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU'. CEIPI Research Paper No. 2019-08 (2019). at 6. *See*, the discussion on policy measures on TDM around the world, SM Fiil-Flynn, B Butler, M Carroll, O Cohen-Sasson, C Craig, L Guibault, P Jaszi, BJ Jütte, A Katz, JP Quintais, T Margoni, AR de Souza, M Sag, R Samberg, L Schirru, M Senftleben, O Tur-Sinai, and JL Contreras. 'Legal reform to enhance global text and data mining research.' *Science (New York, N.Y. 2022)*, *378*(6623), 951–953. https://doi.org/10.1126/science.add6124 *See also,* the seminal work by PB Hugenholtz, 'Auteursrecht op Informatie' (Kluwer, Deventer, 1989).

[49] C Gerrish and AM Skavlan, 'European Copyright Law and Text and Data Mining Exceptions and Limitations: In Light of the Recent DSM Directive, is the EU Approach a Hindrance or Facilitator to Innovation in the Region? (Stockholm Intellectual Property Law Review, Volume 2, Issue 2, 2019) at 58.

[50] C Geiger, G Frosio and O Bulayenko, *ibid*, at 7.

[51] C Gerrish and AM Skavlan, *ibid,* at 58.

[52] *Ibid.*

[53] E Rosati, *supra* note 8, at 200.

[54] K Christensen, *supra* note 11, at 18. *See also,* Recital 33 InfoSoc Directive; See also CJEU, Judgement of 26 April 2017, Stitching Brein v Jack Frederik Wullems, C-527/15, EU:C:2017:300, para 65 and 69, when the pre-installed add-ons permit access to private servers where copyright protected works have been made available to the public without the rightholder's authorization, the CJEU reaffirmed that this exemption cannot be relied upon by users. This was the case when the CJEU considered the term "lawful use" in Article 5(1) InfoSoc Directive.

[55] E Rosati, *supra* note 8, at 206-209. *See also*, Christensen K, *supra* note 11, at 21.

of infringement).[56] In this context, the possibility of copyright infringement occurs since AI depends on processing vast amounts of data derived through TDM, particularly in any GenAI models when TDM is applied to Big Data comprising protectable works like text and images.[57] Moreover, in terms of related rights, the CJEU in Pelham, C-476/17[58] established that "recognisability" rather than "originality" serves as the primary criterion for related rights, meaning that even minor elements of a bigger work may be eligible for related rights protection.[59] TDM will certainly violate related rights of the relevant rightholders because it may involve reproduction that results in the creation of a copy of the protected work without the possibility of choosing specific parts from that work during the TDM process that may not meet the standard for recognisability or additional alteration of the work per se.[60]

From the lens of sui generis database right, TDM may infringe the extraction and the re-utilisation of a substantial part of the contents of a database, when processing Big Data for AI. In this regard, the CJEU has in BHB v. WH, C-203/02 affirmed that the temporary or permanent transfer of data from one media to another and storage of that data is sufficient to be regarded as an extraction. As a result, this right will cover TDM since this operation is

---

[56] Judgement of 16 July 2009, Infopaq International A/S v Danske Dagblades Forening, C-5/08, ECLI:EU:C:2009:465 (Infopaq I), para 45-48. Whereas para 48 provides:

> (48) In the light of those considerations, the reproduction of an extract of a protected work which, like those at issue in the main proceedings, comprises 11 consecutive words thereof, is such as to constitute reproduction in part within the meaning of Article 2 of Directive 2001/29, if that extract contains an element of the work which, as such, expresses the author's own intellectual creation; it is for the national court to make this determination.

See also, reasoning by E Rosati, 'Copyright at the CJEU: Back to the start (of copyright protection)' (2022). Forthcoming in H Bosher and E Rosati (eds), Developments and Directions in Intellectual Property Law. 20 Years of The IPKat (Oxford University Press, 2023), Available at: <https://ssrn.com/abstract=4097316.> accessed July 25, 2022. *See also*, C Gerrish and AM Skavlan, *supra* note 49, at 59.

[57] K Christensen, *supra* note 11, at 21. *See also*, Judgement of 2 May 2012, SAS Institute Inc. v World Programming Ltd, C-406/10, ECLI:EU:C:2012:259, para 66-67, where the CJEU upholds the Infopaq test in regards to reproduction of computer programs.

[58] Judgement of 29 July 2019, Pelham GmbH and others v. Ralf Hutter and Others (Pelham), C-476/17, EU:C:2019:624, para 31 and 39. Whereas para 39 provides:

> (39) In the light of the foregoing considerations, the answer to the first and sixth questions is that Article 2(c) of Directive 2001/29 must, in the light of the Charter, be interpreted as meaning that the phonogram producer's exclusive right under that provision to reproduce and distribute his or her phonogram allows him or her to prevent another person from taking a sound sample, even if very short, of his or her phonogram for the purposes of including that sample in another phonogram, unless that sample is included in the phonogram in a modified form unrecognisable to the ear.

For in-depth analysis, *see also,* M Senftleben, 'Flexibility Grave – Partial Reproduction Focus and Closed System Fetishism in CJEU, Pelham.' (IIC 51, 751–769, 2020). Jütte and Quintais discuss the normative implications of the CJEU's judgment on the interplay between fundamental rights and copyright, particularly in a digital environment. *See*, BJ Jütte and JP Quintais, 'The Pelham Chronicles: sampling, copyright and fundamental rights', (Journal of Intellectual Property Law & Practice, Volume 16, Issue 3, 2021) at 213–225.

[59] E Rosati, *supra* note 8, at 206; *See*, C Geiger, G Frosio and O Bulayenko, *supra* note 48, at 6; *See also*, Christensen K, *supra* note 12, at 22.

[60] K Christensen, *supra* note 11, at 22.

essential for the process.[61] Hence, to lawfully conduct TDM, GenAI models researchers and developers would always need authorisation from the relevant rightholders. However, when TDM may be entitled to protection under the statutory and non-mandatory pre-existing exceptions and limitations provided in the EU *acquis*, such authorization is not necessary.[62] However, the question remains: will the current legal framework (copyright exceptions and limitations) suffice to accommodate the advancement of technologies, especially TDM for the development of GenAI models?

III. Reclassifying Text and Data Mining Exceptions in the EU

1. Introduction

The new EU CDSM Directive was officially published on May 17, 2019.[63] The legislation draft intends to modernize copyright and related rights in the digital era and "contribute to the functioning of the internal market, provide for a high level of protection for rightholders, facilitate the clearance of rights, and create a framework in which the exploitation of works and other protected subject matter can take place."[64] With 86 recitals and 32 articles, the EU CDSM Directive is one of the copyright acquis' longer pieces. It has five titles, comprising title I, 'Measures to Adapt Exceptions and Limitations to the Digital and Cross-Border Environment,' focusing on TDM exceptions. Two different types of exceptions are included in Title II of the EU CDSM Directive as a series of steps to adapt exceptions and limitations to the digital and cross-border environment.

2. Articles 3 and 4 of the EU CDSM Directive: An Overview

Title II of the EU CDSM Directive sets two types of new TDM exceptions and limitations that the EU Member States must provide. Article 3 of the EU CDSM Directive provides legal certainty for the researcher to conduct TDM for scientific research. Moreover, reproductions and extractions of works or subject matter for TDM other than scientific research are exempted under Article 4 of this Directive. Both TDM provisions take into account the need for lawful access which means anyone who uses copyrighted works for TDM purposes must have lawful access thereto and GenAI models are no exception. Both articles will be

---

[61] BHB v. WG, C-203/02, para 65—66. *See*, K Christensen, *supra* note 11, at 23. *See also*, JP Triaille *et al*., 'Study on the Legal of Text and Data Mining (TDM)' (De Wolf and Partners, Funded by the European Commission, 2014) at 28.

[62] K Christensen, *ibid*, at 23; *See*, recital 31 InfoSoc Directive; *See also*, fur further reference, T Chinou, 'Copyright Lessons on Machine Learning: What Impact on Algorithmic Art? (JIPITEC, 2019) at 402. Available at: <https://www.jipitec.eu/issues/jipitec-10-3-2019/5025.> accessed December 10, 2022.

[63] EU CDSM Directive. For an earlier overview of the directive, *see*, T Shapiro and S Hansson, 'The DSM Copyright Directive - EU copyright will indeed never be the same' (E.I.P.R. 41(7), 2019) at 404-414; *See also*, for a critical analysis regarding the new directive, JP Quintais, 'The New Copyright in the Digital Single Market Directive: A Critical Look' (European Intellectual Property Review, 42(1), 2020) at 28-41; S Dusollier, 'The 2019 Directive on Copyright in the Digital Single Market: Some progress, a few bad choices, and an overall failed ambition' (Common Market Law Review, Issue 4, 2020) at 979-1030.

[64] Recital 1 of the EU CDSM Directive.

discussed and introduced in the sections that follow, along with any comments and objections from the literature.[65]

2.1 Article 3 – Scientific Research Exception

Article 3 of the EU CDSM Directive provides an exception for the acts of "reproductions and extractions made by research organisations and cultural heritage institutions to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access." As has been said, only research organizations and cultural heritage institutions are permitted to conduct TDM for scientific research purposes, and they must have lawful access to the works or subject matter in question.[66] Besides research organisations, Article 3 also allows cultural heritage institutions, which encompass publicly accessible libraries and museums, archives, film or audio heritage institutions, and other heritage institutions.[67] To be eligible for the application of Article 3 of the Directive, it is unclear whether research organisations and cultural heritage institutions must be established in the EU.[68] Furthermore, the concept of scientific research is only hinted at in Recital 12 and is intended to include both the natural and human sciences.[69] Article 3 of the EU CDSM Directive is narrowly defined, only several entities as previously mentioned above can benefit from the exception to conduct TDM for scientific research purposes only. Any contractual provision that conflicts with the exceptions provided by Article 3 will be unlawful.[70] The regime of the new exceptions in Article 3 EU DSM Directive is summarised in the following table.

2.2 Article 4 – The Limited Exception

Article 4 of the EU CDSM Directive provides an exception for reproductions and extractions of lawfully accessible works/subject matter for TDM to provide significant legal certainty for both private and public entities undertaking TDM.[71] This means that, unlike the strictly limited beneficiaries in Article 3, any entity can profit from the TDM exception under Article 4 of this Directive. This exemption, however, is subject to rightholders reservations, including through "machine readable means in the case of content made publicly available online."[72] To put it in another way, this is solely an opt-out mechanism in which the relevant rightholders can prevent others from conducting TDM. Moreover, Article 4 is intended to provide legal clarity for the TDM users which do not fulfil all the conditions of the existing exception for temporary acts of reproduction provided for in Article 5(1) of Directive 2001/29/EC by allowing "the copies

---

[65] Recital 14 of the EU CDSM Directive. *See also,* Article 3(1) and 4(1) of the EU CDSM Directive.

[66] Article 2(1) EU CDSM.

[67] Article 2(3) and Recital 13 of the EU CDSM Directive.

[68] E Rosati, *supra* note 19, (2021), at 42-44.

[69] S Dusollier, *supra* note 63, (2019), at 986. *See also,* E Rosati *ibid,* at 43.

[70] EU CDSM Directive, Article 7(1).

[71] Article 4(1) of the EU CDSM. *See*, *Ibid,* at 8. *See also,* Recital 18 of the EU CDSM Directive.

[72] Article 4(3) of the EU CDSM Directive. *See also,* JP Quintais, *supra* note 63, (2020), at 8.

made to be retained for as long as is necessary for those text and data mining purposes."[73] The new exceptions regime in Article 4 of the EU DSM Directive is summarized in Table 4 below.

3. The New TDM Exceptions: Analysis, Responses and Critiques from the Literature

There are numerous advantages to the newly introduced TDM exceptions in the EU. The inclusion of Articles 3 and 4 of the Directive achieves the following primary policy objectives: First, it is intended to create a standardised, uniform level playing field for researchers across the EU to conduct TDM projects lawfully. Second, the Directive focuses on harmonising the legislation of the Member States through a mandatory solution, in which a unified framework for TDM activities under the EU CDSM Directive will accelerate innovation by encouraging EU-wide, coordinated, bigger research programs.[74] However, the new reform continues to have negative consequences.[75] This section will include notable scholars' remarks, responses, and criticisms.[76] They range from the scope of the exception that applies to unqualified beneficiaries through an opt-out mechanism to the numerous restrictions that apply to the research purpose exception. A summary of the key points of assessment is provided below.

3.1 An Overall Assessment of the Reform: Articles 3 and 4

Several copyright scholars contend that TDM should be outside the copyright realm.[77] Margoni and Kretschmer argue that the formulation of the two new TDM exceptions is "conceptually wrong, theoretically flawed and normatively unambitious."[78] As the saying goes, 'the right to

---

[73] Recital 18 of the EU CDSM Directive. *See also,* Article 5(1) of the Directive 2001/29/EC of the European Parliament and Council (InfoSoc Directive). This compulsory exception may apply to the limited TDM method, as specified in Recital 9 of the EU CDSM Directive. *See also*, C Bernault, 'Les nouvelles exceptions au droit d'auteur', (2017), Juris art etc., Vol. 47, at. 22; for a commentary of Art. 5(1) InfoSoc Directive, *see also,* MM Walter, and S Von Lewinski, 'Information Society Directive', (2010), in MM Walter, and S Von Lewinski. (eds), *European Copyright Law: A Commentary* (New York, USA: Oxford University Press) at 968-969 and 1024-1027.

[74] C Geiger, G Frosio and O Bulayenko, *supra* note 48, at 29.

[75] Given the broader TDM exception's uncertain limited applicability, *see also,* for a critical assessment on this matter, R Ducato and A Strowel, 'Limitations to text and Data Mining and Consumer Empowerment: Making the Case for a Right to 'Machine Legibility',' (IIC Vol. 50, No. 6, 2019) at 649.

[76] The need to ensure that the exceptions' usefulness is not unreasonably hampered by technical protection measures is a recurring theme among commentators, but such problems are not addressed here because they are removed from the main topic. *See*, Liber Europe, 'Europe's TDM Exception for Research: Will It Be Undermined by Technical Blocking from Publishers?' (2020). Available at: <https://libereurope.eu/article/tdm-technical-protection-measures/.> Accessed August 9, 2022. *See also,* T Margoni and M Kretschmer, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology,' (2021). Available at: <https://ssrn.com/abstract=3886695.> accessed December 10, 2022.

[77] T Margoni and M Kretschmer, *ibid*, (2021), at 22. *See also*, ECS, 'General Opinion on the EU Copyright Reform Package', (2017), at 5. Available at: <https://europeancopyrightsocietydotorg.files.wordpress.com/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf.> accessed August 9, 2022. *See,* for example, with further references, Sag M. 'The New Legal Landscape for Text Mining and Machine Learning', (66 J. of the Copyright Soc'y of the USA, 3, 9-19, 2019). CJ Craig, 'Globalizing User Rights-Talk: On Copyright Limits and Rhetorical Risks' (Am. U. Int'l L. Rev. Vol. 33(1), 2017).

[78] T Margoni and M Kretschmer, *ibid*, (2021), at 4.

read should be the right to mine,'[79] but not when it is blocked by the requirement of lawful access and restriction to the specific beneficiaries. Article 3 of the EU CDSM Directive should not be restricted to research organizations, but should be accessible to all entities who have lawful access to underlying mined materials, especially to avoid hindering start-ups, GenAI models researchers and developers and independent researchers in AI in general.[80] The difference between commercial and non-commercial purposes was also heavily criticized.[81] Furthermore, Hilty and Richter hold that the requirement of lawful access has the potential to disadvantage smaller or less wealthy research organizations or institutions.[82] By denying lawful access, relevant rightholders can effectively prevent certain parts of existing work, for example, from ever being subject to TDM.[83] Others have challenged the requirement of lawful access, fearing that rightholders may incorporate TDM in their pricing and further escalate overall costs.[84] Higher prices, TDM fees, and the availability of the resource for licensing may result in lower quality and/or quantity of TDM.[85]

To summarize, the new TDM exceptions regime has severe flaws: as Quintais points out, the scope of the articles is too narrow and "this regime will probably not lead to simplification and harmonisation of the system of exceptions in EU copyright law, as it continues to allow significant cherry-picking by the Member States."[86] The Article's emphasis on 'reproduction

---

[79] C Geiger, G Frosio and O Bulayenko. 'Crafting a Text and Data Mining Exception for Machine Learning and Big Data in the Digital Single Market.' (2018), at 109. In X Seuba, C Geiger and J Pénin (eds), *Intellectual Property and Digital Trade in the Age of Artificial Intelligence and Big Data* (CEIPI/ICTSD publication series on 'Global Perspectives and Challenges for the Intellectual Property System,' Vol. 5) at 95-112.

[80] C Geiger, G Frosio and O Bulayenko. *ibid (2018),* at 109. Several scholars have argued for such an expansion of the scope of the limitation. *See,* for example, with further references, N Jondet, 'L'exception pour le data mining dans le projet de directive sur le droit d'auteur: pourquoi l'Union européenne doit aller plus loin que les législations des Etats membres,' (Propriétés intellectuelles 67, 2018), at 33–34. *See also,* P Kollár, *supra* note 17, at 17.

[81] P Kollár, *ibid*, at 17. TDM exceptions should not differentiate between commercial and non-commercial research purposes. The notion of non-commercial is ambiguous and subject to several interpretations. "The lines are becoming less defined between purely non-commercial research and research that has commercial potential or has been funded by commercial entities." *See also*, Liber Europe, 'LIBER Position Statement: Copyright in the Digital Age'. Available at: <https://libereurope.eu/liber-position-statement-copyright-in-the-digital-age/.> accessed December 10, 2022.

[82] RM Hilty and H Reichter, 'Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules,' (Max Planck Institute for Innovation and Competition Research Paper 17-02, 2019) at 9. Available at: <https://pure.mpg.de/rest/items/item_2470998_12/component/file_2479390/content.> accessed January 10, 2023.

[83] European Copyright Society, 'General Opinion on the EU Copyright Reform Package,' (2017). Available at: <https://europeancopyrightsociety.org/.> accessed August 9, 2022. *See also,* C Geiger, G Frosio and O Bulayenko, *supra* note 48, (2019), at 22.

[84] P Kollár, *ibid*, at 17. *See also*, C Gerrish and AM Skavlan, *supra* note 49, at 4; C Geiger, G Frosio and O Bulayenko, *supra* note 48, (2019), at 34.

[85] C Geiger, G Frosio and O Bulayenko, *ibid*, at 22.

[86] JP Quintais, *supra* note 63, (2020), at 11 (address the issues raised by the potential of different national implementations of the voluntary exclusions in Art. 5 of the InfoSoc Directive). *See also*, further reference, L

and extraction' has been seen as potentially problematic for communicating TDM outcomes, especially for GenAI models that work with Natural Language Processing (NLP) such as ChatGPT-4, which is trained on a variety of copyright-protected datasets (i.e. texts) and potentially involves reproduction in part, those models cannot be distributed or communicated publicly since reproducing a work, as little as 11 consecutive words, could be protected by copyright.[87] However, in the case of DALL-E 2 and any other AI image generator systems, the standard for copying might be difficult to meet, since the Infopaq case is irrelevant here.

3.2 Article 3: TDM for Scientific Research and Limited Beneficiaries

The inclusion of Article 3 of the EU CDSM Directive achieves major policy objectives.[88] "It is set to provide a normalised level playing field for researchers across Europe to lawfully carry out TDM projects. The major positive impacts of the proposal lie in its focus on harmonisation of member states' laws, through a mandatory solution."[89] Some praised the reduction in fragmentation of national approaches to TDM,[90] but others noted that the promised harmonization and legal certainty did not occur due to inadequate wording and regulatory process.[91] Furthermore, the prohibition on contractual override is a breakthrough that should be appreciated. This is a critical provision because, as previously stated, to conduct TDM, GenAI models researchers and developers must access numerous databases containing copyrighted materials and accept Terms of Use that frequently limit TDM.[92]

Further, the notion of research organisation received critiques. Some argue that the scope of Article 3 is prohibitively narrow when defining the nature of the research organization.[93] To be eligible for TDM, research organisations "must operate on a not-profit basis, or re-invest all its profits into its scientific research or pursue a public interest mission funded by public funds or public contracts."[94] In this context, commercial-based research organisations such as

Guibault, 'Why Cherry-Picking Never Leads to Harmonisation: The Case of the Limitations on Copyright under Directive 2001/29/EC', (1 J.I.P.I.T.E.C, (2) 2603, 2010).

[87] T Margoni and M Kretschmer, *supra* note 76, (2021), at 25.

[88] C Geiger, G Frosio and O Bulayenko, *supra* note 79, (2018) at 104.

[89] *Ibid.*

[90] C Geiger, G Frosio and O Bulayenko, *supra* note 48, (2019), at 36. C Gerrish and AM Skavlan, *supra* note 49, at 63. *See also,* P Kollár, *supra* note 17, at 18.

[91] JP Quintais, *supra* note 63, (2020), at 23. *See also*, P Kollár, *ibid.*

[92] C Geiger, G Frosio and O Bulayenko, *supra* note 48, (2019), at 25. Ducato and Strowel discusse the prohibition of TDM in T&C, *see also,* for further reference, R Ducato, and A Strowel, *supra* note 75, at 21-24. (Stating: "the analysis focused on the terms and conditions (T&C) of 21 online platforms. 20 out of 21 platforms published the T&C on their website and 14 of them contained specific intellectual property clauses, directly or indirectly, related to TDM activities. More specifically, four platforms expressly prohibit TDM on the website").

[93] RM Hilty and H Reichter, *supra* note 82, at 4; C Geiger, G Frosio and O Bulayenko, *supra* note 48, (2019), at 25; Margoni and Kretschmer, *supra* note 80, (2021).

[94] Recital 12 of the EU CDSM Directive.

OpenAI,[95] the creator of ChatGPT-4 and DALL E 2 or will not be able to conduct their TDM with non-personal data available in the EU, as they are not eligible to do so. The EU legislators explicitly stated that they wanted to ensure that scientific research undertaken for TDM purposes remained neutral and independent from industry. However, keep in mind that public funding and investment are scarce, and many research organizations rely on the private sector to obtain the required funding for cutting-edge research.[96] This narrowly-defined research organization is capable of putting innovation in the EU on the back burner. Last, the EU CDSM Directive does not define the terminology of 'scientific research'. The specific purpose of scientific research in Article 3 has been criticized as possibly generating issues for existing licenses, such as those for educational purposes, and may lead to restrictive interpretations.[97]

3.3 Article 4: The EU Obsession with Licensing?

The reservation or opt-out mechanism in Article 4 has drawn intense criticism because it might hamper the advancement of AI in the EU.[98] The most attention-grabbing point is Article 4(3) which allows the relevant rightholders to reserve the right to perform TDM activities, As of now, reservations are made as mentioned in Recital 18 in an 'appropriate manner'.[99] For this purpose, the recital differentiated between two separate scenarios such as: First, in the case of content that has been made publicly available online, it should only be considered appropriate to reserve the rights in Article 4(1) by the use of machine-readable means.[100] Second, it might be appropriate to reserve the rights by other means, such as contractual agreements or a unilateral declaration.[101]

Overall, rightholders shall only be allowed to reserve the TDM activity for content that is publicly available online by implementing appropriate technological measures, in line with the analogy drawn by the court in *VG Bild-Kunst*, C-392/19.[102] It should be noted, though, that in

---

[95] OpenAI is an AI research and deployment company. Their mission is to ensure that artificial general intelligence benefits all of humanity. *See* OpenAI official website. Available at: <https://openai.com/about/.> accessed January 10, 2023.

[96] C Gerrish and AM Skavlan, *supra* note 49, at 62.

[97] P Kollár, *supra* note 17, at 18. C Geiger, G Frosio and O Bulayenko, *supra* note 48, (2019), at 22. R Ducato and A Strowel, supra note 75.

[98] E Rosati. *supra* note 8, (2019), at 21; P Kollár, *ibid*, at 19; C Geiger, G Frosio and O Bulayenko, *supra* note 48, (2019), at 9; R Meys, 'Data Mining Under the Directive on Copyright and Related Rights in the Digital Single Market: Are European Database Protection Rules Still Threatening the Development of Artificial Intelligence?' (GRUR International 69(5), 2020) at. 457.

[99] E Rosati, *supra* note 19, (2021), at 89.

[100] This includes metadata and terms and conditions of a website or a service. In any event, other uses shall not be affected by the reservation of rights for the purposes of TDM. *Ibid*, at 89-90.

[101] However, in light of the CJEU judgment in VG Bild-Kunst, C-392/19, it is necessary to adopt a corrected reading of the provision, in the sense that reservation by rightholders shall be only possible if done by adopting effective technological measures within the meaning of Article 6(1) and (3) of the InfoSoc Directive 2001/29. This modality, according to the Court, is the one that ensures legal certainty and the smooth functioning of the internet. *Ibid*, at 89-90.

[102] *VG Build-Kunst*, C-392/19, EU:C:2021:181, at [46]. See *Ibid*, at 89-90.

the absence of such robust technological measures in place, it might be daunting for GenAI researchers and developers to determine "whether the concerned rightholders intended to reserve the doing of TDM activities in relation to their copyright works and other protected subject matter, including when these are subject to sub-licenses."[103] However, even with effective technological measures, for most GenAI researchers and developers, Article 4 is a nightmare that comes true. As we know, GenAI researchers and developers need a huge amount of data corpus in the form of text, images etc, and gaining permission to mine from various rightholders can be an exhausting task. Some consider that Article 4 symbolizes the EU CDSM Directive's 'obsession with licensing' and, as a result, favours private ordering above public policy.[104]

IV. The Japanese TDM Exceptions: The New Paradise for AI & Machine Learning

1. Introduction

In 2016, Japan identified AI as one of the most important technological foundations for establishing a super-smart society, well known as 'Society 5.0.' To support the development of AI and technology, an AI Technology Strategy Council was established per instructions from Prime Minister Abe.[105] The Japanese government is getting ready for the 'singularity,' a terminology used to describe the time when AI surpassed human intelligence.[106] Hayashi, the CEO of HEROZ, Inc.,[107] One of the Japanese biggest GenAI model developers underscores the Japanese government's intention to support the development of AI-driven creativity by saying, "although AI engineers are in short supply throughout the world, Japan has a solid number of highly capable AI engineers."[108] To provide legal certainty and flexibility to AI innovators, the government issued the 'IP Strategic Program 2016' on May 9, 2016, which proposed the possibility of introducing 'flexible' provisions on copyright limitations to promote new digital innovations. The following year, on February 24, 2017, the subcommittee on Legal and Fundamental Affairs of the Council for Cultural Affairs' Subdivision on Copyright issued a report recommending the inclusion of such provisions in a bill to amend the

---

[103] E Rosati, *supra* note 19, (2021), at 90.

[104] P Kollár, *supra* note 17, at 19. *See also,* for further reference, P Samuelson, 'Europe's Controversial Digital Copyright Directive Finalized,' (Communications of the ACM 62(11), 2019) at 26; M. Senftleben *et al*., 'Ensuring the Visibility and Accessibility of European Creative Content on the World Market: The Need for Copyright Data Improvement in the Light of New Technologies' (SSRN Publication, 2021) at 6.

[105] The Japanese Government Official Website, 'Artificial Intelligence / the Government of Japan – JapanGov.' Available at: <https://www.japan.go.jp/tomodachi/2018/spring2018/artificial_intelligence.html> accessed January 10, 2023.

[106] This is not the first time the Japanese government mentioned the singularity. Back then in 2014, the Japanese government's research council 'launched a two-year study on how society should manage the singularity,' which includes 'the growth of artificial intelligence and its related technologies such as robots. *Ibid.*

[107] T Hayashi and T Takahashi founded HEROZ, Inc. in 2009, and created the Ponanza software. The company's online application "Shogi Wars" includes a feature that allows a player who is unsure of the best move to make AI choose the best possible strategy on his or her behalf for the next five moves by simply pressing a button. The Government of Japan Official Website, *ibid.*

[108] The Japanese Government Official Website, *ibid.*

copyright law.[109] Finally, on May 18, 2018, the Japanese government passed the bill to amend the Japanese Copyright Act which came into force on January 1, 2019.[110]

The new amendment contains three 'flexible' copyright exceptions, in which Article 30-4 is the 'newly born star' and is important for TDM. The exceptions include the small basket clause, which applies not only to the specific exploitations in the specified items, but also to any other equivalent exploitation and is applied *mutatis mutandis* to the neighbouring rights.[111] The Japanese government, through the introduction of the three exceptions "has insured that copyright cannot be an obstacle for AI"[112] and led Japan to become the new 'Paradise for AI and machine learning.'

2. Article 30-4 of the Japanese Copyright Act

This section focuses on Article 30-4 because it specifically allows TDM and have possible implications for the EU Member States.[113] Overall, Article 30-4 allows any exploitation of works for TDM purposes by classifying the activities into four categories, namely (1) extraction, (2) comparison, (3) classification or (4) other statistical analysis.[114] Someone can conduct TDM without permission from the relevant rightsholders "if the exploitation is aimed

---

[109] T Ueno, 'A General Clause on Copyright Limitations in Civil Law Countries: Recent Discussion on Japanese-Style Fair Use Clause.' (2021), at 213. In S Balganesh, W Loon, and H Sun, *The Cambridge Handbook of Copyright Limitations and Exceptions* (2021).

[110] Act No. 30 of 25 May 2018. *See*, in detail, the Japanese Copyright Research and Information Center (CRIC), 'the history of copyright systems in Japan.' Available at: <https://www.cric.or.jp/english/csj/csj2.html.> accessed August 11, 2022. *See also*, the Japanese Copyright Research and Information Center (CRIC), 'The Copyright of Japan (Official Translation)'. Available at: <https://www.cric.or.jp/english/clj/cl2.html> accessed January 10, 2023.

[111] T Ueno, 'The Flexible Copyright Exception for 'Non-Enjoyment' Purposes – Recent Amendment in Japan and Its Implication.' (GRUR International, 2021) at 147. *See*, Article 10 (1) of the Japanese Copyright Act.

[112] "Japan Amends Its Copyright Legislation to Meet Future Demands in AI" (European Alliance for Research Excellence, September 3, 2018). Available at: <https://eare.eu/japan-amends-tdm-exception-copyright/> accessed January 10, 2023.

[113] The Article provides the following:

> Article 30-4 (Exploitation without the Purpose of Enjoying the Thoughts or Sentiments Expressed in a Work)
>
> It is permissible to exploit a work, in any way and to the extent considered necessary, in any of the following cases, or in any other case in which it is not a person's purpose to personally enjoy or cause another person to enjoy the thoughts or sentiments expressed in that work; provided, however, that this does not apply if the action would unreasonably prejudice the interests of the copyright owner in light of the nature or purpose of the work or the circumstances of its exploitation:
>
> (i) if it is done for use in testing to develop or put into practical use technology that is connected with the recording of sounds or visuals of a work or other such exploitation;
> (ii) if it is done for use in data analysis (meaning the extraction, comparison, classification, or other statistical analysis of the constituent language, sounds, images, or other elemental data from a large number of works or a large volume of other such data; the same applies in Article 47-5, paragraph (1), item (ii));
> (iii) if it is exploited in the course of computer data processing or otherwise exploited in a way that does not involve what is expressed in the work being perceived by the human senses (for works of computer programming, such exploitation excludes.

[114] T Ueno, *supra* note 111, (2021), at 148.

at neither enjoying nor causing another person to enjoy the work unless such exploitation unreasonably prejudices the interests of the copyright holder."[115] The Japanese TDM exception is regarded as the "broadest TDM exception in the world" for the following reasons: (1) TDM applies to both commercial and non-commercial purposes; (2) the Japanese TDM exception applies to any exploitation regardless of the rightholders reservations; (3) exploitation by any means is permitted; and (4) no lawful access is required.[116]

3. The Japanese 'Non-Enjoyment' Purposes

Article 30-4 of the new Japanese Copyright Law has received much international acclaim for the introduction of its 'non-enjoyment' use of TDM.[117] The difference between 'enjoyment' (享受) and 'non-enjoyment' (不見転) under Article 30-4 is the key to understand why the Japanese government has introduced a broad definition of TDM.[118] The Japanese Copyright Office defines the act of 'enjoy,' in this context as, "to accept and appreciate highly emotional things or physical interest, etc."[119] Moreover, the notion of the 'Enjoying the Thoughts or Sentiments Expressed in a Work' under Article 30-4, in the case of literary and artistic works means "enjoying the expression of a work by appreciating it through human senses,"[120] and in the case of a computer program works as "enjoying the function of a computer program work by executing it."[121] Overall, whether an act is considered 'enjoying the ideas or emotions expressed in a work' will be determined by whether the act aims to satisfy the viewer's intellectual or emotional desire.[122]

The question arises as to why TDM exploitation not for 'enjoyment' purposes should be permitted in Japan. This is because copyrighted work satisfies an intellectual or emotional need through enjoyment (e.g. listening to music, looking at a picture, watching a movie, reading a novel or executing a computer program). The purpose of copyright law is to ensure that an author can receive compensation directly or indirectly from those who want to enjoy the work. There is no need for copyright protection if the exploitation of work is aimed at neither enjoying

---

[115] *Ibid*. See also, for further reference, T Ueno, 'Rethinking the Provisions on Limitations of Rights in the Japanese Copyright Act: Toward the Japanese Style "Fair Use" Clause, 07/2009. (Journal of the Japanese Group of AIPPI, 2009) at 159; T Ueno, 'Jinkochino to Kikaigakushu womeguru Chosakukenhojo no Kadai (Copyright Issues on Artificial Intelligence (AI) and Machine Learning, 91(8) Hôristujihô, 2019); *See also,* regarding General Clause on Copyright Limitation (in Japanese),「権利制限の一般規定 ――受け皿規定の意義と課題――」中山信弘・金子敏哉編『しなやかな著作権制度に向けて―コンテンツと著作権法の役割―』（信山社、2017年）141 ~ 182 頁

[116] *Ibid*, at 149.

[117] C Geiger, G Frosio and O Bulayenko, *supra* note 48, (2019); P Kollár, *supra* note 17; K Christensen, *supra* note 11.

[118] T Ueno, *supra* note 111, (2021), at 148.

[119] Japan Copyright Office (JPO), 'Outline of the Amendments to the Copyright Act in 2018.' (4 Patents & Licensing 10, 2019). *See also, ibid,* at 150, note 57.

[120] T Ueno, *supra* note 111, (2021), at 150, note 57.

[121] *Ibid.*

[122] *Ibid.*

it nor causing another person to enjoy it (for example TDM).[123] As Ueno argues, "exploitation of this kind does not prejudice the copyright holder's interests protected by copyright law."[124]

V. Possible Implications of the Japanese TDM Exceptions to the EU Member States

1. Revisiting the Concept of '*Freier Werkgenuss'* under the German Copyright Act

Hugenholtz and Senftleben argue "the need for having more openness in copyright law is almost self-evident in this 'information society' of highly dynamic and unpredictable change."[125] To promote innovation in the advancement of GenAI models and AI in general, EU Member States could consider the Japanese 'non-enjoyment' purposes as an alternative to providing a flexible, but not completely open (i.e., fair-use-like) provision, as this concept will suit the codification-focused EU civil law tradition.[126] One may wonder whether the Japanese 'non-enjoyment' purposes are suitable to be applied in the EU given the different copyright legal systems.

The principle of 'enjoyment' in copyright, on the other hand, is not novel to the EU. The first reference to the 'non-enjoyment' concept can be found in the German Federal Court of Justice or Bundesgerichtshof (hereinafter BGH) judgement of 4.10.1990 - I ZR 139/89,[127] where the defendant used and exploited the system software of the plaintiff in the context of resale in an inadmissible manner; the BGH notes that the pure use, in contrast to the technical rights of use, is not covered by copyright. The use of work as such is not a copyright-relevant process. *This applies to using a computer program as well as reading a book, listening to a piece of music, seeing a work of visual art or watching a movie.*[128] This case illustrates that the BGH employs the notion of 'enjoyment' of the work or well known in Germany as *'Freier Werkgenuss.'* Moreover, in the case of *G. Radio-Werke GmbH., F. i.Bay., v. GEMA*, the BGH rules the following:[129]

> The object of protection of copyright is an intangible good, which, according to its intended purpose, generally serves primarily the intellectual or aesthetic enjoyment of the individual, which by its very nature takes place in the purely private sphere in the case of many intellectual works.

Similar reasoning can be found in this case where the BGH judgement rightly states that the film should be watched and the individual enjoyment of 'watching the movie' is the starting point. Both cases acknowledge that copyright is concerned with intellectual or aesthetic

---

[123] *Ibid*, at 150-151.

[124] *Ibid*, at 151.

[125] PB Hugenholtz and MRF Senftleben, 'Fair Use in Europe. In Search of Flexibilities' (SSRN Publication, 2011) at 29.

[126] *See also*, *ibid*, at 29.

[127] BGH: BGH 04.10.1990 I ZR 139/89 "Operating System" (GRUR 1991, 449) 453.

[128] Raue, '*Das subjektive Vervielfältigungsrecht – eine Lösung für den digitalen Werkgenuss?*', ZGE 2017.

[129] Unofficial translation. BGH 18. Mai 1955 I ZR 8/54 (GRUR 1955, 492).

enjoyment and the exploitation of rights is related to acts that lead to or allow an enjoyment, akin to the Japanese 'enjoyment' concept, in which enjoyment is what gives the ultimate inner justification of copyright protection.[130]

Moreover, in the case of *Grundig-Reporter*,[131] the German court established that the 'non-enjoyment' should not be subject to copyright exclusivity, where "enabling the satisfaction of intellectual needs is what exploitation rights entailed in copyright are concerned with*; if an action does not precede or enable this intellectual satisfaction, it is irrelevant for copyright.*"[132] This case was years before TDM processes were first introduced. However, the notion of 'enable intellectual satisfaction' could be applicable to TDM as well, as TDM does not allow individuals to enjoy someone else's work. Schack echoed this statement by adding:[133]

> If one realizes that the TDM only uses the simple data, but not the intellectual content of the analysed works, then this analysis method does not even interfere with the scope of copyright protection. Technically, there is a reproduction, it does not convey any enjoyment of the work here, and TDM does not trigger a statutory claim to remuneration.

It is worth mentioning that in the BGH judgement of 29 April 2010, I ZR 69/08,[134] the court stipulates that small thumbnails containing copyright-protected works be considered to enable the enjoyment of a work. This is "because the thumbnails are the works concerned of the plaintiff in full, they do not merely represent a public notification or description of their content as of 12 sec. 2 UrhG, more they already enable the enjoyment of the work."[135] In this case, one might argue that no matter how small the copyrighted works presented in the result of TDM, especially in the development of GenAI models, will be categorized as copyright infringement. This article argues that establishing how large and small the copyrighted works influenced the result of creative AI-assisted output, during the TDM processes, is a difficult threshold to meet. Again, as mentioned several times in the previous section, the aim of TDM in the development of GenAI is to find a new pattern from a work. This article suggests putting it this way: "only

---

[130] T Ueno, *supra* note 111, (2021), at 151. *See*, P Kollár, *supra* note 17, at 25. *See also*, BGH 18. Mai 1955 I ZR 8/54 (GRUR 1955, 496).

[131] BGHZ 17, 266/278 = GRUR 1955, 492/496 – Grundig-Reporter.

[132] *See*, Joachim v. Ungern-Sternberg in G Schricker and U Loewenheim, Urheberrecht (6tn edn, CH Beck 2020); Kollár, Péter, *ibid*, at 25. *See also*, for further reference, Leistner, Von '*Grundig-Reporter(n) zu Paperboy(s)*'-*Entwicklungsperspektiven der Verantwortlichkeit im Urheberrecht* (GRUR 2006, 801-814); Lauber-Rönsberg: Autonome „Schöpfung" – Urheberschaft und Schutzfähigkeit (GRUR 2019, 244); BGH, judgment of 18.9.2014 – I ZR 76/13 (OLG Nürnberg).

[133] H Schack, *'Schutzgegenstand, 'Ausnahmen oder Beschränkungen' des Urheberrechts'* (GRUR 2021, 904) at 907. *See also*, for further reference, *Raue/Schöch*, '*Recht und Zugang'* (2020), 118 (124).

[134] BGH, 29 April 2010, I ZR 69/08.

[135] Unoffcial Translation. BGH, 29 April 2010, I ZR 69/08. Available at: <http://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&sid=3a06e12486c979a701f52a24fa4e83cd&nr=51998&pos=0&anz=1> accessed January 10, 2023.

in such cases where truly is no enjoyment, no matter how little, would the non-enjoyment exception be applicable".[136]

A flexible TDM exception, built upon the '*Freier Werkgenuss*' and inspired by the Japanese-style exception, might expressly state the legality of non-enjoyment uses, providing legal assurance for GenAI models researchers and developers in the EU.[137] In France, Article L.122-3 CPI defines reproduction right as "the material fixation of the work by all means that enable to communicate it to the public in an indirect way."[138] We can see here that the French definition of the right of reproduction presupposes communication to the public and the formulation of the article appears to reflect a sense of enjoying or causing another to enjoy as a prerequisite to copyright infringement.[139] In comparison to the fair-use type, the Japanese 'non-enjoyment' doctrine appears to be closer to the European continental view of copyright.[140]

2. The 'Non-Enjoyment' Purposes as an 'Opening Clause' to Drive Innovation

The introduction of a flexible TDM exception, flexible but not too open as the US fair use doctrine, could be effective to boost innovation and the competitiveness of GenAI models in the EU.[141] This is because "an enumerated list of exceptions and limitations has shown little flexibility in adapting to evolving market and technological conditions."[142] The 'non-enjoyment' approach could be the most logical basis for a flexible TDM in the EU because the Japanese exception allows exploitation in any case as long as the purpose is not to cause another person to enjoy the work.[143] This provision thus resembles an 'opening clause,' which "should address uses that are not yet covered by existing exceptions and limitations but are justified by important public interest rationales and fundamental rights such as freedom of expression and the right to information."[144]

Furthermore, the Japanese "non-enjoyment" purposes, which specifically list TDM as one of the permissible uses, offer more legal clarity than the fair use approach does.[145] It is believed

---

[136] P Kollár, *supra* note 17, at 25.

[137] *Ibid*, at 25.

[138] S Dussolier, 'Realigning Economic Rights with Exploitation of Works: The Control of Authors over the Circulation of Works in the Public Sphere', at 166-168 in PB Hugenholtz (ed), 'Copyright Reconstructed: Rethinking Copyright's Economic Rights in a Time of Highly Dynamic Technological and Economic Change' (2018) Kluwer Law International.

[139] P Kollár, *ibid.*

[140] *Ibid*, citing M Senftleben, *et al*, 'Ensuring the Visibility and Accessibility of European Creative Content on the World Market: The Need for Copyright Data Improvement in the Light of New Technologies.' (2021) at 7.

[141] PB Hugenholtz and MRF Senftleben, *supra* note 125, at 29.

[142] C Geiger, G Frosio and O Bulayenko, *supra* note 48, (2019), at 22.

[143] Article 30-4 of the Japanese Copyright Law.

[144] C Geiger, G Frosio and O Bulayenko, *supra* note 48.

[145] P Kollár, *ibid*, at 26.

that the EU would provide a favorable environment for the development of AI-driven creativity with the implementation of the "non-enjoyment" clause.

VI. The 'Non-Enjoyment' Purposes and Three-Step Test: Oh Yes, Test Passed!

Any Berne Convention,[146] WIPO Copyright Treaty (WCY) and TRIPs[147] contracting parties seeking to introduce a new copyright exception should comply with the so-called 'Three-Step' Tests. This test consists of three cumulative conditions in which Article 13 of TRIPs provides that Member States shall confine exceptions and limitations (1) to "certain special cases," (2) "which do not conflict with a normal exploitation of the work", and (3) "do not unreasonably prejudice the legitimate interests of the right holder".[148] In general, the Japanese 'non-enjoyment' purposes doctrine will not conflict with this test, as Ueno argues, "because an exploitation not for 'enjoyment' purposes does not prejudice the opportunities of the copyright holders to receive compensation."[149] TDM may be permitted if an open clause that permits non-enjoyment exploitation of works protected by copyright were included.[150] This section briefly assessed the compatibility of the 'non-enjoyment' doctrine with the three-step test. However, a complete analysis will be required in the future, as well as a factual macro-economic evaluation, which this article is lacking.

1. The First Step: Is TDM allowing 'non-enjoyment' exploitation of copyrighted works a special case?

An open clause allowing non-enjoyment exploitation of copyrighted works passed this test considering that the TDM exception would be '*certain* and *special* cases.' In this context, it is important to differentiate between the two adjectives as the WTO dispute settlement panel[151] considered the word "certain" to indicate that any exception or limitation should be presented in a limited manner and clearly defined,[152] and the term "special" should be understood as the exception or limitation should be narrow in "its scope and reach".[153] Building upon the German doctrine of '*Freier Werkgenuss'*, when the TDM exception allows exploitation of a work which is aimed at neither enjoying nor causing another person to enjoy it should be a special case and

---

[146] Berne Convention for the protection of literary, artistic works, 1886 (Paris Text 1971) (BC)

[147] Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPs), WTO Annex IC, adopted in Marrakesh, 15 April 1994.

[148] Article 13 TRIPs.

[149] T Ueno, *supra* note 111, (2021), at 150.

[150] P Kollár, *ibid*, at 27.

[151] Report of the Panel, United States – Section 110(5) of the US Copyright Act, para. 6.62, WT/DS160/R (June 15, 2000) (hereinafter Report of the Panel, United States – Section 110(5)).

[152] PB Hugenholtz and R Okediji, 'Conceiving an International Instrument on Limitations and Exceptions to Copyright' (Study supported by the Open Society Institute (OSI), March 6, 2008, Amsterdam Law School Research Paper No. 2012-43, 2012). Available at: <https://ssrn.com/abstract=2017629> accessed January 10, 2023 at 22.

[153] Report of the Panel, United States – Section (110)(5), *ibid*, para. 6.112.

"narrow in quantitative as well as a qualitative sense."[154] The 'non-enjoyment' TDM is narrow in scope as it only applies to several ranges of activities.[155] Additionally, by the quantitative criteria, 'non-enjoyment' is constrained in favour of public interest targets, to achieve a balance between copyright and public interests including education, AI research, information transparency, and the right of free expression.[156] The first step is passed.

2. The Second Step: Does TDM allowing 'non-enjoyment' exploitation of copyrighted works conflict with the normal exploitation of the work?

One should realize that TDM is different from other aspects of copyright such as adaptation, in the case of GenAI model and as previously discussed, the TDM process does not allow anyone to enjoy the fruits of intellectual labor, but even if it were, it would not conflict with the normal exploitation of the work in the way rightholders usually exploit their work.[157] TDM build upon the *'Freier Werkgenuss'* concept unproblematically passed the second step as the WTO Panel provides:[158]

> An exception or limitation to an exclusive right […] rises to the level of a conflict with a normal exploitation of the work […] if uses, that in principle are covered by that right, but exempted under the exception or limitation enter into economic competition with the ways that right holders normally extract economic value from that right to the work (i.e., the copyright) and thereby deprive them of significant or tangible commercial gains.

3. The Third Step: Does TDM allowing 'non-enjoyment' exploitation of copyrighted works unreasonably infringe the legitimate interest of the right holder?

Once one considers that TDM and other 'non-enjoyment' purposes should not be relevant for copyright purposes, such exceptions should be permissible under this step. WTO Panel defines a legitimate interest as "relates to lawfulness from a legal positivist perspective, but it has also the connotation of legitimacy from a more normative perspective, in the context of calling for the protection of interests that are justifiable in the light of the objectives that underlie the protection of exclusive rights." As a result, it should not be prohibited to participate in TDM since the goals that underpin protection do not warrant an extension to it. Therefore, TDM exclusivity is not a legitimate interest.[159] Even if it were a legitimate interest, it is unlikely that any copyrighted works used during the TDM process would have independent economic values

---

[154] *Ibid*, at para. 6.145.

[155] P Kollár, *supra* note 17, at 27.

[156] GS Muto, *supra* note 9, at 45, citing S Ricketson, 'The Berne Convention for the Protection of Literary and Artistic Works: 1886-1986' (London 1987), at 535.

[157] P Kollár, *ibid*, at 28.

[158] Report of the Panel, United States – Section (110)(5), *supra* note 159, para. 6.183. *See also*, Hugenholtz, P. Bernt and Okediji, Ruth, supra note 157, at 23.

[159] P Kollár, *ibid.*

derived from it,[160] as unreasonable prejudice would occur only if "an exception or limitation causes or has the potential to cause an unreasonable loss income to the copyright owner."[161]

VII. Good Luck, Europe: What the EU Member States Can Do Without Flexible TDM Exceptions?

It should be noted that the EU CDSM Directive recognizes the potential of both scientific and non-scientific TDM, but due to the restrictive scope of the exceptions, it fails to grasp that potential.[162] Following on from the critiques highlighted in the previous section, numerous scholars and stakeholders have proposed plenty of recommendations that, in their opinion, would improve the legal framework of TDM. Some of these suggestions are summarized shortly as follows: First, numerous critics proposed changes to the list of TDM beneficiaries, ranging from removing purpose-specificity to eliminating any difference between commercial and non-commercial research.[163] Second, as previously discussed, the criterion of lawful access has been widely criticized from a diversity of viewpoints. Why would we need any further restriction if one wants to conduct TDM must have lawful access, to begin with?[164] To comply with lawful access requirements and to avoid unnecessary high licensing costs, this article suggests the establishment of a centralised repository of numerous open access data/information comprising literary and artistic works, similar to LAION-5B datasets used by Stability AI, where the data corpus can be collected, maintained, or exchanged between different market players. This might be an alternative option that Member States can do to foster the development of GenAI models.[165]

Third, Geiger *et al* considered fair remuneration as an alternative for the opt-out mechanism and "might have been considered provided that harm can be demonstrated on the basis of relevant empirical data."[166] This article argues that the remuneration mechanism would indeed be improvements over the current system; however, paying remuneration individually to a collective management organisation for copyrighted works used for TDM is nearly impossible for GenAI models researchers or developers who do not have sufficient financial means. This

---

[160] *Ibid*.

[161] Report of the Panel, United States – Section (110)(5), *ibid*, para. 6.183.

[162] EU CDSM Directive, Recital 18.

[163] C Geiger, G Frosio and O Bulayenko, *supra* note 48, (2019), at 22; L Koschwitz, 'The EU Just Told Data Mining Startups to Take Their Business Elsewhere', (Euractiv.com, 2016). Available at: <https://www.euractiv.com/section/digital/opinion/the-eu-just-told-data-mining-startups-to-take-their-business-elsewhere/> accessed August 16, 2022; F Reda, 'Text and Data Mining Limited'. Available at: <https://felixreda.eu/eu-copyright-reform/text-and-data-mining/> accessed August 16, 2022; PB Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4) - Kluwer Copyright Blog'. (2019). Available at: <http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/> accessed February 2, 2023.

[164] P Kollár, *supra* note 17, at 23.

[165] *See*, C Schuhmann *et al*, supra note 36.

[166] C Geiger, G Frosio and O Bulayenko, *ibid*, at 31.

will lead to the same problems as in the digitised music industry, when 'the winner takes all' and only 'rich' GenAI research organisations and developers will be able to survive in the EU.

If there is no political willingness from the EU Member States to adopt broader TDM exceptions and a welcoming environment for the development of GenAI models such as from the Japanese 'non-enjoyment' style, there are several recommendations that the Member States could do. The key, however, lies in the national implementation, given that Article 25 of the EU CDMS Directive allows the EU Member States to adopt or maintain in force broader provisions compliant with the exceptions and limitations provided for in the InfoSoc Directive and Database Directive. In addition to the solutions provided by the aforementioned commentators, this article recommends the following to the EU Member States who do not wish to adopt broader exceptions:

1. **Advocate for 72 Hours Response if Technological Protection Measures (TPMs) Are Preventing TDM**

The relevant rightholders now can block access for every GenAI researcher and developer planning to do TDM, because according to Article 7(2) of the EU CDSM Directive, both Articles 3 and 4 are subject to technical protection measures (TPMs). Article 3(3) of the Directive defines TPMs as "measures to ensure the security and integrity of the networks and databases where the works or subject matter are hosted." However, the existing mechanisms allowing the circumvention of TPMs at the national level are unclear.[167] In the context of TDM, TPMs have the potential to "limit or prevent access to works altogether for purposes that are not restricted by authors' rights or for uses that are actually privileged."[168] When transposing the EU CDSM Directive to the national level, EU Member States could create a consistent, clear and transparent mechanism regarding TPMs and make a clear definition of the term "appropriate measures" as provided by Article 6(4) of the InfoSoc Directive. However, considering that TPM systems have not shown to be effective since the adoption of the InfoSoc Directive,[169] it will most likely not work for TDM under the EU CDSM Directive and there is a possibility that such technical measures will be used to unlawfully limit TDM.[170]

To prevent unlawful TDM limitation by TPMs, during the national implementation phase, EU Member States could consider a certain period, for example, of a maximum of 72 hours, by which time access must be restored if TPMs are preventing access to resources that are lawfully obtained by GenAI models researcherss or developer. It is important to know that any technical issues are usually solved in 72 hours or less.[171] If access is not given within the predetermined

[167] V Banti-Markouti, 'The Interface between Technological Protection Measures and the Exemptions to Copyright under Article 6 Paragraph 4 of the Infosoc Directive and Section 1201 of the Digital Millennium Copyright Act.' (Issues in Informing Science and Information Technology, 2007) at 582.

[168] C Geiger, G Frosio and O Bulayenko, *supra* note 48, (2019), at 35.

[169] R Ducato and A Strowel, *supra* note 75, at 16-17.

[170] K Christensen, *supra* note 11, at 32.

[171] Research organizations and private companies usually spend thousands or millions of euros per year for access to content. Publishers commonly guarantee 24 hour access to material (except while doing maintenance), and a customer response time of 24 hours is typical. B White and MB Jančič, 'Article 3-4: Text and Data Mining'.

time, a suitable financial penalty should be imposed considering the investment made by research organizations, individuals, and others in acquiring the content for TDM purposes.[172]

2. Robot Exclusion Standard (robot.txt) as a Warning When TDM is Not Allowed on a Website

In the case of GenAI models, when materials such as songs, poetry, paintings, etc., are published online on the website, an automatic way of indicating that a website is not qualified for TDM is required. The robot exclusion standard, for example, robot.txt, which has been extensively utilized since the mid-1990s, might be an option.[173] Almost all websites on the planet follow the standard for restricting what can be mined by robot.txt. Search engine platforms implement this standard to serve as machine-readable terms and conditions.[174] What is envisaged by Article 4(3) of the EU CDSM Directive of "machine readable means in the case of content made publicly available online" is the use of a machine-readable robots.txt file to specify access restrictions.[175] The use of robot.txt will strike a fair balance between the interests of rightsholders and GenAI models researchers and developers wishing to perform TDM on publicly accessible websites.[176]

VIII. Conclusions: The Right to Read Should be the Right to Mine

TDM is one of the building blocks of AI and has attracted much public attention from copyright scholars. The EU CDSM Directive does envision a concrete action to promote research and innovation. The foregoing analysis, however, has shown that the full implementation of the TDM exceptions would be critical to European innovation and research, particularly, in the development of AI-driven creativity. This article argues that the Japanese 'non-enjoyment' purpose is one of the best alternatives for the EU Member States to provide a flexible, but not completely open, TDM exception to foster AI innovation at the national level. A comparable concept is the German '*Freier Werkgenuss'* which acknowledges that copyright is concerned with the intellectual enjoyment of work and such TDM activities should not be subject to copyright exclusivity at all because there is no enjoyment of the work. This is similar to the Japanese concept of 'non-enjoyment' purposes. A TDM exception built upon the German concept of *'Freier Werkgen*uss' could be the opening clause to a flexible, but not too open TDM exception, and offer specific lists of TDM activities as one of the permissible uses.

Once one realizes that TDM processes in the development of AI and machine learning in general, are "copy works not to consume the expression of copyright law protects, but to get

Available at: https://www.notion.so/Articles-3-4-Text-and-data-mining-9be17090ebc545b88ed9ac7d39e4e25a> accessed February 2, 2023.

[172] *Ibid.*

[173] *Ibid.*

[174] Google, Bing, Baidu, DuckDuckGo, Yahoo!, and Yandex are among the search engines. *Ibid.*

[175] *Ibid.*

[176] *Ibid*

access to the facts or structures copyright law dedicates to the public",[177] an opening clause allowing TDM with 'non-enjoyment' purposes could be permissible under the copyright three-step test. If there is no political will amongst the EU Member States to support the development of AI by providing a broader TDM exception, during the national implementation phase, Member States can define carefully the terminology of 'appropriate manner' as per the reservation made in Recital 18 of the EU CDSM Directive. Moreover, to avoid the negative consequences of the lawful requirements, at the national level, Member States could establish a centralized database containing numerous open access to creative works such as texts, poetry, images, songs, etc. With this database, GenAI models researchers and developers can train their AI systems with open source data corpora.

An advocacy of response if TPMs are preventing AI researchers and developers from conducting TDM should be established, considering that both Articles 3 and 4 of the EU CDSM Directive are subject to TPMs. Considering that the current circumvention of TPMs at the national level is unclear, this is the opportunity for the EU Member States to advocate TPMs for TDM. Finally, what is clear from this article is that the EU Member States must be mindful of the future of AI innovation, particularly the development of GenAI, which is dependent on TDM. In a closing remark from the Founder of a UK-based data and analytics company:[178]

> To not have the freedom to access information without infringing on IPRs data science and machine learning would be detrimental to our business and quite frankly stop, or make innovation extremely hard, thus affecting the European tech and start-up economy as a whole.

I am sure that the EU Member States can do something to prevent this from happening. Everything is now in the hands of the EU Member States, whether to protect the interests of rightholders or to create a balance between safeguarding 'the right to read should be the right to mine', protecting rightholders exclusivity, and creating a supportive environment for the GenAI researcher and developers.

***

Acknowledgements

The author's sincere appreciation goes to the anonymous reviewers of the ATRIP Annual Essay Competition 2022. The author also would like to thank the following: Prof. Rosa Ballardini, Prof. Péter Mezei, Prof. Tatsuhiro Ueno, Prof. Jean-Marc Deltorn, Prof. Anne Lauber-Rönsberg, Prof. Ana Ramalho, Dr. David Linke, Dr. Sven Hetmank, Dr. Gabriele Spina Ali, Dr. Daria Kim, Ansgar Kaiser, Natasha Mangal, Ryoko Oshikamo, Alexandre Drouet, Miriam Steinhart, Ana Andrijevic and all of the participants of the Research Atelier AI and IP at CEIPI, the University of Strasbourg (France), and IRGET, TU Dresden (Germany) for the wonderful discussion and their endless support during the writing process.

---

[177] Lemley, Mark A. and Casey, Bryan, 'Fair Learning' (2020), at 785, Available at SSRN: https://ssrn.com/abstract=3528447 or http://dx.doi.org/10.2139/ssrn.3528447

[178] Maryam Mazraei, Founder of Autopsy, interview of April 5, 2019. Available at: <https://www.getautopsy.com/> accessed August 23, 2022. Found in C Gerrish and AM Skavlan, *supra* note 49, at 67.