Essays – *peer-reviewed* i-lex – Rivista di Scienze Giuridiche, Scienze Cognitive ed Intelligenza Artificiale. Vol. 16 n. 2 (2023)
https://doi.org/10.6092/issn.1825-1927/18871 ISSN 1825-1927

Web scraping and Generative Models training in the Directive 790/19

Chiara Gallese[*]

Abstract.

With the rapid development of large generative models, the lack of clarity regarding the possibility of legally scraping the web and using the data set to train AI models, in particular generative ones, has become an urgent issue. Although in the EU web scraping is regulated by Directive 790/2019, AI training is not explicitly mentioned in the text of law.

While for scientific research and teaching web scraping is permitted without exceptions, for other purposes it is allowed if the data is lawfully acquired and if the owner of the copyright did not prohibit so. The Directive allows web scraping for text and data mining for the purpose of gaining new knowledge from the data, but it is not clear if AI training can be considered to fall within this definition. This article aims to analyze the legal dilemma surrounding this topic.

Keywords:

web scraping, AI training, copyright law, intellectual property, generative models, text mining, data mining

1 Introduction

Generative models (GM)[1] such as Midjourney, Dall-e, Stable Diffusion, and Chat GPT have been increasingly used[2] to create new content based on a user's prompt.

For the first time in history, a large number of people have been able to have first-hand access to a unique set of artificial artworks, such as paintings and essays, or to automatically modify existing content at an incredible speed[3], for free or for a small amount of money. AI models have also been used for popular smartphone apps such as Lensa[4] and FaceApp[5].

Tools enabling non-expert users to modify existing content already existed (e.g., Adobe PhotoShop), however, they required a large amount of time and effort, they often had a high cost, and they did not guarantee such an automatic and precise result[6]. In addition, they were not as widespread as the above-mentioned tools, and often required artistic skills of some kind, as it was not sufficient to provide a single prompt in order to obtain a complete result.

---

[*] University of Turin, Department of Law; ✉ chiara.gallese@unito.it

1. Karpathy, A., et al. (2016). *Generative models.* OpenAI blog.
2. Kumar, Satyam, et al. "A Comprehensive Review of the Latest Advancements in Large Generative AI Models." International Conference on Advanced Communication and Intelligent Systems. Cham: Springer Nature Switzerland, 2023.
3. Cao, Yihan, et al. "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt." arXiv preprint arXiv:2303.04226 (2023).
4. Moga, Diana A., and Cosima Rughiniş. "Idealized Self-Presentation through Al Avatars. A Case Study of Lensa Al." *2023 24th International Conference on Control Systems and Computer Science (CSCS).* IEEE, 2023.
5. Neyaz, Ashar, et al. "Security, privacy and steganographic analysis of FaceApp and TikTok." *International journal of computer science and security* 14.2 (2020): 38-59.
6. Knochel, Aaron D. "Midjourney Killed the Photoshop Star: Assembling the Emerging Field of Synthography." *Studies in Art Education* 64.4 (2023): 467-481.

Copyright © 2023 Chiara Gallese
This work is licensed under the Creative Commons BY License.
https://creativecommons.org/licenses/by/4.0/

The quality of the AI-generated works, in fact, is now very impressive[7]: upon a single instance[8] (such as a textual prompt, or the upload of an image), the models can generate all kinds of art styles, imitate specific artists, create realistic photos, mimic or modify existing human faces, write code and specialized essays, and answer specific questions[9], even creating concerns about the future of Google search tools[10].

There is, however, a reason why the output of these models has such a high quality: it does not happen by magic or only by the skills of the programmers, but it is trained on existing materials. This means that the programmer has been feeding existing materials and data to the model in order to train it, often scraping the web to collect a large dataset[11] (in some cases, the materials directly uploaded by users, who are not properly informed about the processing of their data[12], are employed).

Recent cases have highlighted the legal and ethical issues of using data scraped from the web to train generative models[13]. Several artists have been complaining[14] that AI creations inspired by their own works - that were used to train those models without consent or authorization- have created significant damage to their business[15]: once the internet is populated by “artist-inspired” AI images or texts, it is impossible to distinguish the real ones from the generated ones, so the former is eclipsed from web search, resulting in a loss of visibility and related loss of revenue[16]. Similarly, the rise of deep fakes could create consistent reputational damage to individuals[17] and a loss of opportunities for performers[18].

Another problem that may arise from generative AI is the fact that, if copyrighted materials were used to train the model, the output generated after the user’s prompt may clearly display that material (e.g., logos, trademarks, characters), which may then be publicly disclosed, intentionally or by chance[19]. From a legal point of view, it is not easy to determine the applicable law in each case since the different legal systems of the copyright holders might regulate the subject differently[20]. However, in most terms of use issued by the creators of these models, users are warned not to use copyrighted prompts, and the applicable law regulating the contract between the user and the AI creator is pre-determined. Nevertheless, this solution is ineffective and does not solve the issue regarding AI training.

---

[7] Floridi, Luciano. “AI as agency without intelligence: on ChatGPT, large language models, and other generative models.” *Philosophy & Technology* 36.1 (2023): 15.

[8] Suryadevara, Chaitanya Krishna. “GENERATING FREE IMAGES WITH OPENAI’S GENERATIVE MODELS.” *International Journal of Innovations in Engineering Research and Technology* 7.3 (2020): 49-56.

[9] Cousins, S. “The rapid rise of AI art.” *Engineering & Technology* 18.2 (2023): 20-25.

[10] Martens, Bertin. “What Should Be Done About Google’s Quasi-Monopoly in Search? Mandatory Data Sharing Versus AI-Driven Technological Competition.” (2023).

[11] Zhou, Ce, et al. “A comprehensive survey on pretrained foundation models: A history from bert to chatgpt.” *arXiv preprint arXiv:2302.09419* (2023).

[12] Falletti, Elena. “Algorithmic discrimination and privacy protection.” *Journal of Digital Technologies and Law* 1.2 (2023): 387-420.

[13] Dwivedi, Yogesh K., et al. ““So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy.” International Journal of Information Management 71 (2023): 102642; Margoni, Thomas, and Martin Kretschmer. “A DEEPER LOOK INTO the EU text and data mining exceptions: harmonisation, data ownership, and the future of technology.” *GRUR International* 71.8 (2022): 685-701; Crawford, Kate, and Ryan Calo. “There is a blind spot in AI research.” *Nature* 538.7625 (2016): 311-313.

[14] Samuelson, Pamela. “Generative AI meets copyright.” *Science* 381.6654 (2023): 158-161; Kahveci, Zeynep Ülkü. “Attribution problem of generative AI: a view from US copyright law.” *Journal of Intellectual Property Law and Practice* 18.11 (2023): 796-807; Samuelson, Pamela. “Ongoing lawsuits could affect everyone who uses generative AI.”; Abbott, Ryan, and Elizabeth Rothman. “Disrupting Creativity: Copyright Law in the Age of Generative Artificial Intelligence.” *Florida Law Review, Forthcoming* (2022).

[15] Attard-Frost, Blair. “Generative AI Systems: Impacts on Artists & Creators and Related Gaps in the Artificial Intelligence and Data Act.” *Available at SSRN* (2023).

[16] Jiang, Harry H., et al. “AI Art and its Impact on Artists.” *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 2023.

[17] de Rancourt-Raymond, Audrey, and Nadia Smaili. “The unethical use of deepfakes.” *Journal of Financial Crime* 30.4 (2023): 1066-1077; Meskys, E., Kalpokiene, J., Jurcys, P., & Liaudanskas, A. (2020); Meskys, Edvinas, et al. “Regulating deep fakes: legal and ethical considerations.” *Journal of Intellectual Property Law & Practice* 15.1 (2020): 24-31.

[18] Murphy, Gillian, et al. “Face/Off: Changing the face of movies with deepfakes.” *Plos one* 18.7 (2023): e0287503.

[19] Nasr, Milad, et al. “Scalable extraction of training data from (production) language models.” *arXiv preprint arXiv:2311.17035* (2023); Henderson, Peter, et al. “Foundation models and fair use.” *arXiv preprint arXiv:2303.15715* (2023).

[20] Sag, Matthew. “The New Legal Landscape for Text Mining and Machine Learning’(2019).” *Journal of the Copyright Society of the USA* 66: 291; Monterossi, Michael W. “Estrazione e (ri) utilizzo di informazioni digitali all’interno della rete Internet. Il fenomeno del cd web scraping.” *IL DIRITTO DELL’INFORMAZIONE E DELL’INFORMATICA* 2 (2020): 327-369; Zoboli, Laura. *Diritto dei brevetti e intelligenza artificiale*. EGEA spa, 2023; Casonato, Carlo. “L’intelligenza artificiale e il diritto pubblico comparato ed europeo.” *DPCE Online* 51.1 (2022).

The original sin is the lack of clarity regarding the possibility of legally scraping the web[21] and using the data set to train AI models. Although in the EU web scraping is regulated by Directive 790/2019, AI training is not explicitly mentioned in the text of law.

While for scientific research and teaching web scraping is permitted without exceptions, for other purposes, it is only allowed if the data is lawfully acquired and if the owner of the copyright did not prohibit so (for example, by enacting technical measures on their website[22], or by indicating a copyright notice/license).

AI training seems excluded from the definition of text and data mining, especially considering the text of Recital 8. However, the legal scholarship has not provided an answer to this issue**.**

Since AI training is not mentioned, does this mean that it is out of the scope of the Directive, or does it mean that it is prohibited? In the first case, do the general copyright rules apply? How will the forthcoming AI Act regulate the topic? What are the social consequences of the lack of regulation?

The matter certainly needs further investigation and clarification from the legal scholarship. This article aims to provide clarity regarding the possibility of training GM models through web-scraped data according to the discipline of Directive 790/19.

2 Generative Model training

Generative models use various neural network architectures and training methodologies to generate images, code, or text closely resembling human-created content. These models have different applications, from image editing to text-based conversational AI. From a technical point of view, they learn the underlying probability distribution of a dataset and generate new samples that resemble the data distribution. Recently, we have seen substantial advancements in the development of deep learning methodologies, particularly through models like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), or Transformers. However, the most widespread tools in the general public are Midjourney, Dall-e, Stable Diffusion, and Chat GPT. The latter is integrated into many other company's services, such as Bing AI.

The peculiarity of these tools is that they employed a large number of data to train their models. In fact, to reproduce a certain style or subject, they need many different examples, usually labeled and filtered by humans. As known, in machine learning, the training dataset size often correlates with model performance, with larger datasets generally leading to better results, as the overfitting risk decreases[23]. Big data sets are a fundamental resource for training GM effectively: they enable these models to learn patterns in the data and, therefore, generalize, producing realistic outputs. It would be impossible to produce such impressive outputs without employing that amount of data.

Therefore, the creators cannot manually select, review, and collect the desired data set; they must use automatic methods like web scrapers. They often rely on publicly available data[24].

After the data set is created, most companies attempt to eliminate harmful content in various ways, for example, by employing humans[25]. However, developers also employ human "AI trainers" to supervise the fine-tuning process[26].

---

[21] Altobelli, C., et al. "To Scrape or Not to Scrape? The Lawfulness of Social Media Crawling under the GDPR." *Deep Diving into Data Protection; Herveg, J., Ed.; Larcier: Namur, Belgium* (2021).

[22] Gold, Zachary, and Mark Latonero. "Robots welcome: Ethical and legal considerations for web crawling and scraping." *Wash. JL Tech. & Arts* 13 (2017): 275.

[23] Gallese, Chiara, et al. "Investigating Semi-Automatic Assessment of Data Sets Fairness by Means of Fuzzy Logic." *2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB).* IEEE, 2023.

[24] Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018); Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9; Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

[25] As reported by The Times, this process is also raising ethical issues: https://time.com/6247678/openai-chatgpt-kenya-workers (last accessed 26/01/2023).

[26] For example, in the methodological section, Open AI writes the following: "We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format".

The companies that created GM never disclosed the exact corpus of training data they employed to train their models; however, it has become clear sources that many copyrighted materials were employed as part of larger datasets such as the Common Crawl corpus[27]. The content of training datasets can vary widely depending on the specific domain but often includes blog posts, books, online articles, social media posts, and website content.

When such large collections of data are used, it is impossible to manually verify every piece of information to understand whether copyrighted materials are being used, what license is attached to each piece of information, or if offensive and illegal material has been collected. It is also difficult to detect personal data and assess the legal basis for its processing.

When directly asked, ChatGPT replied "ChatGPT, like other language models, has been trained on a diverse range of publicly available text from the internet. However, OpenAI, the organization behind ChatGPT, has not disclosed the specific details of the training data sources or the exact data used to train ChatGPT, including whether it was trained on specific datasets or the proportion of data from different domains. The training data is typically a mixture of licensed data, data created by human trainers, and publicly available text from the internet. The dataset is carefully curated and preprocessed to remove any personally identifiable information (PII) and to ensure compliance with copyright and ethical guidelines."

On the other hand, on the official website of OpenAI, it is stated that "DALL·E 2 is trained on hundreds of millions of captioned images from the internet, and we remove and reweight some of these images to change what the model learns"[28]. While the company did not disclose the full training dataset, Stable Diffusion's company Stability AI released it entirely[29]. Web-scraped data is a substantial part of the training corpus.

2.1 Web scraping

Firstly, the concept of publicly available data has different implications in different countries. Web scraping and secondary use of data sets are not regulated homogeneously around the world, a circumstance that gives rise to a clash of cultures and legal systems, as shown by the recent case law.

As widely known, the US has a legal system that is part of the common law tradition[30], therefore the case law has "the same dignity" as written laws have in the EU. However, a statutory federal law, the Computer Fraud and Abuse Act (CFAA)[31], regulates the criminal aspects of web scraping.

The US case law is particularly noteworthy. In the long legal battle between LinkedIn Corp. and hiQ Labs Inc. regarding web scraping, a California District Court in 2017 and Ninth Circuit panel in 2019 first ruled in favor of the latter[32]. HiQ Labs was a small company that began scraping public profile data and data restricted to registered users from LinkedIn users in order to create a business analytic service. It filed a declaratory judgment request for a preliminary injunction after receiving a cease-and-desist letter from LinkedIn, which demanded that hiQ discontinue its scraping activities. The case went to the Supreme Court, which, in light of the ruling in Van Buren v. United States, granted LinkedIn's petition for certiorari and returned the matter to the Ninth Circuit for further consideration.

In April 2022, the Ninth Circuit upheld the lower court's preliminary injunction in favor of hiQ; however, in October of the same year, a California District Court ruled in favor of LinkedIn. This most recent ruling is noteworthy since it primarily addresses contract liability for violating a website's user agreement and its restrictions on scraping and making fraudulent accounts.

The case has some similarities with another case, Ryanair DAC v. Booking Holdings Inc., in which the famous travel company allegedly acquired information by scraping the ticketing portion of the Ryanair website. According to Ryanair, the action of the defendant violated the CFFA. The Supreme Court's decision in the above-mentioned Van Buren case, which adopted a strict interpretation of "exceeds unauthorized access" under the CFAA, and the Ninth Circuit's earlier decision in the hiQ case, which

---

[27] See Radford, Alec, et al., and Brown, Tom et al., above at 25.

[28] See https://openai.com/research/dall-e-2-pre-training-mitigations last accessed 21/09/2023.

[29] A. Baio, Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator, 2022, https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/ last accessed 21/09/2023.

[30] Liu, Han-Wei. "Two decades of laws and practice around screen scraping in the common law world and its open banking watershed moment." *Wash. Int'l LJ* 30 (2020): 28.

[31] Sellars, Andrew. "Twenty years of web scraping and the computer fraud and abuse act." *BUJ Sci. & Tech. L.* 24 (2018): 372.

[32] See the case description and links to relevant documents in the article on the National Law Review at the following link: https://www.natlawreview.com/article/hiq-and-linkedin-reach-proposed-settlement-landmark-scraping-case (last accessed 27/01/2023).

found that the CFAA's concept of "without authorization" does not apply to "public" websites, were both used by the judge to provide answers to some debated legal issues.

The Court held that "The above cases make clear that in order for the CFAA's "without authorization" and "exceeds authorized access" elements to apply, some sort of authentication mechanism (e.g., the use of usernames and passwords) must be employed to limit access to the website. If the information on the website is publicly available without requiring users to authenticate themselves, a violation of the terms of use or the defiance of a cease-and-desist letter will not give rise to liability under the CFAA".

However, the Court also noted that "Ryanair also alleges that the defendants engaged in screen scraping in violation of the terms of use of the Ryanair website and that the defendants continued to do so even after Ryanair sent each defendant a cease-and-desist letter. Although those allegations would be insufficient to establish liability under the CFAA if the contents of the myRyanair portion of the website were accessible to the public without authentication, courts have found cease-and-desist letters to withdraw authorization to access a protected portion of a website when an authentication mechanism protected access to that portion of the website".

On the contrary, in the EU, the scraping of publicly available content is regulated differently and mostly based on statutory law. Copyright is regulated at the EU level by three directives: Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases; 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; and Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market. Every other aspect is regulated by national law in each Member State. However, as widely known, case law is used in the EU as a guide for the interpretation of laws. Therefore, it remains important for law scholars and should be taken in due account.

In the case between Ryanair and Lastminute.com Srl[33], scraping activity to provide a comparison in ticket prices was deemed lawful by the Court of First Instance[34], the Court of Appeal of Milan[35], and later by the Civil Cassation[36], insofar as copyright protection of the information contained on Ryanair's site was excluded, since "the organization of the data is inspired by management purposes and performs functions of information on the products of the company that manages the site, without an appreciable degree of creative intellectual contribution". The Database Directive's *sui generis* right was excluded as well, because, according to the judge, the fact that Ryanair granted the right to extract its data for a 'symbolic' price of €100.00 would demonstrate not the wish to protect such data, but rather the aim of preventing third parties from marketing its flights.

Although, in the EU, article 7 of Directive790/2019 excludes that contractual arrangements are suitable to prevent web scraping under the conditions set by the Directive, the considerations regarding technical protections to websites remain valid: accessing a website that is password-protected or has other technical protections cannot be considered licit, as shown in the recent case law.

One of the many Rayanair cases, the Ryanair v. PR Aviation BV (C-30/14), was even brought before the Court of Justice of the European Union (CJEU). It involved a dispute between Ryanair and PR Aviation BV, a company that provided online flight booking services. In this case, the central issue was whether PR Aviation BV's practice of "screen scraping" Ryanair's website to obtain flight information and offer it on its own website without Ryanair's permission constituted copyright infringement. The CJEU ruled in 2015 that Ryanair could not claim copyright for the data displayed on its website because such data did not possess the required creativity or originality required for copyright protection, nor claim a *sui generis* right since the construction of the website database did not involve a significant investment.

However, a different scenario presents when web scraping is performed not on publicly available content, but on parts of websites that are only available to subscribers or to registered users. Following the abovementioned case law, the presence of protection in a website (such as a password or a login mechanism) can be considered sufficient to make web scraping illegal, as this activity would not fall under the "lawfully accessible" condition. In some jurisdictions, such as the Italian one, illicitly accessing a website that is not publicly available is even considered a criminal offense[37].

---

33. Iaia, Vincenzo. "La tutela delle concorrenza nell'ambito del trasporto aereo: il caso Ryanair c. Lastminute." *Cammino Diritto* 9 (2020): 1-34.

34. Tribunale di Milano, sez. spec. in materia di impresa, ruling no. 7708 of June 4, 2013.

35. Corte d'Appello di Milano, judgment no. 3585 of October 12, 2015.

36. Corte civile di Cassazione, Sez. I, 12/11/2019, (ud. 18/12/2018, dep. 12/11/2019), judgment no. 29237.

37. Salvadori, Ivan. "Hacking, cracking e nuove forme di attacco ai sistemi d'informazione: profili di diritto penale e prospettive de jure condendo." *Hacking, cracking e nuove forme di attacco ai sistemi d'informazione: profili di diritto penale e prospettive de jure condendo* (2008): 329-366; Salvadori, Ivan. "Il delitto di accesso abusivo ad un sistema informatico o telematico. Sono maturi i tempi per un suo restyling?." *La riforma dei delitti contro la persona*.

More nuanced is the situation of weaker forms of protection, such as copyright logos, terms of service, and licenses. While, in the EU, those mechanisms do not ensure that web-scraping the website content is prohibited, as provided by Article 7 of Directive 790/209, they might be sufficient to exclude it in other jurisdictions. On the contrary, "NO AI" tags and robots.txt might be considered sufficient to prevent TDM. Some authors have argued that website terms and conditions are not contracts but should be classified following property law[38].

Nevertheless, contractual law considerations are interesting for two reasons: on one hand, the materials collected for the training dataset are often covered by contractual clauses (e.g., the dataset license), and, on the other hand, the companies creating GM issued their own terms of use regarding prompts and generated contents.

As for the first matter, GM trainers should be very careful when reusing third-parties datasets (e.g., LAION or Common Crawl), since they often require a non-commercial use of the dataset or have copyleft clauses.

Regarding the second profile, it is interesting to read the Terms and Conditions of some GM providers to find out how they regulate AI outputs and inform users about copyright issues. Since research has shown that, in some particular cases, the AI output is able to reproduce the training data (such as very similar pictures of living persons or registered trademarks), some providers prohibited users from using certain prompts or publishing certain outputs (e.g., Disney characters) and allowed to inquiry the deletion of personal data from the system[39].

2.2 AI Training and Copyright Law

The possibility of using publicly available web-scaped content to train AI models is not clearly regulated in the text of the law, nor has jurisprudence solved the topic. As is often the case, technological progress is faster than legislation, and we face a legal void regarding this issue.

There are several types of AI training, each designed to teach artificial intelligence systems how to perform specific tasks or learn from data, depending on the specific problem they are meant to solve. It is important to consider the difference between supervised learning and unsupervised learning, as the former requires an extensive data-labeling process by a human, and the latter can employ unstructured data, such as web-scraped data. GM typically employ unstructured data, while TDM techniques extract information from structured data.

The issue of collecting publicly available content is not the only problem generated by AI training: often, large models employ copyrighted materials, which are not available on the internet, under the assumption that they are permitted to do so by the fair use exception[40].

Exceptions to copyright are present in many jurisdictions, such as the concept of "fair use", or "non-commercial exception", mainly following the Berne Conventions' three-step test[41], but each country uses a different meaning[42]. Under US legislation, fair use of copyright material is permitted under certain conditions, elaborated and expanded by case law[43].

DiPLaP Editor, 2023. 579-592.

[38] Quarta, Alessandra. *Mercati senza scambi. Le metamorfosi del contratto nel capitalismo della sorveglianza*. ESI, 2020.

[39] From Midjourney's terms of service: "Upon receipt of a notice […], we reserve the right to remove or disable access to the accused material or disable any links to the material; notify the party accused of infringement that we have removed or disabled access to the identified material; and terminate access to and use of the Services for any user who engages in repeated acts of infringement".

[40] Rodriguez Maffioli, Daniel. "Copyright in Generative AI training: Balancing Fair Use through Standardization and Transparency." *Available at SSRN 4579322* (2023); Sandiumenge, Isaac. "Copyright Implications of the Use of Generative AI." *Available at SSRN 4531912* (2023).

[41] See article 9, par. 2: "Right of Reproduction: 1. Generally; 2. Possible exceptions; 3. Sound and visual recordings -(1) Authors of literary and artistic works protected by this Convention shall have the exclusive right of authorizing the reproduction of these works, in any manner or form. (2) It shall be a matter for legislation in the countries of the Union to permit the reproduction of such works in certain special cases, provided that such reproduction does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author. (3) Any sound or visual recording shall be considered as a reproduction for the purposes of this Convention".

[42] Awad, Taysir. "Universalizing Copyright Fair Use: To Copy, or Not to Copy?." *J. Intell. Prop. L.* 30 (2022): 1; Newby, Tyler G. "What's fair here is not fair everywhere: Does the American fair use doctrine violate international copyright law." *Stan. L. Rev.* 51 (1998): 1633.

[43] Samuelson, Pamela. "Unbundling fair uses." *Fordham L. Rev.* 77 (2008): 2537; Netanel, Neil Weinstock. "Making sense of fair use." *Lewis & Clark L. Rev.* 15 (2011): 715.

Recently, jurisprudence has started to balance copyright with technological progress, as shown in Authors Guild v. Google, Inc.[44] The lower court's determination that Google's creation of searchable, digital reproductions of tens of millions of books falls within the parameters of fair use as specified by the Copyright Act was sustained by the Second Circuit[45]. This choice was supported by the claim that Google's search service largely provides information about books and displays a small number of text previews from them, providing a transformative purpose without endangering the book industry.

However, the situation in the US changed with the widespread use of GM. In fact, a group of artists is suing Stability AI Ltd., Midjourney Inc., and DeviantArt Inc. in the first-ever copyright infringement class action regarding the use of protected artwork in training AI models, claiming that this would not fall under the fair use exception[46]. This claim is based on a larger protest within the artistic community, in which several artists complain about the release of GM to the public (instead of limiting its application to research purposes only) and their use in art[47]. According to several digital artists, AI significantly damages their profession[48].

It is not possible to establish a claim based on the overarching presumption that AI systems inherently deprive artists of employment opportunities by supplanting human artistic skills. However, it is feasible to substantiate harm inflicted upon individual artists through specific evidence. In fact, the growing prevalence of AI-generated content labeled as "in the style of" a particular artist on search engines and platforms dedicated to art can overshadow authentic human-created creations[49]. When users, in their quest for original artworks, encounter difficulty discerning genuine works amidst a profusion of other content, artists incur tangible harm[50].

3 Scope of Directive (EU) 2019/790

The Directive offers recommendations for modifying copyright and related rights, emphasizing digital and international contexts in which copyrighted material is used, offering measures to simplify some licensing procedures, and focusing on problems like the circulation of non-commercial works. The major goal is to improve content access in rapidly evolving situations. It also introduces rules that simplify the use of content in the public domain. New laws regulating rights in publications and using works and other materials by online service providers that store and offer access to content posted by users are necessary to keep up with technological progress since Directive 2001/29/EC (the so-called Info-soc Directive) is outdated and unfit to regulate current copyright issues posed by an evolving society[51].

Recital 6 emphasizes that "The exceptions and limitations provided for in this Directive seek to achieve a fair balance between the rights and interests of authors and other rightholders, on the one hand, and of users on the other. They can be applied only in certain special cases that do not conflict with the normal exploitation of the works or other subject matter and do not unreasonably prejudice the legitimate interests of the rightholders".

The Directive disciplines, among others, both research and commercial purposes, providing different rules for each case. This article focuses only on TDM commercial purposes since scientific research is regulated more clearly and does not pose the same legal issues.

---

[44] Campbell, Victoria. "Authors Guild v. Google, Inc." *DePaul J. Art Tech. & Intell. Prop. L* 27 (2016): 59; Mangal, Varsha. "Is fair use actually fair? Analyzing fair use and the potential for compulsory licensing in Authors Guild v. Google." *North Carolina Journal of Law & Technology* 17.5 (2016): On-251.

[45] Diaz, Angel Siegfried. "Fair Use & Mass Digitization: The Future of Copy-Dependent Technologies after Authors Guild v. HathiTrust." *Berkeley Technology Law Journal* 28 (2013): 683-713.

[46] Frosio, Giancarlo. "Generative AI in Court." *Court (September 1, 2023). in Nikos Koutras and Niloufer Selvadurai (eds), Recreating Creativity, Reinventing Inventiveness-International Perspectives on AI and IP Governance (Routledge, 2023, Forthcoming)* (2023).

[47] See, for example: https://arstechnica.com/information-technology/2022/12/artstation-artists-stage-mass-protest-against-ai-generated-artwork/ (last accessed 28/09/2023).

[48] Choi, Eunseo Dana. "Protecting Visual Artists from Generative AI: An Interdisciplinary Perspective." *1st ICML Workshop on Generative AI and Law*. 2023.

[49] See the testimony of Karla Ortiz at https://www.judiciary.senate.gov/imo/media/doc/2023-07-12_pm_-_testimony_-_ortiz.pdf (last accessed 03/12/2023).

[50] See the testimonies at the Senate Judiciary Committee on the 12/07/2023 at https://www.youtube.com/watch?v=uoCJun7gkbA and https://www.judiciary.senate.gov/artificial-intelligence-and-intellectual-property_part-ii-copyright (last accessed 03/12/2023).

[51] Ferri, Federico. "The dark side (s) of the EU Directive on copyright and related rights in the Digital Single Market." *China-EU Law Journal* 7.1-4 (2021): 21-38.

3.1 The meaning of “text and data mining” and the exclusion of GM training

In Computer Science, the practice of identifying patterns and relationships in vast amounts of data is known as data mining. It is referred to also as “knowledge discovery from data (KDD)” by part of the literature, while other authors consider it merely as “an essential step in the process of knowledge discovery”[52]. In order to examine huge digital collections, the area integrates database management with methods from statistics and computational intelligence[53]. According to Han et al., data mining should be named “knowledge mining from data” to better convey its purpose and complexity, at the same way we use the term “gold mining” instead of “rock mining” or “sand mining” when we refer to the mining of gold from rocks or sand[54]. The authors give the following broad definition of data mining: “Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically”[55]. The authors also describe what types of patterns can be mined: “There are a number of data mining functionalities. These include characterization and discrimination; the mining of frequent patterns, associations, and correlations; classification and regression; clustering analysis; and outlier analysis. Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize properties of the data in a target data set. Predictive mining tasks perform induction on the current data in order to make predictions”[56].

To computationally find and extract knowledge from unstructured text is called text mining, also referred to as “knowledge discovery from text”; it is possible to think of text mining as a particular type of data mining, where unstructured textual material is first converted into structured data so that it may be further processed in a more organized manner[57].

According to Han et al., “Text mining is an interdisciplinary field that draws on information retrieval, data mining, machine learning, statistics, and computational linguistics. A substantial portion of information is stored as text such as news articles, technical papers, books, digital libraries, email messages, blogs, and web pages. […] An important goal is to derive high-quality information from text. This is typically done through the **discovery of patterns and trends** by means such as statistical pattern learning, topic modeling, and statistical language modeling. Text mining usually requires structuring the input text (e.g., parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database). This is followed by deriving patterns within the structured data, and evaluation and interpretation of the output.”High quality” in text mining usually refers to a combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity-relation modeling (i.e., learning relations between named entities). Other examples include multilingual data mining, multidimensional text analysis, contextual text mining, and trust and evolution analysis in text data, as well as text mining applications in security, biomedical literature analysis, online media analysis, and analytical customer relationship management”[58].

Jung et al. have performed a review of the most used text mining techniques and describe the process as follows: “Text mining finds new information in human character-based data by extracting context and meaning using natural language and document processing techniques. The typical process of text mining analysis begins with pre-processing the collected text data. Usually at this stage a morphological analysis is performed to sort sentences into parts of speech. The main keywords are extracted based on key topics and words that appear simultaneously in the same paragraphs or sentences. Then the characteristics and frequency of the words are defined and analyzed through a variety of text mining techniques, such as keyword network analysis, association analysis, opinion mining, topic modeling, emotion analysis, and others”[59].

---

[52] Han, Jiawei, et al., *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[53] Clifton, C. Data mining. Encyclopedia Britannica. https://www.britannica.com/technology/data-mining (last accessed 15/09/2023).

[54] Han, Jiawei, et al., at 53.

[55] *Ibidem*.

[56] *Ib*.

[57] Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß. “A brief survey of text mining.” *Journal for Language Technology and Computational Linguistics* 20.1 (2005): 19-62.

[58] Han Jiawei, et al., at 53.

[59] Jung, Hoon, and Bong Gyou Lee. “Research trends in text mining: Semantic network and main path analysis of selected journals.” *Expert Systems with Applications* 162 (2020): 113851.

In summary, data mining and text mining are methods aimed at uncovering patterns, relationships, and insights from large datasets, which can be used for decision-making and predictions, for example, in policymaking.

Article 2 of Directive 790/2019 has its own definition of text mining and data mining. It converges the two terms in a single phrase, “text and data mining”, defining it 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”. According to Recital 8, “New technologies enable the automated computational analysis of information in digital form, such as text, sounds, images or data, generally known as text and data mining. Text and data mining makes the processing of large amounts of information with a view to gaining new knowledge and discovering new trends possible”.

The directive employs a definition of “TDM that in Computer Science is closer to the broad phrase”data mining”, which encompasses a variety of different types of data. It is unclear why the lawmakers decided to restrict the wording of Article 4 only to “text and data mining” instead of maintaining the wording of Recital 7 (“automated computational analysis”) instead of using the general term “data mining”, or expanding the phrase in “data mining and text mining”. The use of a collapsed phrase “text and data mining” might be interpreted as a standalone concept. Separating the legal definition from the one commonly used in Computer Science might create ambiguities and problems in application.

The core and evident features of TDM in the Directive are:

I. automated computational analysis of digital material performed through analytical techniques;

II. gaining new knowledge and discovering new trends;

III. generating information.

It appears evident that the purpose of TDM does not include a) non-digital material; b) non-automated analysis (such as hand-collecting and human analysis); c) non-analytical techniques; d) operations that are merely transformative of pre-existing materials; e) operations that do not generate new knowledge.

Article 4 of the Directive provides that Member States are required to establish a provision that allows for an exception or a restriction to certain aspects of copyright regarding the reproduction and extraction of legally available works for the express purpose of TDM, on the condition that it has not been explicitly reserved by their rights holders through an appropriate method, such as machine-readable means when it comes to publicly available online content.

Article 4 only lists precise copyright rules: 1) Article 5 a) and Article 7 par. 1 of Directive 96/9/EC[60]; 2) Article 2 of Directive 2001/29/EC[61]; 3) Article 4 par. 1 a) and b) of Directive 2009/24/EC[62]; 4) Article 15 par. 1 of the Directive itself[63]. TDM is permitted, then, only within these limits.

Firstly, it is important to note that, because non-digital material is not included, copyright holders might decide to publish their works only in non-digital form to avoid non-authorized use. Or example, some artists might only produce non-digital paintings or sculptures. This might be the only way to protect their work from web scraping. However, third parties sometimes digitalize such work and upload it, in breach of the copyright or relying on the doctrine of fair use (such as Google Books), which is not regulated

---

[60] “In respect of the expression of the database which is protectable by copyright, the author of a database shall have the exclusive right to carry out or to authorize: (a) temporary or permanent reproduction by any means and in any form, in whole or in part”; “1. Member States shall provide for a right for the maker of a database which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database.”.

[61] “Member States shall provide for the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part: (a) for authors, of their works; (b) for performers, of fixations of their performances; (c) for phonogram producers, of their phonograms; (d) for the producers of the first fixations of films, in respect of the original and copies of their films; (e) for broadcasting organisations, of fixations of their broadcasts, whether those broadcasts are transmitted by wire or over the air, including by cable or satellite”.

[62] “(a)The permanent or temporary reproduction of a computer program by any means and in any form, in part or in whole; in so far as loading, displaying, running, transmission or storage of the computer program necessitate such reproduction, such acts shall be subject to authorisation by the rightholder; (b)the translation, adaptation, arrangement and any other alteration of a computer program and the reproduction of the results thereof, without prejudice to the rights of the person who alters the program”

[63] “Member States shall provide publishers of press publications established in a Member State with the rights provided for in Article 2 and Article 3(2) of Directive 2001/29/EC for the online use of their press publications by information society service providers. The rights provided for in the first subparagraph shall not apply to private or non-commercial uses of press publications by individual users. The protection granted under the first subparagraph shall not apply to acts of hyperlinking. The rights provided for in the first subparagraph shall not apply in respect of the use of individual words or very short extracts of a press publication”.

evenly in different countries. Since the Directive explicitly mentions "lawfully accessible" content, what would happen if illicitly shared content is web-scraped is unclear: would it be considered "lawfully accessible" even if it was illicitly digitalized?

Secondly, the Directive excludes non-automated analysis, posing the issue of the multitude of human activities performed manually on web-scraped data. The exceptions or restrictions to copyright laws listed in the Directive seem to exclude such methods; therefore, the usual copyright rules should apply. However, although often non-automatic preprocessing is an essential prerequisite to perform TDM, the Directive is silent on this point.

Thirdly, the difference between GM and TDM should be considered and analyzed.

GM are designed to generate new data instances that have similar statistical properties to the data in the training dataset. These models learn the underlying structure of the data and can create new instances that resemble the training data. **Their scope is not to produce new knowledge** but to capture the distribution of the training data, enabling the generation of realistic outputs. As explained in Section 2, the purpose of scraping the web to collect data for GM training is to build a functioning model, while, in TDM, the data is scraped to be analyzed. The output of the TDM analysis is new knowledge extracted from data and interpreted by a human.

Some particular characteristics of GM are noteworthy:

a. the importance of the dataset in GM is paramount.

b. they do not produce outputs directly from the data, but a user prompt mediates them. This prompt influences the outputs.

c. they can produce so-called hallucinations[64], which are useless to any kind of analysis. The outputs of GM cannot be used for data mining as they are highly inaccurate.

d. they are not able to invent knowledge, they cannot extract information other than that extracted from datasets. Instead, they generate responses based on patterns learned during training. They might be able to generate artistic content re-elaborating the data fed into the model, but they cannot generate new knowledge.

e. they can hardly be considered "analytical techniques", as their primary scope is not to analyze content to discover relations through the scraped data and gain a new understanding of such data since they elaborate it and disaggregate it almost entirely during the training process. GM can mimic an analysis of a new text from a given prompt, but only based on the content that is already present in the training dataset. Any other further analysis cannot be performed on a new text unless a human introduces new knowledge in the prompt.

On the contrary, other types of AI, such as sentiment analysis, fake news detection, statistics, and literary analysis, in our opinion, fully fall within the TDM meaning, although a human is always crucial to produce an insightful interpretation of TDM results. Although TDM models can be trained with data, their use of web-scraped data is not solely finalized to train a functioning model to be used for a different purpose; the goal of the training is to perform data mining.

GM training focuses on capturing sequential dependencies for creative content generation, whereas text mining training aims to extract meaningful patterns or categorize text data for analytical purposes. It appears clear that the intention of the lawmakers was to allow TDM as a tool for creating knowledge about the data, not for using the data to train GM to create human-like content to be sold to users.

In conclusion, the Directive does not include GM training within the meaning of TDM, but it includes other types of AI training, precisely those that are able to generate new knowledge.

3.2 Statements from ChatGPT

Interestingly, when asked if its model could be defined as TDM, ChatGPT correctly replied as follows: "The technique used in ChatGPT-3, which is a large language model based on deep learning, is not typically defined as text and data mining (TDM) in the traditional sense. Text and data mining typically refer to automated computational analysis techniques aimed at extracting structured information, patterns, trends, or knowledge from large volumes of unstructured text or data. TDM often involves techniques like natural language processing, machine learning, and statistical analysis. ChatGPT-3, on the other hand, is primarily designed for natural language understanding and generation. It generates human-like text based on patterns it has learned from a vast corpus of text data but does not typically extract structured information or perform data analysis in the way that TDM

---

[64] OpenAI (2023). GPT-4 technical report. arXiv, 2303-08774.

tools or techniques would. However, it's important to note that the boundary between natural language processing and text and data mining can be somewhat fluid, and the techniques used in ChatGPT-3 may have some overlap with elements of TDM. For example, ChatGPT-3 can be used to extract specific information from unstructured text, summarize text, or answer questions based on text data, which could be considered a form of text analysis. Still, it's not the primary function for which the model was designed. In summary, while ChatGPT-3 uses techniques related to natural language understanding and generation, it is not typically categorized as a text and data mining tool, as its primary focus is on generating human-like text rather than structured data extraction or analysis.". When asked again, the core content remained the same.

When interrogated on the reasons for being able to create artistic content but not new knowledge, ChatGPT replied that "Generative models, including those used for artistic content generation, operate based on patterns and structures learned from existing data during training. These models learn to capture the statistical regularities and features present in the training data and then generate new samples that resemble the learned patterns. However, the generation process is essentially a recombination or synthesis of existing knowledge encoded in the training data rather than the creation of entirely new knowledge. Here are some key reasons why generative models may not produce new knowledge: 1. Training Data Limitation: Generative models are only as good as the data they are trained on. If the training data doesn't contain certain types of knowledge or if the knowledge is not explicitly represented in the data, the model is unlikely to generate it. Generative models lack the capability to independently acquire knowledge beyond what is present in the training dataset. 2. No Inherent Understanding: Generative models lack true understanding or consciousness. They don't possess the ability to comprehend the content they generate. Instead, they rely on statistical patterns and correlations present in the data to produce outputs. 3. No External Knowledge Access: Generative models, including language models like ChatGPT, don't have access to external information or real-time updates. They are limited to the knowledge embedded in the training data up to a certain point in time. They can't access the internet to look up new information or events that occurred after their training cutoff. 4. Creativity vs. Knowledge Synthesis: Generative models excel at creative tasks by combining and reinterpreting existing information. In the case of artistic content, this involves producing novel combinations of visual or auditory elements. Knowledge synthesis, on the other hand, requires the generation of entirely new concepts, principles, or facts, which goes beyond the scope of what generative models can currently achieve. While generative models contribute to creative processes and can produce content with artistic value, the distinction lies in the nature of creativity versus the creation of novel knowledge. Generating new knowledge often involves a deeper understanding, reasoning, and abstraction that current generative models do not possess".

On the other hand, when asked if TDM can produce new knowledge, it replied that "Yes, text and data mining (TDM) have the potential to contribute to the generation of new knowledge. Text and data mining involve the automated analysis of large volumes of textual and structured data to discover patterns, relationships, and insights that may not be immediately apparent to human observers. Here are ways in which TDM can contribute to the creation of new knowledge: 1. Pattern Discovery: TDM techniques can identify patterns and trends within large datasets that may not be evident through manual inspection. By analyzing textual or structured data, TDM algorithms can reveal hidden relationships and correlations, leading to new insights. 2. Knowledge Extraction: TDM can extract valuable information and knowledge from unstructured text, such as articles, research papers, and documents. This process involves identifying key concepts, entities, and relationships, thereby facilitating the synthesis of knowledge from diverse sources. 3. Topic Modeling: Topic modeling is a TDM technique that identifies topics or themes present in a collection of documents. By understanding the main themes within a dataset, researchers can gain insights into prevalent subjects, emerging trends, or areas of interest. 4. Sentiment Analysis: TDM can be used to analyze sentiment in large sets of text data. This can be valuable for understanding public opinion, customer feedback, or the sentiment surrounding specific topics, providing new knowledge about social attitudes and preferences. 5. Prediction and Classification: TDM models, such as machine learning classifiers, can be trained to predict future events or classify data into different categories. This predictive capability contributes to the generation of knowledge by anticipating outcomes based on historical data patterns. 6.Knowledge Integration: TDM allows the integration of information from multiple sources and disciplines. By mining diverse datasets, researchers can bridge gaps between different domains, leading to the creation of interdisciplinary knowledge. It's important to note that the effectiveness of TDM in generating new knowledge depends on the quality of the data, the appropriateness of the algorithms used, and the expertise of the researchers interpreting the results. While TDM can uncover hidden insights, the generation of meaningful knowledge often involves a combination of computational analysis and human interpretation. Additionally, ethical considerations, such as privacy and data ownership, need to be addressed when conducting TDM".

3.3 GM training and the scope of the Directive

Once established that GM training is not included in the definition of TDM in Article 4, we need to consider if it is excluded from the scope of application of the Directive or if it is prohibited by it, as both options have very different legal consequences for AI developers.

If the scope of the directive excludes GM training, we then face a grey area of law, since no other EU regulations mention web scraping, and AI producers that rely on web-scraped data need to reconsider their products. On the contrary, if GM training is included in the scope of the directive but prohibited, then all AI companies that rely on web-scraped data, on the assumption that the directive permits it, need to search for a different legal basis for their products, otherwise they might expect to have them banned from the EU market.

Nonetheless, within the European Union, a fundamental legal principle asserts that in the absence of explicit prohibition within written law (civil law), a presumption of permission arises, with due consideration given to the validity of analogical and extensive legal interpretations. Therefore, when assessing whether web scraping for GM training is permitted in the EU, we must consider the whole corpus of law in the EU and the Member States.

Analyzing the text of the directive, we can see that the terms “generative model”, “large language model”, “artificial intelligence”, “computational intelligence”, and “training” never appear. The only indirect link to AI is the reference to TDM, which can be performed through AI techniques such as Natural Language Processing. In the previous sections, we have argued that TDM models are different from GM models both in their scope and in their functioning. Consequently, we must conclude that the directive excludes GM from its scope entirely, since no provision can be used as an argument to forbid it.

Article 1 par. 2[65] is transparent in stating that the Directive does not in any way affect the other copyright rules that are not included in its provisions: for all other matters related to copyright in the digital society, copyright law is still in place. It is clear that other provisions, such as the Infosoc Directive and internal laws still regulate GM.

Although the internal legal landscape surrounding copyright law is very fragmented, under the general principles of the Infosoc Directive, online content is protected and subject to the decisions of the relevant copyright holders[66]. Consequently, in order to train GM for commercial purposes, companies need to seek permission from each copyright holder before scraping the web or reusing materials that were originally scraped for research, teaching, or cultural purposes.

It is possible to conclude that Midjourney, Dall-e, Stable Diffusion, and Chat GPT creators need to fully comply with the provisions of copyright laws, without the exceptions provided by the Directive, if they want to provide their services in the Member States. Internal copyright laws usually provide exceptions only for specific situations, such as citations, art, teaching, research, and personal use.

It is not possible to argue that web-scraping for the purpose of training commercial GM is permitted in the EU. Although the output of GM is very different from the original work in the dataset, the copyrighted materials were first collected, stored, and fed into the model with the aim of building a commercial tool, which might be considered a violation of copyright. However, this issue is out of the scope of this article.

4 Conclusion

Directive (EU) 2019/790 outlines recommendations to adapt copyright and related rights, particularly in digital and international contexts, simplifying licensing procedures and addressing issues like non-commercial work use. Its primary aim is to enhance content access in evolving scenarios. It also streamlines the use of public-domain content. New laws are required to regulate web scraping and TDM to keep pace with technological advancements, as the existing Directive 2001/29/EC is outdated.

The Directive strives for a fair balance between copyright holders and users, with exceptions and limitations applied in specific cases, such as research, teaching, and cultural purposes.

The Directive defines “text and data mining” as automated computational analysis of digital material to gain new knowledge and generate information. TDM excludes non-digital material, non-automated analysis, and non-analytical techniques. The Directive

---

[65] “2. Except in the cases referred to in Article 24, this Directive shall leave intact and shall in no way affect existing rules laid down in the directives currently in force in this area, in particular Directives 96/9/EC, 2000/31/EC, 2001/29/EC, 2006/115/EC, 2009/24/EC, 2012/28/EU and 2014/26/EU”.

[66] Geiger, Christophe, et al. “The information society directive.” *EU copyright law: a commentary* (2014): 395-527: Renda, Andrea, et al. *The implementation, application and effects of the EU Directive on copyright in the Information Society*. Centre for European Policy Studies, 2015; Senftleben, Martin. “EU copyright 20 years after the InfoSoc Directive–flexibility needed more than ever.” *Reforming Intellectual Property*. Edward Elgar Publishing, 2022. 185-207; Rosati, Eleonora. “From the early days of harmonization to the DSM Directive 2019/790: continuity and complexity of the EU copyright framework.” (2022).

requires Member States to establish provisions for exceptions or restrictions to copyright for TDM, but only within specified limits. There are uncertainties regarding non-digital material protection, non-automated analysis, and the exclusion of AI training from TDM.

This article argues that, while the Directive addresses copyright and related rights in digital and international contexts in a modern way, emphasizing content access and providing text and data mining exceptions, it leaves certain aspects, like GM training, to be regulated by other copyright provisions, which are not entirely clear and differ across jurisdictions.

The differences between GM and TDM have been explored in this work, aiming to show that the Directive intended to restrict its scope only to specific TDM techniques, without addressing the issues posed by GM. After an excursus about the regulation and case law on web scraping, this paper concludes that companies producing or using GM should not base their web scraping practices on the exceptions provided by this Directive.

The EU Digital Strategy would benefit from clearer rules regarding AI training and the conditions under which it is allowed. This requires, however, an impact assessment and a feasibility study, considering the rights of citizens, particularly those in a position of diminished power. The EU has made significant efforts to protect citizens and the EU market from foreign tech companies, but a further step is needed to complete the picture.

5 Bibliography

Abbott, R., & Rothman, E. (2022). Disrupting Creativity: Copyright Law in the Age of Generative Artificial Intelligence. Florida Law Review.

Altobelli, C., E. Johnson, N. Forgó, and A. Napieralski. “To Scrape or Not to Scrape? The Lawfulness of Social Media Crawling under the GDPR.” Deep Diving into Data Protection. Herveg, J., Ed.; Larcier: Namur, Belgium (2021).

Attard-Frost, B. (2023). Generative AI Systems: Impacts on Artists & Creators and Related Gaps in the Artificial Intelligence and Data Act. Available at SSRN.

Awad, T. (2022). Universalizing Copyright Fair Use: To Copy, or Not to Copy?. J. Intell. Prop. L., 30, 1.

Baio, A. Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion’s Image Generator, 2022

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Campbell, V. (2016). Authors Guild v. Google, Inc. DePaul J. Art Tech. & Intell. Prop. L, 27, 59.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of AI-generated content (AIGC): A history of generative AI from gan to chatgpt. arXiv preprint arXiv:2303.04226.

Casonato, C. (2022). L’intelligenza artificiale e il diritto pubblico comparato ed europeo. DPCE Online, 51(1).

Choi, E. (2023). Protecting Visual Artists from Generative AI: An Interdisciplinary Perspective. In 1st ICML Workshop on Generative AI and Law.

Clifton, C. Data mining. Encyclopedia Britannica.

Cousins, S. (2023). The rapid rise of AI art. Engineering & Technology, 18(2), 20-25.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. Nature, 538(7625), 311-313.

de Rancourt-Raymond, A., & Smaili, N. (2023). The unethical use of deepfakes. Journal of Financial Crime, 30(4), 1066-1077.

Diaz, A. S. (2013). Fair Use & Mass Digitization: The Future of Copy-Dependent Technologies after Authors Guild v. HathiTrust. Berkeley Technology Law Journal, 28, 683-713.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., & Wright, R. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642.

Falletti, E. (2023). Algorithmic discrimination and privacy protection. Journal of Digital Technologies and Law, 1(2), 387-420.

Ferri, F. (2021). The dark side (s) of the EU Directive on copyright and related rights in the Digital Single Market. China-EU Law Journal, 7(1-4), 21-38.

Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative models. Philosophy & Technology, 36(1), 15.

Frosio, G. (2023). Generative AI in Court, in Nikos Koutras and Niloufer Selvadurai (eds), Recreating Creativity, Reinventing Inventiveness-International Perspectives on AI and IP Governance (Routledge, 2023, Forthcoming).

Gallese, C., Scantamburlo, T., Manzoni, L., & Nobile, M. S. (2023). Investigating Semi-Automatic Assessment of Data Sets Fairness by Means of Fuzzy Logic. In 2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (pp. 1-10). IEEE.

Geiger, C., Schönherr, F., Stamatoudi, I., & Torremans, P. (2014). The information society directive. EU copyright law: a commentary, 395-527.

Gold, Z., & Latonero, M. (2017). Robots welcome: Ethical and legal considerations for web crawling and scraping. Wash. JL Tech. & Arts, 13, 275.

Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan Kaufmann.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., & Liang, P. (2023). Foundation models and fair use. arXiv preprint arXiv:2303.15715.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. Journal for Language Technology and Computational Linguistics, 20(1), 19-62.

Iaia, V. (2020). La tutela delle concorrenza nell'ambito del trasporto aereo: il caso Ryanair c. Lastminute. Cammino Diritto, (9), 1-34.

Jiang, H. H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., & Gebru, T. (2023, August). AI Art and its Impact on Artists. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (pp. 363-374).

Jung, H., & Lee, B. G. (2020). Research trends in text mining: Semantic network and main path analysis of selected journals. Expert Systems with Applications, 162, 113851.

Kahveci, Z. Ü. (2023). Attribution problem of generative AI: a view from US copyright law. Journal of Intellectual Property Law and Practice, 18(11), 796-807.

Karpathy, A., P. Abbeel, G. Brockman, P. Chen, V. Cheung, R. Duan, I. Goodfel-low, D. Kingma, J. Ho, R. Houthooft, T. Salimans, J. Schulman, I. Sutskever, W. Zaremba, et al. (2016). Generative models. OpenAI blog.

Knochel, A. D. (2023). Midjourney Killed the Photoshop Star: Assembling the Emerging Field of Synthography. Studies in Art Education, 64(4), 467-481.

Kumar, S., Musharaf, D., Musharaf, S., & Sagar, A. K. (2023, June). A Comprehensive Review of the Latest Advancements in Large Generative AI Models. In International Conference on Advanced Communication and Intelligent Systems (pp. 90-103). Cham: Springer Nature Switzerland.

Liu, Han-Wei. Two decades of laws and practice around screen scraping in the common law world and its open banking watershed moment. Wash. Int'l LJ 30 (2020): 28.

Mangal, V. (2016). Is fair use actually fair? Analyzing fair use and the potential for compulsory licensing in Authors Guild v. Google. North Carolina Journal of Law & Technology, 17(5), On-251.

Margoni, T., & Kretschmer, M. (2022). A deeper look into the EU text and data mining exceptions: harmonisation, data ownership, and the future of technology. GRUR International, 71(8), 685-701.

Martens, B. (2023). What Should Be Done About Google's Quasi-Monopoly in Search? Mandatory Data Sharing Versus AI-Driven Technological Competition.

Meskys, E., Kalpokiene, J., Jurcys, P., & Liaudanskas, A. (2020). Regulating deep fakes: legal and ethical considerations. Journal of Intellectual Property Law & Practice, 15(1), 24-31.

Moga, D. A., & Rughiniş, C. (2023, May). Idealized Self-Presentation through Al Avatars. A Case Study of Lensa Al. In 2023 24th International Conference on Control Systems and Computer Science (CSCS) (pp. 426-430). IEEE.

Monterossi, M. W. (2020). Estrazione e (ri) utilizzo di informazioni digitali all'interno della rete Internet. Il fenomeno del cd web scraping. Il diritto dell'informazione e dell'informatica, (2), 327-369.

Murphy, Gillian, Didier Ching, John Twomey, and Conor Linehan. "Face/Off: Changing the face of movies with deepfakes." Plos one 18, no. 7 (2023): e0287503.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., … & Lee, K. (2023). Scalable Extraction of Training Data from (Production) Language Models. arXiv preprint arXiv:2311.17035.

Netanel, N. W. (2011). Making sense of fair use. Lewis & Clark L. Rev., 15, 715.

Newby, T. G. (1998). What's fair here is not fair everywhere: Does the American fair use doctrine violate international copyright law. Stan. L. Rev., 51, 1633.

Neyaz, A., Kumar, A., Krishnan, S., Placker, J., & Liu, Q. (2020). Security, privacy and steganographic analysis of FaceApp and TikTok. International journal of computer science and security, 14(2), 38-59.

OpenAI (2023). GPT-4 technical report. arXiv, 2303-08774.

Quarta, A. (2020). Mercati senza scambi. Le metamorfosi del contratto nel capitalismo della sorveglianza (pp. 1-422). ESI.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Renda, A., Simonelli, F., Mazziotti, G., Bolognini, A., & Luchetta, G. (2015). The implementation, application and effects of the EU Directive on copyright in the Information Society. Centre for European Policy Studies.

Rodriguez Maffioli, D. (2023). Copyright in Generative AI training: Balancing Fair Use through Standardization and Transparency. Available at SSRN 4579322.

Rosati, E. (2022). From the early days of harmonization to the DSM Directive 2019/790: continuity and complexity of the EU copyright framework.

Sag, M. (2018). The new legal landscape for text mining and machine learning. J. Copyright Soc'y USA, 66, 291.

Salvadori, I. (2008). Hacking, cracking e nuove forme di attacco ai sistemi d'informazione: profili di diritto penale e prospettive de jure condendo. Hacking, cracking e nuove forme di attacco ai sistemi d'informazione: profili di diritto penale e prospettive de jure condendo, 329-366.

Salvadori, I. (2023). Il delitto di accesso abusivo ad un sistema informatico o telematico. Sono maturi i tempi per un suo restyling?. In La riforma dei delitti contro la persona (pp. 579-592). DiPLaP Editor.

Samuelson, P. (2008). Unbundling fair uses. Fordham L. Rev., 77, 2537.

Samuelson, P. (2023). Generative AI meets copyright. Science, 381(6654), 158-161.

Samuelson, P. Ongoing lawsuits could affect everyone who uses generative AI.

Sandiumenge, I. (2023). Copyright Implications of the Use of Generative AI. Available at SSRN 4531912.

Sellars, A. (2018). Twenty years of web scraping and the computer fraud and abuse act. BUJ Sci. & Tech. L., 24, 372.

Senftleben, M. (2022). EU copyright 20 years after the InfoSoc Directive–flexibility needed more than ever. In Reforming Intellectual Property (pp. 185-207). Edward Elgar Publishing.

Suryadevara, C. K. (2020). Generating free images with OpenAI's generative models. International Journal of Innovations in Engineering Research and Technology, 7(3), 49-56.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., … & Sun, L. (2023). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. arXiv preprint arXiv:2302.09419.

Zoboli, L. (2023). Diritto dei brevetti e intelligenza artificiale. EGEA spa.