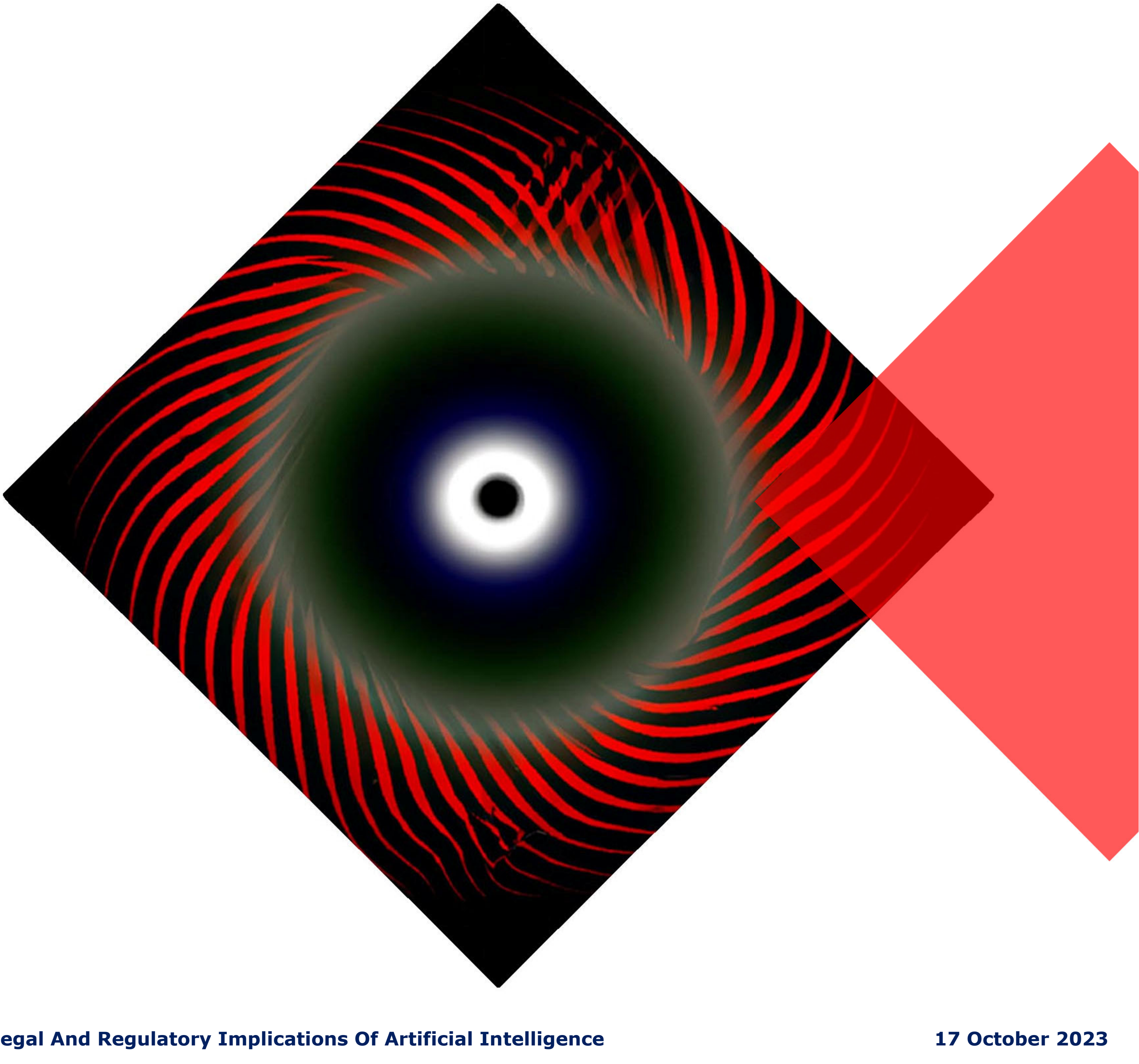
AI-REGULATION.COM

AI-Regulation Papers 23-10-2

EU Copyright Directive: A ‘Nightmare’ for Generative AI Researchers and Developers ?

By Dr. Theodoros KARATHANASIS



Chair Legal And Regulatory Implications Of Artificial Intelligence **17 October 2023**

Dr. Theodoros Karathanasis holds a PhD in European law from the Faculty of Law of the University of Grenoble Alpes. He is attached to the Centre for International Security and European Cooperation Studies (CESICE) and his research work in the field of cybersecurity has been funded by the Grenoble Alpes Institute of Cybersecurity. He is a member of the cyber experts network of the European Centre of Excellence for Combating Hybrid Threats (HybridCoE), as well as the EU CyberNet expert network. He is a Research Fellow at the AI-Regulation.com Chair.

Cover Photo : Created by **DALL-E,** prompt by **Theodore Christakis : "***Generate an image that resembles a circular painting and conveys the fusion of AI and art, with futuristic elements and a forward-looking aesthetic***"**

To cite this article: T. Karathanasis, EU Copyright Directive: A 'Nightmare' for Generative AI Researchers and Developers? AI Regulation Papers 23-10-2, AI-Regulation.com, October 17th, 2023.

These statements are attributable only to the author, and their publication here does not necessarily reflect the view of the other members of the AI-Regulation Chair or any partner organizations.

This work has been partially supported by MIAI @ Grenoble Alpes, (ANR-19-P3IA-0003) and by the Interdisciplinary Project on Privacy (IPoP) of the Cybersecurity PEPR (ANR 22-PECY-0002 IPOP).

EU Copyright Directive: A ‘Nightmare’ for Generative AI Researchers and Developers ?

Drawing on intense criticism from online publishers across the European Union (EU) against Generative AI (GAI), the present article aims to highlight the highly debated copyright issue of data collection for Generative AI training. Three questions are therefore addressed: To what extent is scraping data for GAI training considered to be a copyright issue; How Data scraping and data mining are regulated under EU Law and; How the future AI Act intends to deal with the use of training data.

On September 7th, 2023, Microsoft announced the Microsoft Copilot Copyright Commitment according to which Microsoft will assume legal responsibility for their customers should they be sued for copyright infringement whilst using the company’s Copilot AI services, provided “the customer uses the guardrails and content filters integrated into the AI products”.

In response to the growing concerns about Generative AI and copyright infringement, the European Publishers Council (EPC) released on 6 September 2023 Global Principles for Artificial Intelligence (AI), which is aimed “at ensuring publishers’ continued ability to create and disseminate quality content, while facilitating innovation and the responsible development of trustworthy AI systems”. This release was followed eight days later by a position paper by the French Group of online service publishers (GESTE), which highlights the necessity of “*setting up licences within a negotiated framework*”, in order to protect authors against Generative AI (GAI) data and text mining (DTM). While the rightsholders are calling for the inclusion of transparency provisions in relation to copyright in future European Regulation on Artificial Intelligence (the AI Act), the only solution for the time being seems to be the use of the right to ‘opt-out’ from DTM that the EU Copyright in the Digital Single Market (CDSM) Directive provides.

This summer, amid the ongoing trilogues between the Council of the EU (Council), the European Parliament (EP) and the European Commission (Commission), convened to reach consensus on the AI Act, the French media (e.g., France Médias Monde, TF1, Les Echos) decided to put the brakes on OpenAI data mining. Despite the fact that the opt-out mechanism was supposed under the CDSM Directive to enable publishers to retain control over the use of their content, questions still remain as to the extent to which opting out may be respected in the case of GAI; as well as to the extent to which the EP proposal on the transparency requirements for GAIs under the AI Act may bring in additional safeguards concerning copyright infringements.

Scraping data for Generative AI training: A Much-Debated Copyright Issue

Generative AI took the world by storm in the months after ChatGPT, a chatbot based on OpenAI’s GPT-3.5 neural network model, was released on November 30, 2022. GAI is a relatively new form of AI that, unlike its predecessors, can create new content by extrapolating it from its training data. Its extraordinary ability to produce human-like writing, images, audio, and video have captured the world’s imagination since the first generative AI consumer chatbot was released to the public. Providing such outputs involves obtaining millions of people’s information from the internet in order to train GAIs. It's no secret that AI training involves the use of publicly available data, including texts, images, videos, and other content.

It has been widely reported in the media that Google updated its privacy policy on July 1st, 2023, “to allow the company to collect and analyse information people share online to train its AI models”.

Indeed, in order to provide, maintain, improve and develop Google products, services, and machine learning technologies, publicly accessible sources and Bard user information are processed to the extent it is necessary for the legitimate interests of Google. Meta AI also uses several sources to train its Generative AI models, such as publicly available online and licensed information, as well as information from Meta products and services. OpenAI's ChatGPT is also being developed using, among other things, information that is publicly available on the internet. Midjourney's privacy policy also states that "data collected from third party sources. (…) may include, but not be limited to: public databases, commercial data sources, and the public internet." Collecting publicly available data from the internet is also known as data scraping.

Data scraping automatically gathers data from online resources such as websites, databases, APIs, and documents. Information may be present in online resources in a structured, semi-structured, or unstructured format. Data scraping aims to parse through these data and transform them into a structured format for further processing, analysis or storage. For example, OpenAI has explained that scraping data from websites "can help AI models become more accurate and improve their general capabilities and safety." The concerns that are being expressed today about data scraped from the internet to train AI systems are similar to those expressed by website owners in the past about search engine web crawlers[1]. "Just as search engine companies need to scrape data to provide accurate and up-to-date search results, so too do AI companies need to scrape data to train their AI systems"[2].

Web scraping is legal in the United States, but there is a risk that policymakers could decide to intervene. Indeed, "data protection regulators from a dozen countries—including Australia, Canada, Mexico, China, and the UK—recently published an open letter to website operators urging them to implement measures to protect against "unlawful data scraping". With twitter's data being openly scraped to train AI models, Elon Musk also announced that the platform will have "rate limits" (or limits on how many tweets a user can look at each day) in order to "address extreme levels of data scraping & system manipulation". The new terms of X (formerly Twitter), which was introduced on 29 September, states that " (…) crawling [systematic indexing] and scraping [extraction for exploitation purposes] of the Services, in any form and for whatever purpose is expressly prohibited without our prior written consent".

Due to the involvement of large databases, especially those of a public nature, intellectual property issues may arise. Generative AI systems may be trained on copyrighted data and produce content that infringes upon the intellectual property rights of others, posing legal and ethical challenges in terms of ownership and attribution. However, the act of scraping data does not inherently infringe on copyright, as copyright law protects original works and not the raw data itself. That said, the way this data is used could potentially infringe on copyright. With the help of ChatGPT Plus and Google's Bard, Neville Hobson researched the topic and went on to state that:

"The use of generative AI for transformative purposes is likely to fall within the fair use exception to copyright law. This means that you can scrape data from the Internet and use it to train a generative AI model without infringing copyright. However, it is important to note that fair use [and fair dealing] is a complex area of law, and there is no guarantee that a court will find a particular use to be fair. If you are unsure whether a particular use is fair, it is always best to consult with an attorney."

In January 2023, the artists Sarah Andersen, Kelly McKernan and Karla Ortiz filed a class-action lawsuit

[1] The distinction between scraping and crawling therefore needs to be clarified. Scraping is about extracting data from one or more websites while crawling is about finding or discovering URLs or links on the web. See K. Moaiad (2021) 'Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application', International Journal of Advances in Soft Computing and its Applications 13(3):145-168.

[2] Stevens, M. and D. Castro, (2023) 'In the Wake of Generative AI, Industry-Led Standards for Data Scraping Are a Must', Center for Data Innovation. Available at https://datainnovation.org/2023/09/in-the-wake-of-generative-ai-industry-led-standards-for-data-scraping-are-a-must/#:~:text=Just%20as%20search%20engine%20companies%20need%20to%20scrape,accurate%20and%20improve%20their%20general%20capabilities%20and%20safety.%E2%80%9D

against Stability AI, Midjourney, and DevianArt,[3] accusing them of committing mass copyright infringement by "scraping the internet" to copy and store billions of copyrighted images without obtaining consent or licenses from artists, and then using the copied images as inputs to train their AI platforms, without the artists' knowledge or consent. During the hearing held on July 19th in relation to *Andersen et al v. Stability AI Ltd. Et al.*, the Court dismissed the claims of the plaintiffs by ruling that they failed to present the facts clearly and demonstrate how each defendant could be held liable for copyright infringement.

In February 2023, Getty images, a global digital media provider and supplier of stock images, editorial photography, video, and music content, filed a lawsuit against Stability AI at the U.S. District Court of Delaware[4] claiming that their AI art tool had copied and processed 12 million images and associated text and metadata in order to train their AI model, without obtaining a license to do so. In May, a second lawsuit was filed at London's High Court of Justice, to prevent Stability from selling its AI-image generator tool in the UK.[5]

On June 28th, 2023, a class action was filed by two authors on behalf of themselves and other parties in relation to the class action complaint[6] against OpenAI Inc., at the U.S. District Court for the Northern District of California, claiming that they never authorised OpenAI to copy their books, make derivative works, publicly display copies (or derivative works), or distribute copies (or derivative works). Copyright lawsuits that are currently underway in the United States have substantial implications for the future of generative AI systems. The results of such class actions will be very interesting, since they go to the heart of certain criticism targeted at the Generative AI business model. The following are all of the US lawsuits that are taking place in relation to OpenAI and ChatGPT concerning copyright infringement in the US:

- Authors Guild et al v. OpenAI Inc. et al - September 19, 2023[7]
- Chabon v. OpenAI, Inc. - September 8, 2023[8]
- Walters v. OpenAI LLC - July 14, 2023[9]
- Silverman, et al v. OpenAI Inc. - July 7, 2023[10]
- Tremblay v. OpenAI Inc. - June 28, 2023[11]
- Getty Images (US), Inc. v. Stability AI, Inc. – October 7, 2023

Whatever the outcome of these cases, they certainly represent major precedents not only in terms of the solution of other legal cases but also in terms of how lawmakers are likely to regulate the use of copyright by AI. The only solution for the time being seems to be the use of the opt-out.

On July 6th, 2023, a spokesperson pointed to a recent blog post by Google in which the company said it wanted a discussion around creating a community-developed web standard similar to the robots.txt system that allows publishers to opt out of parts of their sites being crawled by search engines. Google's comments come as news companies such as News Corp have already reportedly been initiating conversations with AI companies about payment for scraping news articles. Publishers should be able to opt out of having their works mined by generative artificial intelligence systems, according to Google, but the company has not said how such a system would work.

On the other side of the Atlantic, Article 4 of the EU CDSM Directive somewhat explicitly provides that the use of copyrighted content for text and data mining (TDM), including content used for AI training, is permissible and merely gives rightsholders the

[3] Andersen et al. v. Stability AI Ltd. et al., case no. 3:23-cv-00201, U.S. District Court for the Northern District of California.

[4] Getty Images (US), Inc. v. Stability AI, Inc., case no. 1:23-CV-00135, U.S. District Court District of Delaware.

[5] Getty Images (US) Inc. and others v. Stability AI Ltd., case no. IL-2023-000007, High Court of Justice of England and Wales.

[6] Tremblay v. OpenAI Inc., Case no. 4:2023-cv-03223, U.S. District Court for the Northern District of California.

[7] Authors Guild et al v. OpenAI Inc., Case no. 1:23-cv-8292, U.S. District Court Southern District Of New York

[8] Chabon et al v. OpenAI, Inc. et al., Case no. 3:2023cv04625, US District Court for the Northern District of California

[9] Walters v. OpenAI LLC, Case no. 1:23-cv-03122, US District Court for the Northern District of Georgia

[10] Silverman, et al v. OpenAI Inc., Case no. 3:23-cv-03416, US District Court for the Northern District of California

[11] Tremblay v. OpenAI Inc., Case no. 4:2023-cv-03223, U.S. District Court for the Northern District of California

option to reserve the right for their works to be used in such a way.

EU Copyright in the Digital Single Market Directive, Data Scrapping *versus* Data Mining

Data scraping and text/data mining are both techniques used to extract information from digital sources, but they serve different purposes and can have different legal implications under EU law, particularly with regard to data protection and copyright regulations.

> ***Data scraping** involves extracting data from websites or other digital sources using automated tools or scripts. This data can be structured or unstructured and may include text, images, or other forms of content. Therefore, scrapers work by parsing the HTML source code of a website in order to extract and retrieve specific elements within the page's code.*

Mass data scraping of personal information can constitute a reportable data breach in many jurisdictions. Under EU law, data scraping may be subject to the General Data Protection Regulation (GDPR)[12] if the data being scraped contains personal information. In such cases, the data scraper may need to comply with GDPR requirements, such as obtaining user consent or ensuring that the data is processed lawfully and securely.

In March 2020, the Polish Data Protection Authority (DAP) issued its first fine under the GDPR against Bisnode, a Swedish-headquartered company that specialises in business intelligence and data analytics. Apparently, "Bisnode had scraped data from publicly available government databases about individuals' prior registrations as sole proprietors and other related corporate activities and produced certain reports for its clients"[13]. The Polish DPA issued a fine in response to this violation. Instead of complying with their request to mail out millions of notices, Bisnode reportedly stated it would delete the data involved and appeal against the Polish DPA's order.

Data scraping may be subject to copyright law if it involves copying and using copyrighted content without permission. EU copyright law protects the rights of content creators, and scraping copyrighted material without authorisation can lead to copyright infringement claims. Contrary to data scraping, TDM is considered a means of research. It forms one of the techniques used for collecting information from an indefinite number of digital data ('Big Data'), which focuses on particular words, themes or subject matter with the help of an automated tool. Indeed, under EU law, there is an exception to copyright law that allows for TDM for research purposes.

> ***Article 2 of the CDSM Directive** provides that '"text and data mining" means any automated analytical technique aimed at analyzing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations'.*

According to Article 3 of the EU CDSM Directive, the "reproductions and extractions made by research organizations[14] and cultural heritage institutions" can only be carried out for the purpose of scientific research and they must have lawful access to the works or subject matter in question.[15] In other words, "once a copyright work has been legitimately accessed, the right to read should be the right to mine when it comes to research and machine

[12] European Parliament and the Council, Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016, p. 1–88, ELI: http://data.europa.eu/eli/reg/2016/679/oj

[13] Martinier, S., Pépin, M., Neuburger J. & J. Mollod, (2020) 'French DPA Issues Guidance Surrounding Practice of Web Scraping', Proskauer.

[14] According to recital 11, startups operating in the digital environment, which are the source of important innovations, particularly in the field of artificial intelligence, are not taken into consideration, and therefore their data mining activities remain subject to the exclusive right.

[15] The GDPR may also apply to TDM if the data being mined contains personal information. Researchers must ensure compliance with GDPR requirements when using personal data for TDM.

learning"[16]. According to Recital 14, 'lawful access' covers access to content pursuant to contractual arrangements (e.g. subscriptions or open access licenses), as well as to "content that is freely available online". The collection and storage of data for the purpose of text and data mining of works protected by copyright requires that the user grants lawful access, and this can be done without obtaining the prior authorisation of the copyright owners. However, the requirement of 'lawful access' does not imply that rightsholders may contractually rule out text and data mining in their terms of agreement.

Article 4 of the EU CDSM Directive provides an exception for reproduction and extraction of lawfully accessible works, irrespective of whether or not they are for commercial gain. But the most attention-grabbing point is Article 4(3), which allows the relevant rightsholders to reserve the right to perform TDM activities. As things stand, such a right may be reserved, as mentioned in Recital 18, in an "appropriate manner", such as via machine-readable means. In other cases, it may be appropriate to reserve the right by other means, such as contractual agreements or a unilateral declaration.

In France, for example, the CDSM Directive was transposed into French law by means of ordinances. Ordinance No. 2021-1518 of 24 November 2021, which completed this transposition, introduced this option in order for rightsholders to expressly object to text and data mining in Article L. 122-5-3 of the French Code of Intellectual Property. This article states that "Without prejudice to the provisions of II, digital copies or reproductions of lawfully accessed works may be made for the purpose of text and data searches carried out by any person, regardless of the purpose of the search, unless the author has objected in an appropriate manner, in particular by machine-readable processes for content made available to the public online". Decree no. 2022-928 of 23 June 2022 specifies that this "opt-out" does not have to be justified and may be expressed by any means (specifying for content placed online: "by means of machine-readable processes, including metadata, and by recourse to the general terms and conditions of use of a website or service").

Rightsholders will therefore only be allowed to reserve the right to use TDM for content that is publicly available online if they implement appropriate technological measures. In line with the analogy drawn by the Court of Justice of the European Union (CJEU) in the *VG Bild-Kunst* case,[17] such technological measures should be understood as follows: "the copyright holder cannot be allowed to limit his or her consent by means other than effective technological measures". The application of an access control or protection process, such as encryption, scrambling or other means of altering the work or other subject matter or copy control mechanism, are considered by the CJEU to constitute technological measures.

In its willingness to provide its members with the tools they need to exercise their right to "opt out", the French Publishing Union – Le Syndicat national de l'édition (SNE) - is proposing that a new model clause be included in publishers' websites' terms of use or, failing that, in their legal notices. Publishers wishing to express their desire to "opt out" thus have an intermediary solution that can be implemented immediately. In addition, the SNE recommends using the technical tool proposed by EDRLab, which enables this opt-out to be exercised through the use of metadata (TDM Reservation Protocol (TDMRep) (w3.org). The use of this metadata, designed to fall into the category of

[16] WIPO, WIPO Conversation O
n Intellectual Property (Ip) And Artificial Intelligence (AI), Third Session, Geneva, 4 November 2020. Available at https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_3_ge_20/wipo_ip_ai_3_ge_20_inf_5.pdf

[17] CJEU, Judgement of the Court (Grand Chamber) of 9 March 2021. VG Bild-Kunst v Stiftung Preußischer Kulturbesitz. Request for a preliminary ruling from the Bundesgerichtshof. Reference for a preliminary ruling – Intellectual property – Copyright and related rights in the information society – Directive 2001/29/EC – Article 3(1) – Concept of 'communication to the public' – Embedding, in a third party's website, of a copyright-protected work by means of the process of framing – Work freely accessible with the authorisation of the copyright holder on the licensee's website – Clause in the exploitation agreement requiring the licensee to introduce effective technological measures against framing – Lawfulness – Fundamental rights – – Article 11 and Article 17(2) of the Charter of Fundamental Rights of the European Union. Case C-392/19. Court reports – general, ECLI identifier: ECLI:EU:C:2021:181.

machine-readable processes, is an effective technical addition to the tools for harvesting data.

The reservation of rights or 'opt-out' mechanism of the EU CSDM Directive might hamper the advancement of AI in the EU. At a time when the EU is trying to adopt the first comprehensive legal framework on AI with its AI Act, the provisions of the CDSM Directive instead paradoxically favour "the development of biased AI systems due to price and accessibility conditions for training data that offer the wrong incentives. To avoid licensing, it may be economically attractive for developers to train their algorithms on older, less accurate, biased data, or import AI models already trained on unverifiable data"[18].

The AI Act and the 'Lack of Clarity' on the Disclosure of the Use of Training Data

The emergence of generative AI has disrupted the legislative process of the proposed AI Act and has forced lawmakers to reconsider how they categorise and assign responsibilities to providers and users of AI systems. In its negotiating position adopted at the Strasbourg's plenary session of June 14th, 2023, on the AI Act, the EP proposes adding a clause requiring providers of generative AI systems to "make publicly available a summary disclosing the use of training data protected under copyright law".

Article 28b of the EP approach to the AI Act provides that:

4. Providers of foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video ("generative AI") and providers who specialise in a foundation model within a generative AI system, shall

(c) without prejudice to national or Union legislation on copyright, document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law.

Unlike the United States, there is no copyright register, and copyright laws vary among member states. This makes it challenging to determine whether content is protected by copyright, often requiring legal analysis and even litigation. Additionally, there is uncertainty regarding what constitutes a "sufficiently detailed summary of the use of training data" and how frequently such summaries should be updated. This uncertainty may result in both over-inclusion and under-inclusion in disclosures.

The primary purpose of the disclosure obligation is to empower rightsholders to take legal action against unauthorised use. However, an unclear scope of disclosure may increase the risk of unfounded claims and reduce transparency for rightsholders. As already highlighted previously, the EU's Copyright Directive already provides an option for publishers to opt out of TDM. Since publishers can easily opt out of TDM ex ante, introducing an ex-post transparency requirement is unnecessarily burdensome. The extent to which imposing new disclosure requirements via the AI Act is warranted due to the existence of a regulatory gap for copyright protection in relation to text and data mining (TDM) is therefore questionable. Especially when there are also long-established IP enforcement mechanisms that publishers can use to obtain a court order to compel alleged infringers to disclose relevant information. It should also be kept in mind that AI providers may be hesitant about exposing their intellectual property or trade secrets to the public. Any summary of the use of training data in foundation models represents valuable know-how that constitutes a trade secret. It is therefore commercially unreasonable and a violation of existing Member State IP protections to require public disclosure of such information without a court order tailored to a specific claim. Such a disclosure could also be misused by malicious actors, or lead to leakage of technology that could be misused by malicious actors.

Aside from this, the EP's proposed amendment to the AI Act also contains a somewhat ambiguous

[18] Thomas Margoni, Martin Kretschmer, A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology, GRUR International, Volume 71, Issue 8, August 2022, Pages 685–701, https://doi.org/10.1093/grurint/ikac054

compliance obligation in Art 28b para 4(b), according to which providers are required to "*train, and where applicable, design and develop the foundation model in such a way as to ensure "adequate safeguards" against the generation of content in breach of Union law in line with the generally acknowledged state of the art, and without prejudice to fundamental rights, including the freedom of expression*". It is not clear what standard of diligence "adequate safeguards" entails, especially in relation to potential breaches of copyright laws.

It should be noted here that there is disagreement about whether or when using copyrighted works to develop datasets to train AI models (in both generative and non-generative systems) constitutes an infringement. According to the Notice of Inquiry published by the United States Copyright Office (USCO), which is undertaking a study of the copyright law and policy issues raised by AI systems, "in some cases, a non-generative AI model may be trained on copyrighted material. In other cases, the same AI model may be capable of being deployed in both a generative AI system and a non-generative one". This Notice seeks information about whether permission by and/or compensation to copyright owners is or should be required when their works are included in the training dataset, as well as information about the records that need to be retained in order to identify underlying training materials and the availability of this information to copyright owners and others.

Consequently, more guidance on the proposed article 28b para 4(b) will have to be included in the final agreement on the AI Act, given the substantial fines associated with non-compliance. However, AI providers should prepare for these disclosure obligations, possibly by tracking and documenting their training data.