Fair Compensation for Copyrighted Data Used in AI training

Amanda Coelho Della Giustina
(SRN: 2108352)

Master Thesis
LLM. Law & Technology

Supervised by Pratham Ajmera and Dr. Anuj Puri

Tilburg Institute for Law, Technology, and Society (TILT)
Tilburg University
2024

TABLE OF CONTENTS

1. CHAPTER 1 – INTRODUCTION

1.1. OVERVIEW

Artificial intelligence (AI) tools are now essential for solving complex problems and helping humans in making more efficient decisions by enhancing their information-gathering and evaluation capabilities.[1] However, integrating AI systems into our daily basis brings a range of complexities, particularly concerning intellectual property rights.[2] AI systems are no longer just generators of art; they are also evolving into 'consumers' of it, introducing what is known as "upstream problems".[3] This refers to the use of pre-existing copyrighted works for training AI systems.[4]

Since AI systems require a large amount of training data to generate high-quality outputs,[5] there are concerns about potential violations of copyright law[6] and the infeasibility of obtaining individual licenses from all rightsholders.[7] To address this issue, there is a need for a comprehensive framework to safeguard intellectual property rights and ensures that content creators receive fair compensation in the rapidly evolving area of AI technology. In this context, this study aims to explore the intersection of European Copyright legislation and the use of copyrighted works for AI training, specifically focusing on how image-generative models such as Stable Diffusion[8] handle copyrighted works in their training process. Additionally, the research aims to analyze challenges to copyright licensing in the AI context and assess the feasibility of integrating an alternative compensation remuneration system within the European legislative framework.

[1] Mohammed Hossein Jarrahi, 'Artificial intelligence and the future of work: Human-AI symbiosis in organization decision making' (2018) Science Direct 61/4 <https://www.sciencedirect.com/science/article/abs/pii/S0007681318300387> accessed 20 September 2023.

[2] European Commission, Directorate-General for Communications Networks, Content and Technology, 'Study on copyright and new technologies: copyright data management and artificial intelligence' (Study) Publications Office of the European Union (2022) <https://data.europa.eu/doi/10.2759/570559 > accessed 20 September 2023.

[3] Upstream issues refer to the legal challenges emerging when the AI system is trained utilizing pre-existing works safeguarded by copyright. Isaac Sandiumenge Torres, 'Copyright implications of the use of generative AI' (2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4531912 > accessed 20 September 2023.

[4] *Ibid*

[5] Jenny Quang, 'Does training AI violate copyright law? (2021) 36/4 Berkeley Technology Law Journal <https://doi.org/10.15779/Z38XW47X3K> accessed 29 September 2023.

[6] Nicola Litchi, 'ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems' (2023) European Journal of Risk Regulation <http://dx.doi.org/10.2139/ssrn.4483390> accessed 20 September 2023.

[7] Juba Vesala, 'Developing Artificial Intelligence-Based Content Creation: Are EU Copyright and Antitrust Law Fit for Purpose?' (2023) 54 IIC <https://doi.org/10.1007/s40319-023-01301-2> accessed 7 October 2023.

[8] Stable Diffusion is a generative artificial intelligence model that produces realistic images from text and image prompts. See more information: <https://stablediffusionweb.com/> accessed 20 September 2023.

1.2. BACKGROUND AND PROBLEM STATEMENT

1.2.1. AI technology and its relationship with Intellectual Property Rights

"Generative AI" (GenAI) refers to a category of artificial intelligence that regenerates information based on training sets.[9] Unlike traditional AI systems that rely on predefined rules,[10] Generative AI is characterized by its capacity to accurately interpret data, learn from that information and apply those acquired insights to achieve specific goals and tasks.[11] In this context, one of the most prominent AI has been launched by OpenAI, called ChatGPT.[12][13] Similarly, "Stability AI" has developed an image-generative model called "Stable Diffusion".[14] These AI models demonstrate their primary strength in crafting original texts and images that, at first, seem to stand apart from what was used in their training data.[15] Such developments have given rise to a proliferation of AI-generated content across various creative domains, such as music, literature, and more.[16]

Nonetheless, the practices employed by AI companies in training these models have sparked concerns regarding the transparency of data used on AI training processes.[17] Currently, the sources used are kept confidential, which makes it difficult for creators to ensure that AI-driven decision-making complies with copyright law.[18] To ensure transparency in the training process, creators should have access to information on how AI technologies are used in the

[9] Ömer Aydin, Enis Karaarslan, 'is ChatGPT leading Generative AI? What is Beyond Expectations?' (2023) 11(3) Academic Platform Journal of Engineering and Smart Systems <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4341500> accessed 20 September 2023.

[10] Subhajit Panda, Navkiran Kaur, 'Exploring the viability of ChatGPT as an alternative to traditional chatbot systems in library and information centers' (2023) 40(3) Library Hi Tech News <https://www-emerald-com.tilburguniversity.idm.oclc.org/insight/content/doi/10.1108/LHTN-02-2023-0032/full/html> accessed 7 October 2023.

[11] Fiona Fui-Hoon Nah and others, 'Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration' (2023) 25 (3) Journal of Information Technology Case and Application Research <https://doi-org.tilburguniversity.idm.oclc.org/10.1080/15228053.2023.2233814> accessed 7 October 2023.

[12] ChatGPT is an artificial intelligence in form of chatbot. It can generate human-like written texts closely mirroring natural human language patterns. See more information: <https://openai.com/blog/chatgpt> accessed 20 September 2023.

[13] Nishith Reddy Mannuru and others, 'Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development. Information Development' (2023) 0/0 SageJournals <https://doi-org.tilburguniversity.idm.oclc.org/10.1177/02666669231200628> accessed 30 September 2023.

[14] *Ibid* (n. 6)

[15] Nicholas Carlini and others, 'Extracting Training Data from Diffusion Models' (2023) <https://arxiv.org/abs/2301.13188> accessed 30 September 2023.

[16] Krzysztof Walczak, Wojciech Cellary, 'Challenges for higher education in the era of widespread access to Generative AI' (2023) 9(2) Economics and Business Review <https://intapi.sciendo.com/pdf/10.18559/ebr.2023.2.743> accessed 25 November 2023.

[17] Committee on Legal Affairs, *resolution on intellectual property rights for the development of artificial intelligence technologies* (2020) Report 2020/2015(INI) para J and 7 <https://www.europarl.europa.eu/doceo/document/A-9-2020-0176_EN.html > accessed 25 September 2023.

[18] *Ibid* para 18

production of artistic works. This would enable them to create safeguards to protect their intellectual property rights.[19]

For instance, in February 2023, a lawsuit was filed by Getty Images against Stability AI in The United States.[20] It has been argued that Stability AI has copied "more than 12 million photographs from Getty's image collection" without proper permission or licensing authorization. According to Getty Images, the output promoted by AI seems to resemble the original copyrighted images from their website, thereby infringing on copyright law.[21] In parallel, a class action lawsuit was filed by three artists against two AI model creators Stability AI and Midjourney as well as an 'art online community' called DeviantArt[22]. The lawsuit has been presented before the High Court of Justice in London similarly, claiming that copyrighted images were used to train their software without proper permission, subsequently utilizing them to create derivative works.[23]

Traditionally, copyright is recognized as a form of intellectual property right. It is granted fundamental rights protection under Article 17(2) of the Charter of Fundamental Rights.[24] Copyright involves distinct rights about specific uses of artistic works, such as literary works, sound recordings, films and paintings.[25] According to the principle of exclusivity, the owner of copyright – or someone granted exclusive benefits – holds the exclusive right to reproduce, adapt, communicate, make the work available to the public or authorize others,[26] particularly in the digital environment.

In this context, there are significant concerns regarding the protection of artist's intellectual property rights. The main concern relates to the potential for copyright infringement during the AI training phase.[27] AI requires vast amounts of data in the learning process, which often includes copies of existing works without permission from their rightful owners.[28] In this

[19] *Ibid*

[20] *Getty Images (US) Inc v Stability AI Inc* Case 1:23-cv-00135-UNA (US District Court for the District of Delaware, 2023).

[21] *Ibid* 1

[22] *Sarah Andersen, Kelly McKernan, Karla Ortiz v Stability AI Ltd, Stability AI Inc, Midjourney Inc, DeviantArt Inc* Case 3:23-cv-00201 (United States District Court Northern District of California San Francisco Division, 2023).

[23] *Ibid* 14

[24] Charter of Fundamental Rights of the European Union [2012] C326/02

[25] Justine Pila, Paul Torremans, *European intellectual property law* (2nd edition, Oxford University Press, 2019) 221-223

[26] *Ibid*

[27] Pamela Samuelson, 'Generative AI meets copyright' (2023) 381(6654) Science <https://www-science-org.tilburguniversity.idm.oclc.org/doi/full/10.1126/science.adi0656> accessed 9 October 2023.

[28] Enrico Bonadio, Plamen Dinev, Luke McDonagh, 'Can artificial intelligence infringe copyright? Some reflections' in Ryan Abbott (ed), *Research Handbook on Intellectual Property and Artificial Intelligence* (Edward Elgar Publishing Limited 2022) 246

sense, the use of such copyrighted works during the training process could potentially infringe upon the owner's exclusive right of reproduction.[29]

1.2.2. Legal gap and legislative and technical alternatives frameworks

In order to address risks and challenges imposed by the AI on exclusive rights - specifically on right of reproduction - the European Parliament already started deliberating changes to the existing legislation framework to align better with these models.[30] According to the regulation on AI ("AI Act"),[31] companies developing AI models must align with principle of transparency.[32] This means that AI systems must publish detailed summaries of the content used for training to ensure that every decision and action taken by the AI can be traced back to its source (traceability), and inform users when content is generated by AI, allowing users to understand how and why the AI makes certain decisions (explainability).[33]

Moreover, Articles 3 and 4 of the Directive on the Digital Single Market (CDSM Directive) play a significant role in determining the rules around AI and copyright. It introduces two exceptions for Text and Data Mining (TDM) in the context of EU law, allowing a significant portion of copyrighted material to be used in TDM processing activities.[34] TDM is an essential step for machine learning [35] that involves extracting information from machine-readable content. [36] It aims to collect diverse datasets, including copyrighted materials. [37] However, engaging in TDM processes without proper licensing or exceeding agreed-upon terms may violate right of reproduction and lead to copyright issues.[38]

[29] Christopher T. Zirpoli, 'Generative Artificial Intelligence and Copyright Law' (2023) Congressional Research Service (CRS) 3 < https://crsreports.congress.gov/product/pdf/LSB/LSB10922> accessed 30 September 2023.

[30] Committee on Legal Affairs (n. 17)

[31] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 [2024] OJ L 1689/44 (Artificial Intelligence Act)

[32] Artificial Intelligence Act, recital n. 27

[33] Artificial Intelligence Act, recital 27 and Article 50.

[34] Council Directive (EU) 2019/790 of 17 April 2019 on Copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] OJ L 130/92, articles 3 and 4. (The Digital Single Market Directive).

[35] Sean M. Fiil-Flynn and others, 'Legal reform to enhance global text and data mining research: Outdated copyright laws around the world hinder research' (2022) 378/6623 Science <https://doi.org/10.1126/science.add6124> accessed 25 September 2023.

[36] Máxime Barnwell, 'Balancing the benefits of TDM against copyright protection' (Master Thesis, Tilburg University 2018) <https://arno.uvt.nl/show.cgi?fid=146392> accessed 25 September 2023.

[37] Quang (n. 5) 1430

[38] European Parliament, 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspect' (PE 604.942, 2018) <https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf> accessed 6 August 2024.

Additionally, specific tools have been launched lately that technically would allow the removal of copyrighted work from AI systems. Artists and creators have raised concerns about the lack of compensation for the human work used in training AI systems and the way companies collect data through TDM processing.[39] To address this issue, 'Spawning.ai'[40] launched a new AI tool, which technically would allow the removal of copyrighted work from AI systems. Through their application program interface (API) "HaveIBeenTrained",[41] artists can have more control and determine how and where their artwork is employed. In March 2023, this tool facilitated the removal of approximately 80 million pieces of artwork from Stable Diffusion.[42] However, this tool has limitations[43] when it comes to cross-border collaboration among diverse stakeholders, and there may be a need for regulatory intervention.[44]

The widespread use of copyright data in machine learning has created a legal gap in the existing legal framework. Copyright law has traditionally been built upon principles of exclusivity and enforcement.[45] However, it now struggles to address the accessibility and affordability of online copyrighted works. As a result, the current copyright system appears outdated, and it is now challenged in preventing unauthorized content usage and providing fair compensation to creators.[46] Generative AI exacerbates this issue by gathering datasets of copyrighted works through web scraping for training purposes without permission from rightsholders[47] This possibility infringes on copyright and does not compensate copyright owners accordingly for the use of their works. Therefore, addressing these gaps is complex, particularly given the impracticality of compensating for each individual work.[48] Hence, it is essential to

[39] Paul Keller, 'Protecting creatives or impeding progress? Machine learning and the EU copyright framework' (*Institute for Information Law (IViR)*, 20 February 2023) <https://copyrightblog.kluweriplaw.com/2023/02/20/protecting-creatives-or-impeding-progress-machine-learning-and-the-eu-copyright-framework/> accessed 27 June 2023.

[40] Spawning.ai empowers artists and creators to have control over the usage of their works in machine learning training sets. See more information on <https://spawning.ai/> accessed 28 June 2023.

[41] See how the AI works on <https://haveibeentrained.com/> accessed on 28 June 2023.

[42] Kyle Wiggers, 'spawning lays out plans for letting creators opt out of generative AI training' (*TechCrunch*, 3 May 2023) <https://techcrunch.com/2023/05/03/spawning-lays-out-its-plans-for-letting-creators-opt-out-of-generative-ai-training/> accessed 28 June 2023.

[43] Keller (n. 39)

[44] Janna Anderson, Lee Rainie, 'Solutions to address the AI's anticipated negative impacts' (2018) <https://www.pewresearch.org/internet/2018/12/10/solutions-to-address-ais-anticipated-negative-impacts/> accessed 1st October 2023.

[45] Joao Pedro Quintais, *Copyright in the age of online access: Alternative compensation systems in EU copyright law* (40, Law International BV, 2017) 3

[46] *Ibid* 7

[47] Christopher T. Zirpoli (n. 29) 3-4.

[48] Niloufer Selvadurai, Rita Matulionyte, 'Reconsidering creativity: copyright protection for works generated using artificial intelligence' (2020) 15/7 Journal of Intellectual Property Law & Practice 15 <https://academic.oup.com/jiplp/article/15/7/536/5837190> accessed 30 September 2023.

explore legal avenues to enable AI companies to use digital content information while ensuring fair compensation for rights holders and fostering innovation.[49]

In this sense, this research aims to explore how copyrighted material, exemplified by image-generative AI like Stable Diffusion, is implemented during the AI training phase. Then, the study aims to reflect on the implications of incorporating copyrighted data into AI training and on the impact on right of reproduction, given the uncertainty and limited research in this intersection. Moreover, the study will explore the *Austro-Mechana vs. Strato AG*[50] case to assess whether the activities of AI systems could also be considered in the realm of 'reproduction medium' since it involves the temporary storage of data for training purposes. Furthermore, considering the need to ensure proper remuneration in the digital age and the necessity of technological advancement, the study aims to investigate an alternative compensation system beyond traditional licensing schemes. The objective is to address a new model and analyze its feasibility within the EU legal landscape concerning AI, thereby fostering fair compensation for rightsholders.

1.3. LITERATURE REVIEW

The question of whether using copyrighted material during AI training processes constitutes copyright infringement has sparked debate in the literature. Some argue that AI does not replicate the specific characters of existing works. Others, however, believe that such utilization infringes on copyright.

In general, copyright protection is automatically acquired due to human creativity. According to Pila and Torremans,[51] copyright and *related rights* are defined as 'limited-term exclusionary rights that subsist automatically in authorial works'. However, establishing exclusive rights in the "digital realm" has been a challenge for rightsholders.[52] The ease of reproducing and sharing digital content, the global nature of the internet, and the anonymity present obstacles to effective control and enforcement of exclusive rights.[53]

With the development of artificial intelligence and machine-generated works, the European Commission addressed the need to merge AI in the industry, economy, society, and legal frameworks based on the Union's values and fundamental rights.[54] On the other hand, the European Parliament, concerned about the impact and implications of AI on fundamental rights,

[49] Marcello Mariani, 'Generative Artificial Intelligence and Innovation: Conceptual Foundations' (2022) <https://ssrn.com/abstract=4249382> accessed 30 September 2023.

[50] Case C-433/20 [2022] ECLI:EU:C:2022:217.

[51] Pila and Torremans (n. 25) 221

[52] Christian Handke, 'Compensation Systems for Online Use' (2020) <https://doi.org/10.1007/978-3-030-44850-9_15> accessed 30 September 2023.

[53] *Ibid*

[54] Commission 'Artificial Intelligence for Europe' (Communication) COM(2018) 237 final. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN> Accessed 2 March 2023.

suggested several resolutions regarding intellectual property rights and its need for adequate rules with 'transparency, accountability, and verification of AI decision-making.'[55]

The Studies on the Impact of Artificial Intelligence[56] further explain machine learning processes and conclude regarding input content, - that AI needs a large amount of high-quality training data, and questions about authorization for authorial works are raised. In this sense, right of reproduction hamper AI development, and measures related to both rights and text and data mining (TDM) exceptions in the EU should be taken into consideration.[57] TDM involves the automated processing, recognition, and extraction of large volumes of data and text.[58] This is a critical step in preparing datasets for machine learning. By extracting and organizing relevant information from vast datasets, TDM allows machine learning algorithms to identify patterns in the data, which are crucial for model training.[59]

Whether AI companies infringe the copyright of existing works when utilizing them to train AI machines has become a topic of discussion in Europe and worldwide.[60] Matthew Sag argues that the use of copyrighted material might fall within the fair use doctrine. Therefore, the usage of copyrighted data would not infringe on IP rights.[61] Other authors, however, believe that such utilization infringes on copyright. They suggest that further actions to prevent such infringement should be considered, including enforcing exclusive rights and ensuring copyright protection.[62] Moreover, Thomas Margoni draws parallels between human learning and machine

[55] Committee of Legal Affairs, 'Report on intellectual property rights for the development on artificial intelligence technologies' (Report - A9-0176/2020) para 17 <https://www.europarl.europa.eu/doceo/document/A-9-2020-0176_EN.html> accessed 2 March 2023.

[56] European Commission, Directorate-General for Communications Networks, Content and Technology, 'Study on copyright and new technologies: copyright data management, and artificial intelligence' (2012) Publications Office of the European Union <https://data.europa.eu/doi/10.2759/570559> accessed 3 March 2023.

[57] *Ibid* 272-273

[58] Rosana Ducato, Alain Strowel, 'Ensuring Text and Data Mining: Remaining issues With the EU Copyright Exceptions and Possible Ways Out' (2021) Intellectual Property Review <https://ssrn.com/abstract=3829858> Accessed 8 August 2024.

[59] Eleonora Rosati, 'Copyright as an obstacle or an enabler? A European perspective on text and data mining and its role in the development of AI creativity' (2019) 27(2) Asia Pacific Law Review <https://doi-org.tilburguniversity.idm.oclc.org/10.1080/10192557.2019.1705525> accessed 8 August 2024.

[60] See e.g. European Parliament, 'Resolution of 20 October 2020 on Intellectual Property Rights for the Development of Artificial Intelligence Technologies' (2020/2015(INI)) <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020IP0277>; Intellectual Property Office, Consultation outcome Artificial intelligence and intellectual property: call for views (UK, 2021) <https://www.gov.uk/government/consultations/artificial-intelligence-and-intellectual-property-call-for-views>; US Patent and Trademark office (USPTO), 'Public Views on Artificial intelligence and Intellectual Property (2020) AI-Report_2020-10-07 <https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf> accessed 20 July 2023.

[61] Matthew Sag, 'Copyright safety for Generative AI' (2023), Forthcoming in the Houston Law Review 8 <https://ssrn.com/abstract=4438593> accessed 30 September 2023; Carys Craig, 'AI and Copyright' in Florian Martin-Bariteau & Teresa Scassa (eds), *Artificial Intelligence and the Law in Canada* (LexisNexis 2021) 11-13

[62] Selvadurai and Matulionyte (n.48); Keller (n. 39)

data acquisition. He explains that machine learning can memorize information from training datasets. Hence, also constituting infringement.[63]

The Copyright legislation appears inadequate for addressing the complexities of this innovative field. Nevertheless, significant progress is made toward establishing a clear direction with the AI Act.[64] Still, Copyright in the AI context is a subject of ongoing discussion and legal scrutiny. Further research is required to ascertain whether the training phase that uses copyrighted material infringes on the copyright of existing works, as outlined by the Directive on Copyright and Related Rights in the Information Society.[65] Therefore, this study seeks to explore copyright rules and highlight the complexities of enforcing them on AI, particularly in the training phase, to provide valuable insights into the ongoing discourse in this domain.

Furthermore, copyright enforcement in the digital realm fails to remunerate individual authors for their creative works, and it may be undesirable.[66] Alternative Compensation Systems (ACS) emerge as a response to challenges faced by traditional copyright enforcement, aiming to explore alternative ways of remunerating creators in the digital age, particularly in the context of challenges posed by online sharing and distribution.[67] In essence, ACS aims to enhance the remuneration of rights holders by ensuring they receive financial compensation for the use of their work in AI training. This approach is preferred over seeking injunctions, which could lead to numerous individual court cases, an undesirable outcome.

The legal perspective of ACS has been further discussed in the literature regarding *individuals' online* use of copyrighted content. The online use of data by individuals has created a mismatch between law and its applicability for two decades already.[68] On one hand, we have copyright enforcing exclusivity, and on the other, we have the internet with massive amounts of data being downloaded, copied, shared, and remixed, disseminating the data at a very low cost.[69] Expanding copyright in the digital realm signifies a substantial shift from the professional relationships and tangible materials to the virtual relationship and online activities.[70]

[63] Thomas Margoni, 'Artificial Intelligence, Machine learning and EU copyright law: who owns AI?' (2018) CREATe Working Paper, 2 <10.5281/zenodo.2001763> accessed 19 November 2023.
[64] Congressional Research Service (CRS) (n. 29)
[65] Council Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L 167/10 (The InfoSoc Directive)
[66] Quintais, *Copyright in the age of online access: Alternative compensation systems in EU copyright law* (n. 45) 6
[67] Christian Handke, Joao Pedro Quintais, Bodo Balazs, 'The Economics of Copyright Compensation systems for Digital Use' (SERCI Annual Congress 2013, Paris, 8th and 9th July 2013) <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f963447a040555b50c5081020f35dd1e381f0fff> accessed 01st October 2023.
[68] Handke, Quintais and Balazs (n. 67)
[69] Quintais, *Copyright in the age of online access: Alternative compensation systems in EU copyright law* (n. 45) 3
[70] *Ibid*

Moreover, the technological progress of machine learning marks a transformative phase for copyright systems, necessitating a reassessment of copyright applicability.[71] For this reason, this research aims to contribute to the ongoing discourse on the use of copyright content for AI training purposes, mainly to study an Alternative Compensation System - in the context of AI - to promote financial alternative recognition for the losses incurred through the use of their copyrighted works in AI training processes as a forward-looking approach. There is a need to address challenges posed by AI to expand its scope, govern online activities, and adapt to the changing technology landscape.

1.4. RESEARCH QUESTION AND SUB-QUESTIONS

How can an alternative compensation system be implemented to address challenges associated with obtaining licenses for the extensive use of copyrighted content during the generative AI training process?

1) To what extent does the use of extensive datasets in the training of generative AI systems intersect with Intellectual Property Rights, particularly Copyright? *(Chapter 2)*

2) How does generative AI training impact the exclusive right of reproduction, and what are the specific challenges and limitations on Copyright law and on the recent AI Act associated with the use of copyrighted content in generative AI training? *(Chapter 3).*

3) To what degree are AI training activities involving copyrighted materials considered within the scope of "reproduction on any medium," as established by the Austro-Mechana vs. Strato AG case? *(Chapter 4)*

4) Considering the complexities in incorporating copyrighted materials into AI training, the substantial need for mass amounts of data, and the practical challenges in acquiring individual licenses for copyright works, how may an Alternative Compensation System be established to promote financial alternative recognition for losses incurred through the use of copyrighted works in AI training processes? *(Chapter 5)*

1.5. METHODOLOGY

This thesis will use a doctrinal legal research methodology to analyze different legal sources, with a primary focus on the EU, in order to investigate the impact of AI technology on intellectual property (IP) rights, particularly copyright law. The analysis examines how current

[71] Reto Hilty, Jörg Hoffmann, Stefan Scheuerer, 'Intellectual Property Justification for Artificial Intelligence' (2020) Draft chapter. Forthcoming in: J.-A. Lee, K.-C. Liu, R. M. Hilty (eds.), *Artificial Intelligence & Intellectual Property* (Oxford University Press, 2020), Max Planck Institute for Innovation & Competition Research Paper No. 20-02 <https://ssrn.com/abstract=3539406> accessed 13 October 2023.

EU copyright laws, specifically those outlined in the EU Directive 2001/29/EC (InfoSoc Directive), the EU Directive 2019/790 (CDSM Directive), and the AI Act (Regulation (EU) 2024/1689), apply to the use of copyrighted materials in training generative AI. To highlight the challenges and implications of using copyrighted data in AI training, this thesis also references a specific US case, *Getty Images vs Stability AI*. This case serves as an illustrative example to underscore the broader legal issues that arise when copyrighted works are utilized in AI development, even though the focus of the legal analysis remains on the EU context.

The chosen methodology allows for a detailed and systematic examination of relevant legal texts, case law, and legislation. The primary objective is to identify and clarify legal principles, rules, and issues related to the use of copyrighted materials in AI training under the current EU legal framework. The primary sources include EU copyright legislation and relevant case law from the Court of Justice of the European Union (CJEU), such as the *Austro Mechana vs Strato* case. Secondary sources include legal books on EU copyright law, online articles, law review journals, and interdisciplinary perspectives from scientific journals and online legal platforms. These sources provide critical analysis and commentary on the intersection of AI and copyright, offering insights that complement the primary legal analysis.

The fast-evolving nature of AI technology, this research acknowledges the possibility of new legal developments beyond its scope. Despite this limitation, the chosen methodology provides a comprehensive framework for understanding the current legal landscape and its adequacy in addressing the challenges posed by generative AI.

1.6. ROADMAP OF THE ARGUMENTS

Chapter 2 will address Artificial Intelligence (AI) technology, focusing on the training processes of generative image models with Stable Diffusion as a reference. The chapter explores the relationship between generative AI and intellectual property (IP) rights, specifically Copyright rules. It will cover the fundamentals of Generative AI, emphasizing the significance of Machine Learning and Deep Learning in building AI systems. To better understand how Generative AI works, the study will analyze the Stable Diffusion input phase and training, examining how it generates new images, evaluating its training process, and determining how it differs from other AI models. Additionally, the chapter will explore the relationship between Copyright and AI image generative models, emphasizing the importance of addressing dataset usage during the training phase and its implications for Copyright.

Chapter 3 will address the intersection of AI and Copyright in more detail. It will introduce the scope and concepts of European Copyright rules, with a specific focus on how temporary copying and the right to reproduction are affected during the AI training phase. This chapter will emphasize the challenges related to the online use of works and their implications across different stages of AI training. The chapter will also explore exceptions and limitations under the InfoSoc Directive, including an overview of the Text and Data Mining exceptions outlined in the CDSM Directive since Generative AI relies on TDM activities during its training. It will specifically analyze the application of these exceptions in AI contexts, exploring whether using copyrighted materials in AI-related activities qualifies for exceptions. Furthermore, a brief exploration of the AI Act will assess regulatory considerations of copyright rules and their implementation in AI contexts, particularly for AI training purposes. Lastly, the chapter will explore the practicality of data deletion and unlearning practices in AI to evaluate the feasibility of opting-out mechanisms in this context of AI training and data usage.

Chapter 4 outlines the Copyright licensing system, emphasizing the need for updated licensing practices, legal exceptions, and compensation mechanisms to balance innovation and the right holder's interests. The chapter addresses the challenges of obtaining licenses for AI training, exploring potential solutions through legal exceptions and fair compensation by analyzing the private use exception under Article 5(2) b of the InfoSoc Directive. Moreover, the *Austro-Mechana vs. Strato AG* case will be explored in detail to address the digital acts of reproduction and fair compensation in the digital age. The study examines the CJEU's decision regarding whether cloud services fall under the scope of 'reproduction medium' as defined under Article 5(2) of the InfoSoc Directive. This analysis will help determine whether AI companies should be considered as a 'reproduction medium' and to what extent copies stored and reproduced in cloud services require the compensation of rightsholders.

In Chapter 5, the study will focus on arriving at a viable Alternative Compensation System (ACS) to the traditional copyright model, particularly tailored to address the challenges posed by AI and digital content usage. The chapter explores different ACS frameworks, their attributes, examples of implementation in some countries, and their benefits. ACS is assessed as an innovative solution for the losses associated with using copyrighted works in AI training processes, aiming for fair compensation while promoting innovation in AI development.

Chapter 6 will conclude the main research question and sub-questions, synthesizing the presented findings throughout the thesis.

2. CHAPTER 2 — GENERATIVE AI TECHNOLOGY AND THE USE OF DATASETS FOR TRAINING AI PURPOSES

The following subchapters will explore the fundamentals of AI and its training phase. The analysis aims to offer a comprehensive understanding of Generative AI and its distinctions from other AI technologies. Nonetheless, this thesis primarily focuses on a Stable Diffusion, also known as “diffusion-based generative model”,[72] which is used to generate high-quality images. This type of AI allows for a clearer understanding of the training phase, the input and output data. This analysis is important for identifying and mitigating copyright concerns in the AI training context.

2.1. Generative AI and Stable Diffusion

Generative AI (GenAI) models focus on creating new content like images, text, music, and videos.[73] These models learn patterns and styles from existing data and use this knowledge to generate new content.[74] Unlike other machine learning systems that predict outcomes,[75] generative AI learns patterns, transforms input information, and creates new content.[76] Generative AI is trained by recognizing patterns within datasets and generating new outputs based on provided prompts.[77] Additionally, Generative AI can handle massive amounts of data, mapping input information from multiple datasets to create new content that resembles the training data.[78] This AI can process complex information, like text, and convert it into different types of outputs, such as videos or audio.[79] It achieves this by understanding patterns in data and using that understanding to create new content in various formats. Along with its transformative capabilities, generative AI also raises significant legal considerations for Intellectual Property rights, particularly copyright, which will be explored further in the chapter.

[72] Seongmin Lee and others, ‘Diffusion Explainer: Visual Explanation for Text-to image Stable Diffusion’ (2023) <https://arxiv.org/abs/2305.03509> accessed 21 November 2023.

[73] Roberto Gozalo-Brizuela, Eduardo C. Garrido-Merchan, ‘ChatGPT is not all you need. A State of the Art Review of large Generative AI models’ (2023) <https://arxiv.org/abs/2301.04655> accessed 19 November 2023.

[74] Milad Hakimshafaei, ‘Survey of Generative AI in Architecture and Design’ (2023) 16 <https://escholarship.org/uc/item/47x6k9j8> accessed 19 November 2023.

[75] The generative AI has advanced in recent years, leading to the development of sophisticated algorithms like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion models such as Stable Diffusion. In summary, there are different types of generative models. For instance, Generative Adversarial Networks (GANs) focus on generating realistic data by training a generator to receive a discriminator. Variational Autoencoders (VAEs) aim to learn a probabilistic representation of input data in a lower-dimensional space, enabling the generation of new samples through random sampling in that space. Diffusion models are generative models that can be used to generate data similar to the data that was used to train them. They use Gaussian noise in the process. Further explanation will be provided in the study. Hakimshafaei (n. 74) 16-25

[76] Brizuela and Merchant (n. 73) 1-2

[77] Henrik Skaug Sætra, ‘Generative AI: here to stay, but for good?’ (2023) 75/102372 Technology in Society <https://www.sciencedirect.com/science/article/pii/S0160791X2300177X> accessed 19 November 2023.

[78] Hakimshafaei (n. 74) 16

[79] Brizuela and Merchan (n. 73) 2

Nonetheless, all those abilities are possible through machine learning techniques, particularly deep learning, which form the foundation of AI learning.

Generative AI has potential in diverse areas, including art, design and research.[80] It enables the automated creation of novel and diverse content, facilitating creativity and innovation.[81] These features are made possible by machine learning and deep learning, which play a crucial role in the development of AI.[82] Deep learning allows computer systems to analyze huge amounts of data and recognize complex patterns.[83] Deep learning algorithms use neural networks - similar to human brain[84] - with multiple layers to extract features from data, resulting in accurate predictions and decisions.[85] As a result, AI relies heavily on deep neural networks that are extensively trained on diverse datasets. Such capability plays an important role in image recognition and classification models, as they learn from its mistakes, making AI models more accurate and result-oriented.[86]

Generative AI has advanced in recent years. Traditionally, generating images from text was a challenging task for AI,[87] and it was uncommon for algorithms to perform this task.[88] However, recent advancements in generative models, such as Stable Diffusion, have changed this process.[89] Stable Diffusion (also known as *denoiser*[90]) emerged as a powerful tool, significantly advancing AI capabilities when introduced by Stability AI in August 2022.[91] This model has distinct features that enhance its functionality, enabling the model to generate high-quality images from text.[92] Particularly, Stable Diffusion utilizes a specific feature[93] to perform

[80] Stefan Feuerriegel and others, 'Generative AI' (2024) 66 Business Information System Engineering <https://doi-org.tilburguniversity.idm.oclc.org/10.1007/s12599-023-00834-7> accessed 26 February 2024.
[81] Feuerriegel and others (n. 80)
[82] Brizuela and Merchan (n. 73) 1
[83] Md Nazmus Saadat, Muhammad Shuaib, 'Advancements in Deep Learning Theory and Applications: Perspective in 2020 and beyond' in Marco Antonio Aceves-Fernandez (ed) *Advances and Applications in Deep Learning* (IntechOpen, 2020) 13 <DOI: 10.5772/intechopen.92271> accessed 19 November 2023
[84] Laith Alzubaidi and others, 'Review of deep learning: concepts, CNN architectures, challenges, applications, future directions.' (2021) 8/53 Journal of Big Data <https://doi.org/10.1186/s40537-021-00444-8> accessed 29 January 2024.
[85] Andreas Häuselmann, 'Disciplines of AI: An overview of approaches and Techniques' (2021) 48 in Bart Custers and Edward Fischer Villaronga (ed), *Law and Artificial Intelligence: Regulating AI and Applying in Legal practice* (T.M.C Asser Press, 2022) 55 <https://doi.org/10.1007/978-94-6265-523-2> accessed 26 November 2023.
[86] Saadat and Shuaib (n. 83) 9
[87] Hakimshafaei (n. 74) 29
[88] Lee and others (n. 72) 2
[89] Lee and others (n. 72) 2
[90] Carlini and others (n. 15) 2
[91] Nassim Dehouchea, Kullathida Dehoucheb, 'What's in a text-to-image prompt? The potential of stable diffusion in visual arts education' (2023) 9/el6757 Heliyon < https://www.cell.com/heliyon/pdf/S2405-8440(23)03964-6.pdf> accessed 17 November 2023.
[92] Chenshuang Zhang and others, 'Text-to-image Diffusion Models in Generative AI: A Survey' (2023) <https://arxiv.org/abs/2303.07909> accessed 21 November 2023.
[93] Also called "lower dimensional latent space". Hakimshafaei (n. 74) 30

modifications on input data in order to transform the text in high-quality images.[94] Moreover, Stable Diffusion utilizes a approach of adding noise into the images, in a prolonged, slowly and gradual process, which ensures more stability and data quality[95] throughout the processing of billions of images.[96]

Generating new content through generative AI requires extensive datasets for training, making data an essential tool in AI development.[97] Advancements in 'Big Data Analytics,' Deep Neural Networks algorithms, and robust hardware/software infrastructure have enabled AI to collect, analyze, and process a vast amount of data[98] much faster and more efficiently than humans. The performance and effectiveness of generative AI systems are intricately linked to the quality, quantity, representativeness, and diversity of these training data.[99] Consequently, the algorithms heavily rely on information collected to learn and improve. This interdependence between data and AI performance raises pertinent issues regarding copyright and intellectual property rights, particularly relating to its training phase.

2.2. The use of datasets on training AI and implications on Copyright

Joao Pedro Quintais proposes a distinction in addressing copyright aspects within generative AI, suggesting a division in the process into two domains: input or training *vs* output.[100] This delimitation is essential as legal inquiries assume different perspectives when examining copyright rules in both scenarios. While this thesis focuses primarily on the input phase, exploring the legal and economic implications surrounding the use of copyright materials in training generative AI models, it does not exclude occasional observation of output results.

Generative AI technologies rely heavily on datasets containing source materials to operate effectively.[101] This dependency raises significant concerns about intellectual property, mainly because AI models are trained on extensive datasets obtained through internet scraping and extraction.[102] The main issue is the potential inclusion of copyrighted protected works in the

[94] Hakimshafaei (n. 74) 30

[95] Brizuela and Merchan (n. 73) 6

[96] Gowthami Somepalli and others, ‘Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models’ (2023) <https://openaccess.thecvf.com/content/CVPR2023/papers/Somepalli_Diffusion_Art_or_Digital_Forgery_Investigating_Data_Replication_in_Diffusion_CVPR_2023_paper.pdf> accessed 25 November 2023.

[97] Greg Allen, ‘Understanding AI Technology’ (2020) 7 <https://www.ai.mil/docs/Understanding%20AI%20Technology.pdf> accessed 19 November 2023.

[98] Saadat and Shuaib (n.83) 5; Brizuela and Merchan (n. 73) 6

[99] Allen (n. 97) 7

[100] Joao Pedro Quintais, ‘Generative AI, Copyright and the AI Act’ (Kluwer Copyright Blog, 9 May 2023) < https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/> accessed 8 January 2024.

[101]Bonadio, Dinev, McDonagh (n. 28)

[102] Bonadio, Dinev, McDonagh (n. 28)

datasets.[103] For instance, Stable Diffusion relies on sources provided by LAION-5B,[104] a dataset containing approximately 5.86 billion CLIP-filtered image texts.[105] LAION sources its data from Common Crawl,[106] a nonprofit organization that collects data from the web, creating extensive archives and datasets. Despite relying on open sources,[107] copyright issues can still arise. The lawsuit filed by Getty Images against Stability AI (the creator of Stable Diffusion), suggests that copyrighted material was used to train AI models, indicating a reliance beyond open sources.[108] Getty Images claims that over 12 million photographs from their collection were copied without proper permission and that the AI-generated output closely mirrors their copyrighted images.[109] Similarly, three artists have also filed a class action against Stability AI, Midjourney, and the 'art online community' DeviantArt.[110] This legal action refers to the unauthorized use of copyrighted images in training software and subsequently creating derivative works.[111] Therefore, even when AI systems rely on publicly accessible data, the use of copyrighted material without proper authorization raises copyright implications.[112] This is especially significant when copyrighted works are used without securing licenses from the rightsholders, potentially infringing on the exclusive rights granted by copyright law.[113] The unauthorized use of such materials not only risks legal actions but also challenges existing frameworks for copyright protection, leading to debates over how traditional copyright principles should be considered in the context of AI.

As AI technology learns and generates content based on existing copyrighted works, it becomes essential to carefully assess copyright law in AI training. Including copyrighted works in datasets complicates identifying data sources due to the extensive amount of collected information.[114] This makes it difficult for copyright holders to identify their works in the context of AI-generated content, as it is not always clear how to distinguish between copyrighted

[103] Torres (n. 3) 9

[104] LAION-5B is the largest dataset for research purposes. <https://laion.ai/blog/laion-5b/> accessed 25 November 2023.

[105] Christoph Schuhmann and others, 'LAION-5B: An open large-scale dataset for training next generation image-text models' (2022) 2 <https://proceedings.neurips.cc/paper_files/paper/2022/file/a1859debfb3b59d094f3504d5ebb6c25-Paper-Datasets_and_Benchmarks.pdf> accessed 22 November 2023.

[106] Common Crawl is a nonprofit organization systematically exploring the web, gathering extensive archives and datasets. <https://commoncrawl.org/> accessed 26 November 2023.

[107] *Getty Images* (n. 20) para 53.

[108] *Ibid*

[109] *Getty Images* (n.20) para 1.

[110] Sarah Andersen (n. 22*)*

[111] *Ibid*

[112] *Sarah Andersen* (n. 22) para 53-58, para 129.

[113] Lin Yin, 'Copyright Infringement in Al-Generated Artworks' (2024) <https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9158262&fileOId=9158279> accessed 6 August 2024.

[114] Somepalli and others (n. 96) 1

material and original content,[115] posing a challenge for rightsholders, particularly when AI models remain undisclosed. Consequently, proving infringement often places a heavy burden on copyright holders.[116] Additionally, since AI-generated works transform copyrighted content, it can be challenging to demonstrate that AI has reproduced a significant part of the original work.[117]

In order to address concerns around the use of copyrighted materials by AI companies, the AI Act's Article 50[118] requires these companies to fulfil transparency and disclosure obligations for both providers and deployers of AI systems. The AI Act emphasizes the need for transparency by requiring providers to publicly share a summary of their usage of copyright-protected training data.[119] Article 53(1)(d) further specifies that providers must prepare and make available a detailed summary of their usage of such data, following a template provided by the AI Office.[120] Additionally, recital 108 introduces safeguards, including content moderation obligations, to ensure responsible and accountable AI data processing practices. The AI Office is responsible for overseeing whether these obligations are being fulfilled by the providers.[121]

[115] Joris M. Roos, 'Artificial inteligence: Copyright & Consequences' (2023) 20-22 <https://studenttheses.uu.nl/handle/20.500.12932/44493> accessed 3 January 2024.

[116] Joris M. Roos (n. 115)

[117] Joris M. Roos (n. 115)

[118] "(1). Providers shall ensure that AI systems intended to interact directly with natural persons are designed and developed in such a way that the natural persons concerned are informed that they are interacting with an AI system, unless this is obvious from the point of view of a natural person who is reasonably well-informed, observant and circumspect, taking into account the circumstances and the context of use. This obligation shall not apply to AI systems authorised by law to detect, prevent, investigate or prosecute criminal offences, subject to appropriate safeguards for the rights and freedoms of third parties, unless those systems are available for the public to report a criminal offence. (2). Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. Providers shall ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible, taking into account the specificities and limitations of various types of content, the costs of implementation and the generally acknowledged state of the art, as may be reflected in relevant technical standards. This obligation shall not apply to the extent the AI systems perform an assistive function for standard editing or do not substantially alter the input data provided by the deployer or the semantics thereof, or where authorised by law to detect, prevent, investigate or prosecute criminal offences.". Artificial Intelligence Act, Article 50 1 and 2.

[119] "In order to increase transparency on the data that is used in the pre-training and training of general-purpose AI models, including text and data protected by copyright law, it is adequate that providers of such models draw up and make publicly available a sufficiently detailed summary of the content used for training the general-purpose AI model. While taking into due account the need to protect trade secrets and confidential business information, this summary should be generally comprehensive in its scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, for example by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used. It is appropriate for the AI Office to provide a template for the summary, which should be simple, effective, and allow the provider to provide the required summary in narrative form." Artificial Intelligence Act, recital n. 107.

[120] "Providers of general-purpose AI models shall: d. draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office." Artificial Intelligence Act, article 53, 1.

[121] Artificial Intelligence Act, recital n. 108

Despite the outlined requirements for transparency, technical documentation and record-keeping of AI systems,[122] the resolution of copyright issues related to training data remains uncertain. In summary, the complexities introduced by the training process of generative AI require further investigation into how the legal framework can effectively accommodate copyright owners and creators. Accordingly, the next chapter will refer to the relevant legal frameworks, specifically on the European Union's perspective regarding copyright in using copyrighted works within AI training data.

[122] Artificial Intelligence Act, recital n. 9

3. CHAPTER 3 – COPYRIGHT RULES FOR AI TRAINING DATA PROCESSING

3.1. Scope

Copyright constitutes a time-limited exclusive right automatically granted to authors for their intellectual creations, including literary, artistic, musical, or other creative expressions.[123] As part of the intellectual property rights, it aims "to protect the fruits of person's creative efforts from exploitation by others."[124] The copyright legal framework is designed to provide creators control over their works serving as incentive and safeguard for the effort and resources invested in their creations.[125] It acts as a protective mechanism for rightsholders,[126] providing them exclusive rights to authorize or restrict the copy, communication, reproduction, distribution, performance, adaptation and display of their original works by others. To qualify for copyright protection, a work must meet two criteria: it must be original and expressed in a tangible form.[127] Furthermore, copyright also functions as an economic legal framework by allowing creators to profit from their works, thereby motivating them to continue creating works.[128] In this sense, copyright rules aim to find a balance, protecting the rights of creators, ensuring profitability for their works, and simultaneously, facilitating the "free flow of information, ideas and creativity" within the public.[129]

3.2. European copyright law and the right of reproduction

The EU acknowledges international copyright instruments. However, each Member State has its own national laws outlining copyright rules. To harmonize copyright law within the EU, thirteen Directives and two regulations have been introduced,[130] including the InfoSoc Directive (2001/29/EC)[131] and its 2019 revision, the CDSM Directive (2019/790)[132] also known as The Directive on Copyright in the Digital Single Market. This latest legislation adapts copyright law in for the digital age.

The InfoSoc Directive mainly requires EU Member States to acknowledge and protect the copyright of authors concerning their works. Among the rights, reproduction is considered

[123] Pila and Torremans (n. 25) 221

[124] Robin Jacob, Matthew Fisher, Lynne Chave, *Guidebook to intellectual property* (7th edition, 2022) 143

[125] Roos (n. 115) 11-12

[126] *Ibid* 13

[127] Jacob, Fisher and Chave (n. 124) 150

[128] Roos (n. 115) 13

[129] Roos (n. 115) 13

[130] See details about the Directives on the European Commission's website available at <https://digital-strategy.ec.europa.eu/en/policies/copyright-legislation> accessed 6 January 2024.

[131] The InfoSoc Directive (n. 65)

[132] The Digital Single Market Directive (n. 34)

the fundamental element of copyright and related rights.[133] Article 2 of InfoSoc Directive[134] explicitly addresses the right of reproduction, which applies to all authorial works, granting the rightsholders exclusive rights to "authorize or prohibit the complete or partial, direct or indirect, temporary or permanent reproduction by any means and in any form". This is particularly relevant in AI training processes due to its potential to infringe upon reproduction rights.[135] Based on the cited article, *any* replication of content, in whole or in part, either directly or indirectly, constitutes reproduction and may lead to copyright issues.

3.3. AI training and impact on right of reproduction

While humans store information learned in the brain - and such activity is beyond traditional copyright scope - machines need to 'train models' that represent their memory.[136] Incorporating copyrighted material into the data used for training AI processes may impact exclusive rights, particularly reproduction.[137] In general, developers use data sourced from the internet to train machine learning models, storing this data as copies on hard drives and cloud storage.[138] During the training process, AI creates *temporary copies* of materials from datasets to learn specific characteristics and improve information over time. [139] In this sense, storing temporary copies of copyrighted works to memorize and improve AI performance suggests that such actions should be regarded as acts of reproduction. Furthermore, the volume of data required to train machine learning models is often massive and typically scraped from the internet in large quantities. It is likely that some of this training data may be protected by copyright law.[140]

This is relevant in the *Austro Mechana v Strato AG case,*[141] which will be explored further in the next chapter. However, the decision is important from the perspective of copying and storing protected works in cloud services, which might constitute right of reproduction in the digital age. The CJEU held that reproductions could occur in 'any medium' including the servers used in cloud computing.[142] According to the Court, any act of uploading and downloading

[133] Pila and Torremans (n. 25) 279

[134] "Member States shall provide for the exclusive right to authorize 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 organizations, of fixations of their broadcasts, whether those broadcasts are transmitted by wire or over the air, including by cable or satellite". The InfoSoc Directive, Article 2.

[135] Roos (n. 115) 22

[136] Margoni (n. 63)

[137] Bonadio, Dinev, McDonagh (n. 28) 247

[138] Quang (n. 5) 1413

[139] Torres (n. 3)

[140] Quang (n. 5) 1413

[141] *Austro-Mechana* (n. 51) para 17-18

[142] *Austro-Mechana* (n. 51) Para 37-43 and 74

protected works to the cloud devices or media, if not considered as an exception under law, constitutes a reproduction of the content and potentially infringes right of reproduction.[143] Based on that, creating temporary copies of copyrighted materials for AI training should also be considered as acts of reproduction.

Furthermore, copyright implications might arise in different stages of AI training.[144] Firstly, *during the pre-processing stage*[145] (or data gathering and preparation stage), where the acquisition of unlicensed data and its conversion to a format suitable for neural network training might infringe on right of reproduction.[146] Secondly, *during the training process*, where storing copyrighted content for training purposes could result in the unauthorized copying of training data. Lastly, *during the storage of the 'learned information'*, neural networks may inadvertently replicate features from the training data, potentially reproducing substantial portions and giving rise to copyright concerns.[147] In this context, there is a likelihood that entire works may be duplicated and reproduced during AI training. Those actions might infringe on copyright rules unless they fall within copyright exceptions, are authorized by obtaining a license, or are justified by another legal basis.[148]

There are specific provisions under EU copyright law that allow the copying and reproduction of copyrighted materials under certain circumstances without explicit authorization from the copyright owner.[149] Articles 5(1) and 5(2) of the InfoSoc Directive outline the exceptions and limitations that provide a legal basis for using copyrighted works without obtaining a license from the rightsholder.[150] Furthermore, the CDSM Directive also introduced two exceptions for Text and Data Mining activities, which is also relevant in the context of AI training activities. In the following subchapters, we will explore whether using copyrighted materials in AI-related activities is included as an exception and consider the implications on copyright.

3.4. Copyright exceptions: article 5(1) and 5(2) of the InfoSoc Directive

[143] *Austro-Mechana* (n. 51) Para 66

[144] Vesala (n. 7) 353-355

[145] Christophe Geiger, Giancarlo Frosio, Oleksandr Bulayenko, ‘Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU’). Concepción Saiz García and Raquel Evangelio Llorca (eds), *Propiedad intelectual y mercado único digital europeo* (Valencia,Tirant lo blanch, 2019) ., Centre for International Intellectual Property Studies (CEIPI) Research Paper No. 2019-08 <https://ssrn.com/abstract=3470653 or http://dx.doi.org/10.2139/ssrn.3470653> Accessed 31 January 2024.

[146] Vesala (n. 7) 353-355

[147] Vesala (n. 7) 353-355

[148] Vesala (n. 7) 353-355

[149] European Parliament (n. 38) 8

[150] European Parliament (n. 38) 8

EU Member States have the option to choose whether to implement exceptions and limitations to copyright law under Article 5(2) of the InfoSoc Directive.[151] These exceptions are subject to the three-step test outlined in Article 9 of the Berne Convention,[152] and later adopted by the InfoSoc Directive under Article 5 (5). This test ensures that exceptions do not interfere with the normal use of the work or unreasonably harm the author's legitimate interests,[153] and exceptions and limitations are only allowed if the test is observed.[154]

The 'transient copy' exception under article 5(1) of the InfoSoc Directive allows temporary reproduction if it meets specific criteria.[155] The copying must be incidental or transient (purely temporary), essential for the technological process, enable lawful use and lack independent significance.[156] These cumulative requirements, if they are satisfied, can be relevant in context of AI.[157] For example, during training, AI might generate temporary copies of copyrighted material to facilitate the process without retaining the material beyond what is necessary for technological development.[158] If these copies are necessary for lawful use and automatically deleted after use, they may fall under this exception, causing no economic harm to rightsholders.[159] However, Generative AI and Stable Diffusion rely on large, often web-scraped

[151] Eleonora Rosati, 'Copyright and the Court of Justice of European Union' (Oxford, 2019) 128 <https://doi-org.tilburguniversity.idm.oclc.org/10.1093/oso/9780198837176.003.0006> accessed 24 January 2024.

[152] Berne Convention for the Protection of Literary and Artistic Works (adopted 9 September 1886, entered into force 5 December 1887, as revised at Paris 24 July 1971, and amended in 1979) 1161 UNTS 3 (The Berne Convention). The Berne Convention is an international treaty that establishes minimum standards for copyright. These standards cover aspects such as the categories of works protected, the duration of copyright protection, and the scope of exceptions and limitations. Article 6bis of the Berne Convention recognizes copyright as individual moral rights. This provision recognizes the right of the authors to be identified as creators of their works, confirming their authorship. Additionally, it recognizes the right of creators to prevent others from subjecting their works to derogatory treatment throughout the entire duration of copyright. Pila and Torremans (n. 25) 223 and 226.

[153] "(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. The Berne Convention (n. 152), Article 9.

[154] Electronic Frontier Foundation, 'The Three-Step Test' (study) <https://www.eff.org/files/filenode/three-step_test_fnl.pdf> accessed 15 January 2024.

[155] Bonadio, Dinev, McDonagh (n. 28) 251

[156] "Temporary acts of reproduction referred to in Article 2, which are transient or incidental [and] an integral and essential part of a technological process and whose sole purpose is to enable: (a) a transmission in a network between third parties by an intermediary, or (b) a lawful use of a work or other subject-matter to be made, and which have no independent economic significance, shall be exempted from the reproduction right provided for in Article 2." The InfoSoc Directive, Article 5(1); *Infopaq International A/S v Danske Dagblades Forening* Case C-5/08 [2009] ECLI:EU:C:2009:465 *Infopaq International A/S v Danske Dagblades Forening* Case C-302/10 [2012] ECLI:EU:C:2012:16; and Margoni (n. 63).

[157] Bonadio, Dinev, McDonagh (n. 28) 251; Margoni (n. 63) 18.

[158] Margoni (n. 63)

[159] Margoni (n. 63)

datasets.[160] Since digital data can be stored indefinitely, these datasets can persist even after the original data is deleted,[161] excluding them from the Article 5(1) exception. Moreover, the requirement of “independent economic significance’’ becomes a concern when using the data to train commercial AI conflicts with the normal exploitation of the work.[162] If AI training substantially impacts the market for the original works, it may fail the three-step test, prejudicing the interests of the rightsholders.

The rapid progress of technology has outpaced existing exceptions,[163] given rise to Text and Data Mining (TDM) processing activities.[164] In response, the European Commission introduced mandatory exceptions and limitations in the 2019 CDSM Directive.[165] This aims to effectively regulate text and data mining (TDM) activities, addressing copyright challenges in the digital age.

3.5. Text and Data Mining activities and its importance for AI training

Text and Data mining (TDM) is an essential technique for artificial intelligence.[166] TDM automates the processing, recognition, and extraction of large amounts of data and text,[167] uncovering patterns essential for a deep understanding of the extracted information.[168] However, conflicts may arise between intellectual property rights and employment of TDM techniques,[169] especially for AI training for creative purposes.[170] While TDM itself does not fall under exclusive rights granted by copyright law,[171] the automated processing involved may raise copyright concerns, particularly when it includes repeated copying of copyrighted works. This creates a “copyright paradox,[172] precisely in processes aimed at extracting information from copyrighted works.

[160] Robyn Trigg, Catherine Hammon, Arty Rajendra, Will James, ‘Generative AI: can intellectual property infringements in training data be avoided?’ (2023) Lexology < https://www.lexology.com/library/detail.aspx?g=c0f6e9d0-96d9-431a-9bfc-206ed024e06e> accessed 18 January 2024.

[161] Margoni (n. 63)

[162] Trigg, Hammon, James (n. 152)

[163] Ducato and Strowel (n. 58); Torres (n.) 17

[164] Ducato and Strowel (n. 58)

[165] European Parliament (n. 38) 19

[166] Ducato and Strowel (n. 58)

[167] Ducato and others (n. 58) 3

[168] Rosati (n. 59)

[169] European Parliament (n. 38) 5

[170] Rosati (n. 59)

[171] European Parliament (n. 38) 5

[172] Copyright paradox exists where automated processing involving repeated copying of works with the purpose of information extraction, creates a *prima facie* case for infringement. Maurizio Borghi, Stravroula Karapapa, *Copyright and Mass Digitization* (Oxford University Press, online edn, 2013) 51 <https://doi-org.tilburguniversity.idm.oclc.org/10.1093/acprof:oso/9780199664559.001.0001> accessed on 19 January 2024.

Data analysis and pattern extraction typically lie outside the scope of traditional copyright.[173] However, the risk of infringement arises when protected materials are digitalized, formatted, and compiled into datasets for mining and analysis without proper authorization from rightsholders.[174] For instance, during the 'mining stage' of the TDM process (where data is fully extracted), copyright restrictions may apply depending on the software and extraction techniques employed.[175] Therefore, obtaining proper authorization through licensing agreements or legal provisions - such as TDM exceptions within the copyright frameworks - allows for the use of a significant portion of copyrighted material in TDM processing activities[176] and AI training. This is essential as such training often relies on TDM to extract information from datasets that usually include copyright works.[177]

3.6. Text and data mining exceptions - art. 3 and art. 4 of the CDSM Directive

The CDSM Directive has two exceptions related to Text and Data Mining (TDM) activities. Article 3 has a narrower focus, applying specifically to scientific research conducted by certain entities—namely research organizations and cultural heritage institutions—and does not allow rightsholders to opt-out.[178] In contrast, Article 4 has a broader scope, permitting any lawful use of TDM by anyone, but it allows rightsholders to opt-out and prevent their content from being mined.[179]

Article 3 covers TDM activities conducted by research organizations and cultural heritage institutions conducted for scientific research purposes.[180] This exception applies as long as the

[173] Borghi and Karapapa (n. 173).
[174] Flynn and others (n. 35)
[175] Geiger and others (n. 145) 8
[176] Flynn and others (n. 35)
[177] Quintais, *Generative AI, Copyright and the AI Act* (n. 100)
[178] "1. Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access. 2. Copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results (…) " The Digital Single Market Directive, article 3.
[179] "1. Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining. 17.5.2019 EN Official Journal of the European Union L 130/113 2. Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining. 3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online. 4. This Article shall not affect the application of Article 3 of this Directive." The Digital Single Market Directive, article 4.
[180] Ducato and Strowel (n. 58) 3

entity has lawful access to the copyrighted works,[181] without needing additional permission from rightsholders or being restricted by contractual provisions.[182] Article 4, however, introduces a broader TDM exception. It applies to any lawful use of content and is available to anyone engaging in TDM activities,[183] including business and individuals.[184] The key condition is under article 4(3) where rightsholders have the option to opt-out, meaning they can explicitly reserve their rights and restrict the use of their works for TDM purposes.[185]

In this context, Stable Diffusion may be subject to regulations outlined in Article 3, while other generative AI models, such as MidJourney[186] and Dalle-E[187] may qualify for TDM exception under Article 4[188] due to differences in their sources and training methods.[189] Stable Diffusion AI is trained on sources provided by LAION dataset, sourced from Common Crawl, a non-profit web data scraper.[190] In theory, considering its identification as a "non-profit web data scraper," it could potentially fall under the exception of Article 3, as it might be classified as a research organization. Nevertheless, LAION argues that they rely on license-free works and share a database with links to the image files themselves.[191] However, the moment they have the right to use these works license-free and share a database with links for commercial purposes, reproducing copyrighted works would not fall under the Article 3 exception but rather under the scope of Art. 4.[192]

Moreover, Article 4(3) provides the rightsholders with the option of an "opt-out" mechanism, allowing them to restrict the use of their works in the TDM activities, including

[181] Recital 14 of The Digital Single Market Directive stipulates that access is permitted through an open access policy or contractual agreements between rightsholders and research organizations or cultural heritage institutions. This can include subscriptions or other lawful means, considering content that is freely accessible online.

[182] Art. 7 (1) of the CDSM Directive states that contractual limitations cannot override the exceptions outlined in article 3.

[183] Ducato and Strowel (n. 58) 7-8

[184] Torres (n. 3) 18

[185] Torres (n. 3) 19

[186] Midjourney is designed to improve and transform images, utilizing advanced image processing algorithms. This capability allows users to adjust colors, apply artistic filters, add special effects, and create (unique) visual experiences. Midjourney allows users to express their creativity through image manipulation. <https://www.simplilearn.com/dalle-vs-midjourney-vs-stable-difussion-article> and <https://www.midjourney.com/explore> accessed 26 January 2024.

[187] DALL-E, developed by OpenAI, utilizes the power of Generative Adversarial Networks (GANs) to generate images based on textual descriptions. Trained on a vast image dataset, it uses unsupervised and reinforcement learning techniques to generate creative images in response to various prompts and descriptions. <https://www.simplilearn.com/dalle-vs-midjourney-vs-stable-difussion-article> and <https://openai.com/research/dall-e> accessed 26 January 2024.

[188] Torres (n. 3) 28

[189] Mohamed Abduljawad, Abdullah Alsalmani, *Towards Creating Exotic Remote Sensing Datasets Using Image Generative AI* (IEE, 2022)

[190] LAION (n. 104)

[191] Torres (n. 3) 27-28

[192] Torres (n. 3) 28

machine learning.[193] Rightsholders can explicitly prevent their works from being employed on AI training purposes[194] through methods like machine-readable formats, metadata, contractual agreements, unilateral declarations,[195] and tools such as Spawning.ai. However, reserving rights for individual works has become critical to prevent unrestricted reproduction and extraction, especially for commercial TDM purposes like training AI models. Obtaining documented and proven consent from the rightsholders for data extraction purposes becomes mandatory for those seeking to retain control over the utilization of their works for AI training.[196] Rightsholders must be aware of whether their works are being used for training purposes,[197] as information about training data is often ambiguous and unclear. A collective call for transparency from copyright holders is essential to address this issue effectively.[198]

3.7. AI Act: A solution?

The AI Act is a significant step in regulating artificial intelligence in the digital era. It emphasizes transparency, particularly in Text and Data Mining (TDM) activities for AI training. [199] Recognizing the intersection between AI and intellectual property rights, the European Parliament incorporated specific copyright-related provisions into the Act.[200]

One of the key documents in the legislative process of AI Act is the "Compromise Proposal on General-Purpose AI Models/General-Purpose AI Systems," [201] published on December 8, 2023. Article C (1) of this proposal requires that providers of AI models and systems must: a) create and maintain technical documentation, outlining the AI model's training and testing process along with the results; b) develop and maintain information and

[193] Torres (n. 3) 19
[194] Torres (n. 3) 19
[195] Torres (n. 3) 19
[196] Aikaterini Simopoulou, 'Text and Data Mining under EU Copyright law' (2020) 11 <https://repository.ihu.edu.gr/xmlui/bitstream/handle/11544/29743/Text%20and%20Data%20Mining%20under%20EU%20Copyright%20Law.pdf?sequence=1> accessed 31 January 2024.
[197] Quintais, *Generative AI, Copyright and the AI Act* (n. 100)
[198] Katharina Uppenbrink, Matthias Hornschuh, Thomas Höppner, 'Our call for safeguards Around Generative AI' (Initiative Irheberrecht, 2023) <https://urheber.info/media/pages/diskurs/call-for-safeguards-around-generative-ai/c93a5ab197-1681904353/final-version_authors-and-performers-call-for-safeguards-around-generative-ai_19.4.2023_12-50.pdf> accessed 29 January 2024.
[199] Recital 27 of the AI Act states that transparency is a fundamental AI Act principle relating to all AI-based systems. Providers are encouraged to integrate this principle to offer clearer and more considerate insights into the functioning of their AI-based systems, including details about the model's operation, data utilized in training, and accurate information. Thus, transparency obligations are outlined under Article 13 for High Risk AI Systems, and Article 50 for general purposes AI.
[200] European Parliament (PT) 'Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI' COM (20231206IPR15699) <https://www.europarl.europa.eu/news/pt/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai> accessed 29 January 2024.
[201] Global Partnership on Artificial Intelligence, 'GPAI Compromise Proposal' (2023) < https://www.openfuture.eu/wp-content/uploads/2023/12/231206GPAI_Compromise_proposalv4.pdf> accessed 29 January 2024.

documentation accessible to AI system providers; c) implement a policy adhering to EU copyright law, specifically Article 4(3) of the CDSM Directive; d) Prepare and publicly release a detailed summary of the content used to train the model or system.[202]

The following provisions were ultimately included in the final version of the AI Act. They are designed to strengthen copyright rules under the AI Act, ensuring that providers consider the opt-out mechanism outlined in Article 4(3) when using copyright data for AI training. Article 53(1)(c)[203] of the Act requires providers to establish a policy that ensures compliance with EU law on copyright and related rights. It requires any provider introducing a general-purpose AI model to the Union market to adhere to the specified opt-out mechanism.[204] The AI Act also requires providers to disclose detailed information used in their training datasets,[205] which helps to comply with transparency principle and cooperate with copyright rules and the opting-out mechanism, thereby strengthening the rights of rightsholders.[206]

3.8. Unlearning the AI

Although mechanisms are available for opting out, the effectiveness of AI in forgetting or erasing accessed and trained data must be addressed. Whether AI companies are able to comply with previous rules depends on the feasibility of AI being able to forget or erase the acquired information[207] rather than solely relying on the tools or mechanisms provided for users to opt out.

Unlearning and deleting acquired data in AI systems is a challenge,[208] primarily due to their capacity to learn and develop from previous datasets,[209] which makes data revocation and

[202] *Ibid*

[203] "Providers of general-purpose AI models shall: … c. put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790;" Artificial Intelligence Act, Article 53 (1) (c).

[204] Recital 106 of the AI Act specifies that providers (including the ones placing the AI into the EU market) must establish policies to comply with copyright law, particularly Article 4(3). "(…) To that end, providers of general-purpose AI models should put in place a policy to comply with Union law on copyright and related rights, in particular to identify and comply with the reservation of rights expressed by rightsholders pursuant to Article 4(3) of Directive (EU) 2019/790. Any provider placing a general-purpose AI model on the Union market should comply with this obligation, regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place. (…)."

[205] Artificial Intelligence Act, Article 53 (1) (d).

[206] Paul Keller, 'Generative AI and copyright: convergence of opt-outs?' (Kluwer Copyright Blog, 2023) < https://copyrightblog.kluweriplaw.com/2023/11/23/generative-ai-and-copyright-convergence-of-opt-outs/> accessed 29 January 2024.

[207] Bjorn Aslak Juliussen, Jon Petter Rui, Dag Johansen, 'Algorithms that forget: Machine unlearning and the right to erasure' (2023) 51/105885 Computer Law & Security Review <https://www.sciencedirect.com/science/article/pii/S026736492300095X?via%3Dihub> accessed 5 February 2024.

[208] Lucas Bourtoule and others, 'Machine Unlearning' (2021) <doi: 10.1109/SP40001.2021.00019> accessed 5 February 2024.

deletion complicated. This challenge is intensified by the unrestricted dissemination of online data through practices like web scraping.[210] Online data is regularly scraped for activities such as web crawling, and AI models are often trained using this data.[211] For instance, Stable Diffusion is trained on LAION dataset provided by Common-Crawl. If rightsholders opt-out using tools like 'spawning.ai,' it becomes essential to remove that the data from the current dataset and refrain from including it in future training sets. Nevertheless, the issue at hand may require more than simply removing the data from the datasets.

AI systems learn by analyzing large datasets to identify complex patterns and relationships within the data. These patterns are embedded in the AI parameters and weights during training.[212] Even if specific data points are removed, the overall patterns and relationships learned from the data may still persist within the AI's framework, allowing it to retain its learned behavior.[213] This can lead to potential copyright concerns, as patterns and information may have be integrated within AI. Therefore, addressing potential copyright issues associated with AI-generated content requires other approaches that go beyond data deletion. Several approaches are suggested in the literature to train AI in unlearning data.[214] However, implementing such techniques may be challenging for most AI applications. Retraining an AI to unlearn data is complex, time-consuming, costly, computationally intensive, and energy consuming. Additionally, if data deletion is performed, it requires retraining the model from scratch.[215]

Given these challenges, when dealing with large datasets and incorporating copyrighted content into AI training, it is essential to highlight that learned patterns can become deeply ingrained. Retraining AI models to unlearn specific patterns may prove challenging for companies due to time and cost constraints. Overall, the current copyright provisions do not adequately address the complexities of this activity. It becomes imperative to explore the (in)feasibility of obtaining individual licenses in the AI context and ensure rightsholders are compensated for their contribution to AI learning.

[209] Juliussen, Rui and Johansen (n. 207)

[210] Dawen Zhang and others, 'Tag your Fish in the Broken Net: A Responsible Web Framework for Protecting Online Privacy and Copyright' (2013) < https://doi.org/10.48550/arXiv.2310.07915> accessed 5 February 2024.

[211] Dawen Zhang and others (n. 210) 7

[212] Dawen Zhang and others (n. 210) 7

[213] Dawen Zhang and others (n. 210) 7

[214] For instance, 'SISA training' is a framework designed to accelerate the unlearning process by strategically constraining the impact of individual data points during the training procedure. Bourtoule and others (n. 208). In addition to training AI to unlearn data, researchers are exploring other techniques like "Q-k-Means". This method involves 'quantizing centroids algorithms' in iterative processes, thereby enhancing the efficiency of data deletion by increasing the likelihood of erasing information effectively. Antonio A. Ginart and others, 'Making AI Forget You: Data Deletion in Machine Learning' (33rd Conference on Neural Information Processing Systems NeurIPS, 2019) 4 <https://proceedings.neurips.cc/paper_files/paper/2019/file/cb79f8fa58b91d3af6c9c991f63962d3-Paper.pdf> accessed 8 August 2024.

[215] *Ibid* 2

4. CHAPTER 4 – FAIR COMPENSATION

Despite the AI Act regulations requiring providers to clarify the content of data used in AI training to enhance compliance with copyright laws, significant uncertainty remains regarding how the licensing framework should be applied in the context of AI. This is particularly challenging given the way copyrighted works are scraped and used in AI training. This uncertainty extends to questions about how rightsholders will be financially compensated for the use of their works in AI training. Given the gaps in current legal frameworks on this issue, the following chapters will explore the financial perspective of rightsholders, focusing on the economic implications they face under the existing licensing system in the context of AI.

4.1. Licensing copyright material for AI training

Copyright rules were originally established to protect the creative works of individuals, providing them with an incentive and reward for their efforts while also allowing society to benefit from the free exchange of ideas and information.[216] Licensing enables rightsholders to authorize or prohibit others from using their works while receiving compensation. Traditionally, licenses are the primary legal mechanism for granting permission to use copyrighted materials.[217] They are contractual agreements between the owner of IP rights and the licensee and the terms, determined by the IP owner, may only apply to certain IP rights.[218] For instance, in the context of copyright, a license may be necessary to copy, reproduce or distribute a work, as they are exclusive rights of the rightsholder.[219] The CJEU has clarified that making a copy, even by an individual acting privately, can cause prejudice to the rightsholders if done without the prior authorization.[220] Thus, anyone seeking to copy, reproduce, or distribute a protected work must obtain authorization. Without an agreement between the author and the other party and in the absence of exceptions, unauthorized use of the work constitutes a violation of the exclusive rights,[221] thereby constituting copyright infringement.[222]

In the context of AI training, developers use data to expand machine learning and improve AI performance.[223] Therefore, obtaining licenses and permissions is essential for using

[216] Quang (n. 5) 1412

[217] Flynn and others (n. 35)

[218] Danish Contractor and others, ‘Behavioral Use Licensing for Responsible AI’ in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* (Association for Computing Machinery, New York 2022) 781<https://dl.acm.org/doi/abs/10.1145/3531146.3533143> accessed 8 August 2024.

[219] Barnwell (n. 36) 20

[220] *Padawan SL v Sociedad General de Autores y Editores de España (SGAE)* Case C-467/08 [2010] ECLI:EU:C:2010:620 para 45.

[221] Barnwell (n. 36) 20-21

[222] Jenny Quang argues that copyright infringement is not limited to copying a work's material form but also, it constitutes by the unauthorized use of the work for its expressive purpose. Quang (n. 5) 1408.

[223] Quang (n. 5) 1413

copyright works. However, securing these licenses is challenging due to the large volume of data involved in TDM activities.[224] Licensing individual works is impractical,[225] especially when copyright owners are either unidentifiable or unwilling to grant permission.[226] This creates a significant issue with the unauthorized copying and reproduction of copyrighted content, given the extensive use of datasets for AI training, thus (economic) harm to rightsholders. In this regard, the European Union (EU) has previously confronted similar challenges regarding reproduction rights in private copying for non-commercial use.

4.2. Copyright exception on right of reproduction and conception of fair compensation

The concept of 'Copyright Levies' was introduced to offset losses from unauthorized reproductions of music and films[227] with the introduction of sound recording equipment's.[228] This involved imposing a levy on equipment sales to balance consumers' rights to make private copies while compensating rights holders.[229] In 2001, the EU implemented a harmonized exception for private copying and copyright levies under Article 5(2)(b) of the InfoSoc.[230] This aimed to address the economic harm to copyright holders from private copying, allowing non-commercial use, prevent copyright infringement, and ensure fair compensation through the levies system.[231] Article 5(2)(b) clearly states this:

> 2. Member States may provide for exceptions or limitations to the reproduction right provided for in Article 2 in the following cases: (…) (b) in respect of reproductions *on any medium* made *by a natural person for private use* and for

[224] Torres (n. 3) 20. Barnwell (n. 36) 20-21

[225] Quang (n. 5) 1413

[226] Flynn and others (n. 35)

[227] Martin Kretschmer, 'Private Copying and fair compensation: a comparative study of copyright levies in Europe'(An Independent Report Commissioned by the UK Intellectual Property Office, (2011) <https://microsites.bournemouth.ac.uk/cippm/files/2014/11/levy-final-report-jun-2011.pdf> accessed 20 July 2023

[228] IRIS plus 2011-4, Who Pays for Private Copying? in Susanne Nikoltchev (ed), *European Audiovisual Observatory* (Strasbourg, 2011) 8-9 <https://rm.coe.int/1680783bc7> accessed 10 June 2024.

[229] "In certain cases of exceptions or limitations, rightsholders should receive fair compensation to compensate them adequately for the use made of their protected works or other subject-matter. When determining the form, detailed arrangements and possible level of such fair compensation, account should be taken of the circumstances of each case. When evaluating these circumstances, a valuable criterion would be the possible harm to the rightsholders resulting from the act in question. In cases where rightsholders have already received payment in some other form, for instance as part of a license fee, no specific or separate payment may be due. The level of fair compensation should take full account of the degree of use of technological protection measures referred to in this Directive. In certain situations where the prejudice to the rightsholder would be minimal, no obligation for payment may arise." The InfoSoc Directive, recital 35.

[230] "(…) 2. Member States may provide for exceptions or limitations to the reproduction right provided for in Article 2 in the following cases: (…) (b) in respect of reproductions on any medium made by a natural person for private use and for ends that are neither directly nor indirectly commercial, on condition that the rightsholders receive fair compensation which takes account of the application or non-application of technological measures referred to in Article 6 to the work or subject-matter concerned." The InfoSoc Directive, Article Art 5(2)(b).

[231] Kretschmer (n. 227).

> ends that are neither directly nor indirectly commercial, *on condition that the rightsholders receive fair compensation* which takes account of the application or non-application of technological measures referred to in Article 6 to the work or subject-matter concerned.

Private copying has led to a new market for the exploitation of copyrighted works, benefiting private users and equipment manufacturers.[232] Levies were originally imposed to ensure rightsholders received fair compensation, addressing their economic interests[233] while also aiming to address the uncontrolled and mass copying and reproduction facilitated by sound and video recording equipment. This is particularly illustrated in the recent *Austro-Mechana vs. Strato AG*[234] case, which will be further illustrated.

4.3. Fair compensation and digital acts of reproduction

4.3.1. Case analysis

Technological advancements, particularly in internet access and storage, have transformed the reproduction and distribution of protected works,[235] prompting shifts in content usage.[236] The Austrian case of *Austro-Mechana vs. Strato AG* addresses the complexities of copying and storing protected works, which may constitute right of reproduction in the digital age. The decision aims to extend copyright protection to remain relevant amid technological progress.[237] The European Court of Justice (CJEU) ruled that the 'private copying' exception under Article 5 (2)(b) of the InfoSoc Directive covers cloud computer (storage) services. The Court held that such an exception applies to reproduction of copyrighted works *on any medium.*[238] Furthermore, the CJEU held that if Member States adopt this exception, they must ensure fair compensation, prioritizing the economic interests of rightsholders.[239]

The Court examined whether reproducing content in an online storage space provided by a third party, commonly known as cloud services, constitutes reproduction in any medium under exception on Art. 5(2)(b). The Court's interpretation of the law, particularly the phrase "reproduction in any medium," led to the conclusion that the private copying exception under the InfoSoc Directive extends beyond physical form, including to digital forms of reproduction and

[232] IRIS plus 2011-4 (n. 228) 8-9
[233] *Ibid* 8-9
[234] *Austro-Mechana* (n. 51)
[235] National Research Council, et al. *The digital dilemma: Intellectual property in the information age*. (National Academies Press, 2000) 3-8
[236] European Digital Rights, 'Private Copying levies' (Edris) <https://www.edri.org/files/privatecopyinglevies.pdf> accessed 9 June 2024.
[237] *Austro-Mechana* (n. 51) Para. 25-26
[238] *Austro-Mechana* (n. 51) Para 30-33, 55
[239] *Austro-Mechana* (n. 51) Para 40-41, 55

those made via cloud services.[240] Moreover, the Court held that unless an exception applies, each instance of uploading and downloading copyrighted content to the cloud from devices or media (i.e., such as smartphones) constitutes a reproduction, potentially infringing on Article 2 of InfoSoc Directive.[241] Additionally, the Court emphasized that the introduction of the exception under Article 5(2)(b) by the legislator was intended to prevent undue harm to rightsholders resulting from uncontrolled and generalized private copies done by individuals.[242] The Court highlighted that Member States have the authority to decide whether to implement the exception outlined in Article 5(2)(b) into their national legislation. However, compensation under Article 5(2)(b) becomes mandatory once implemented. The Court stated such compensation serves as means to estimate the harm caused to rightsholders.[243] It also highlighted the impracticality of determining whether a protected work was reproduced by each user and on which medium.[244] Given the inherent challenges in monitoring or identifying such reproductions, the Court held that EU allows Member States to establish specific presumptions regarding private copying to quantity and assess the harm caused, thus determining equitable remuneration.[245] Overall, reproduction of copyrighted material can be carried out through digital media, such as cloud computing services, which enable storing and access of protected works. These services facilitate the reproduction of content in a cloud environment that is made available.[246] However, the Court stated that such (cloud) services qualify as exceptions to right of reproduction if equitable remuneration is provided.[247] The Court's rationale in this matter can be *partially* extended to the context of AI, as it will be illustrated further.

4.3.2. Fair compensation in the context of AI

Despite the legal provision allowing for private copying, there is an increasing challenge in monitoring and regulating the reproduction of protected works in digital formats. Copying and storing data for AI training may involve activities similar to private copying (i.e., mass use of works) since it has been used for large-scale reproduction of protected content for training purposes. However, they serve a distinct purpose in advancing technological innovation and knowledge dissemination. Therefore, it is important to carefully consider whether such practices should be subject to a (copyright) exception to prevent undue harm to rightsholders while fostering innovation in AI. This would allow acts of copying for AI training purposes if a payment of equitable remuneration is observed.

[240] *Austro-Mechana* (n. 51) Para 21, 30-33
[241] *Austro-Mechana* (n. 51) Para 24
[242] *Austro-Mechana* (n. 51) Para 37
[243] *Austro-Mechana* (n. 51) Para 49-51
[244] *Austro-Mechana* (n. 51) Para 44
[245] *Austro-Mechana* (n 51) Para 54
[246] *Austro-Mechana* (n. 51) Para 51
[247] *Austro-Mechana* (n. 51) Para 54

According to the Court's interpretation of Article 5(2)(b) and its decision, reproductions can take place in *any medium*, and unless an exception applies, each step of uploading and downloading copyrighted content to the cloud from devices or media constitutes a reproduction of that content, potentially infringing on Article 2 of InfoSoc Directive. Nowadays, developers use data sources to train machine models, download and store data copies into the cloud, as well as store temporary copies of materials to improve AI learning. Thus, storing any protected work in the digital form on hard drives or cloud services[248] constitutes reproduction in the digital age.

Historically, private copying practices gave rise to the reproduction of copyright materials in physical forms, which subsequently led to the creation of exceptions in copyright law.[249] The purpose of Article 5(2)(b) is to reduce the risk of unauthorized reproduction of content within the private sphere.[250] By allowing specific uses of copyrighted materials for personal consumption meanwhile ensuring fair compensation to rightsholders, the exception seeks to address the economic interest of rightsholders.[251] Fair compensation under Article 5(2)(b) of Directive 2001/29 is triggered by the presumption, rebuttable under certain circumstances, of harm caused to rights holders, which generally entails the obligation to compensate them.[252] Member States that adhere the exception under their national law can determine the payment of an equitable compensation[253] Levies, for instance, emerged to ensure right holders receive fair compensation.[254] This operates under the assumption that rightsholders suffer harm from the private copying, particularly in response to uncontrolled nature of copying and reproducing works though sound and video recording equipment.[255]

When assessing the harm suffered by rightsholders on the context of AI, there is a rebuttable presumption that AI - and the companies behind their production - fully exploit copying and reproduction of copyrighted material and storage capacity at their disposal, similarly to how copyrighted works are explored in a private sphere for private use.[256] Furthermore, considering that copyrighted material becomes deeply ingrained within AI learning processes, the removal of such protected content becomes difficult.[257] The copyrighted material becomes deeply embedded within the models, making it difficult to unlearn the acquired patterns.[258] This

[248] Quang (n. 5) 1413

[249] Francisco Javier Cabrera Blázquez, 'Private Copying at the Crossroads' on *Who Pays for private Copying?* (Audiovisual Observatory, 2011) 8 <https://rm.coe.int/1680783bc7> accessed 15 April 2024

[250] Case C-433/20 *Austro-Mechana v Strato AG* [2022] ECLI-2021:763, Opinion of AG Hogan, para 59-60.

[251] Blázquez (n. 273) 8-9

[252] *Ibid* Para 57 and Blázquez (n. 249) 10-13.

[253] *Ibid*

[254] Martin Kretschmer (n. 227) 7

[255] Blázquez (n. 249)

[256] Case C-433/20 *Austro-Mechana v Strato AG* [2022] ECLI-2021:763, Opinion of AG Hogan, para 69

[257] Lucas and others (n. 208)

[258] Dawen Zhang and others (n. 210) 7

raises an extra concern for rightsholders since they are not receiving any financial compensation for such usage.

According to the Court, the concept of fair compensation is central to maintaining a balance, deemed essential to safeguard the financial remuneration of creators.[259] The Court has emphasized the importance of balancing the interests of rightsholders with the need to prevent undue harm resulting from uncontrolled and mass private copying and reproductions.[260] Therefore, the economic impact on rightsholders must be carefully considered in this context by ensuring fair compensation for rightsholders in AI training.

One must further consider whether an exception could be created or extended on copyright law and/or on the AI Act, given the extensive widespread use of copyrighted works on AI training and the resulting harm to rightsholders. This could potentially qualify as an exception to right of reproduction if equitable remuneration is provided to them. A great example of the rule governing mass usage of copyrighted works is the privilege granted to end-users for reproducing copyrighted material for non-commercial purposes within their private sphere, coupled with the assurance of fair compensation to rightsholders.[261] Considering this, establishing exceptions to right of reproduction for AI training purposes would empower Member States to incorporate relevant provisions into their national laws. If pursued, Member States should establish a mechanism whether they could assign organizations engaged in AI activities to collect a designated fee (which could be potentially determined based on the revenue generated from generative AI) and determine an external organization to manage and distribute a ‘compensation right’, or manage the distributions themselves.

Alternative Compensation System (ACS) refers to a “permitted-but-paid system”,[262] where usually there is Collective Management Organization (CMOs) involved which would be responsible for managing the licenses and the remuneration or compensation rights.[263] In general, CMOs are private entities entrusted with tasks such as licensing, monitoring, enforcing copyright rules, and distributing royalties to rightsholders.[264] They serve as the collective interests of their members, providing an alternative to exclusive rights management.[265] The CMO would operate as intermediaries in the market between rightsholders and interested entities,[266] by managing and collecting the sum (which could be based on the revenue generated from works generated from

[259] Case C-433/20 *Austro-Mechana v Strato AG* [2022] ECLI-2021:763, Opinion of AG Hogan, para 59

[260] Case C-433/20 *Austro-Mechana v Strato AG* [2022] ECLI-2021:763, Opinion of AG Hogan, para 58

[261] Joao Pedro Quintais, *Copyright in the age of online access: Alternative compensation systems in EU copyright law* (n. 45) 52

[262] *Ibid* 86

[263] *Ibid* 91-93

[264] *Ibid* 95

[265] *Ibid* 91-93

[266] *Ibid* 93

AI-produced works) and further, distributing among rightsholders in order to ensure a fair compensation among rightsholders.[267] An ACS serves as a unified platform interested parties to access copyrighted works online, with funds collected distributed to rightsholders. CMOs play a pivotal role in this system, ensuring tighter control over access to copyrighted works and overseeing the collection and distribution of compensation among rightsholders. By implementing a payment fee based on the revenue generated and collected from AI companies, CMOs can effectively manage copyright works and ensure fair compensation for creators. Such an alternative system will be further analyzed in the next chapter.

[267] *Ibid* 92

CHAPTER 5 – Alternative Compensation System and Copyright

The Alternative Compensation System (ACS) is a copyright management approach where the obligation to pay royalties is linked not directly to the use of copyrighted works, but to the acquisition or use of related goods and services.[268] This system proposes a legal framework that allows individuals to use works online without explicit authorization from rightsholders, provided they are compensated.[269] The objective of ACS is to create a model where access to copyrighted works online is regulated through compensation,[270] offering an alternative to the current paradigm of copyright which is bounded by exclusivity, individual management, and the potential for stringent enforcement against individuals.[271]

5.1. Key attributes:

Several attributes help define the scope and characteristics of an ACS. These attributes include *subject matter scope, substantive rights, compensation type, management system, compensation target, and burden of compensation.*[272] The composition of each attribute varies depending on the type of ACS or legalization proposals. Nonetheless, these attributes are useful in illustrating the benefits, costs, and their impact on rightsholders.[273]

About the *subject matter*, an Alternative Compensation System (ACS) can be applied to any digital content that is protected by copyright or related rights.[274] It focuses on allowing the use of copyrighted material on digital networks, such as the Internet. *Substantive rights* refer to online activities like downloading, uploading, streaming, reproduction (copying), public communication (sharing with the public), making works available online, distributing works online, and creating adaptations (derivative works). These rights involved in these activities are clearly stated or directly specified or understood through their association with a broader category of rights.[275] Furthermore, the types of compensations in ACS models can vary. It depends on the context in which it is involved. Rightsholders can receive different types of *compensation*, including payment, tariff, royalty, license fee, remuneration, compensation, levy,

[268] Christian W. Handke, João Pedro Quintais, and Balázs Bodó, 'Truce in the Copyright War? The Pros and Cons of Copyright Compensation Systems for Digital Use' (2018) 15/2 Review of Economic Research on Copyright Issues <https://ssrn.com/abstract=3311019> accessed 5 July 2024

[269] João Pedro Quintais, 'Alternative Compensation Models for Large-Scale Non-Commercial Online Uses' (2015). (ALAI International Congress, Bonn, June 2015) 1 <https://ssrn.com/abstract=2625492> accessed 5 July 2024.

[270] Quintais, *Copyright in the age of online access: Alternative compensation systems in EU copyright law* (n 45) 138

[271] Handke, Quintais, and Balázs (n. 268) 24

[272] Quintais, *Copyright in the age of online access: Alternative compensation systems in EU copyright law* (n 45) 138-139

[273] *Ibid*

[274] *Ibid* 139

[275] *Ibid* 140-143

contribution, or tax.[276] The type of compensation depends on the nature of the being authorized. For instance, "license fee" or "royalty" are common in the Voluntary Collective Licensing model, while "taxes" and "rewards" are associated with State System proposals.[277] *Management system* relates to the calculation, collection, and distribution of compensation.[278] There are two types of compensation: one generated by an ACS and the other by the usage of copyrighted material, which involves user payment. ACS-generated compensations are typically calculated based on objectives such as providing fair compensation, addressing market failures, or incentivizing creation and access to work.[279] The *compensation target* refers to the specific goods or services subject to payment obligations aimed at providing remuneration to copyright owners.[280] For example, in AI training, the targeted constitutes copyright material, and compensation could be a percentage of the revenue generated by the AI. Direct and indirect taxes are also used in State Systems to generate this compensation. Finally, the *burden of compensation* refers to the party responsible for payment liability, usually the user or an intermediary.[281]

Quintais suggests that, regardless of the ACS model, it operates as a single-interface system through which users can decide or are competed to pay for online content usage rights, with the funds distributed to rightsholders.[282] Such an ACS system could be implemented in the context of AI, which involves copying and reproducing digital works. An ACS would manage access to copyrighted digital works on the condition of compensation. This compensation, managed by a Collective Rights Management organization (CMO) or specialized agencies depending on the ACS model, aims to recoup lost profits incurred from using these materials for AI training and ensure these profits are distributed to the rightsholders.

5.2. Type of ACS models:

The private copying limitation of Article 5 (2)(b) of the InfoSoc Directive, along with statutory licenses, serves as an inspiration for ACS models.[283] This exception allows users to reproduce copyrighted material for non-commercial use within their private sphere while ensuring fair compensation to rightsholders. It allows end-users to reproduce copyrighted material for non-commercial purposes within their private sphere while ensuring fair compensation to rightsholders.[284] It addresses challenges arising from technological disruptions

[276] *Ibid* 143
[277] *Ibid*
[278] *Ibid* 144
[279] *Ibid* 144-146
[280] *Ibid* 146-147
[281] *Ibid*
[282] *Ibid* 147-148
[283] *Ibid* 24
[284] It refers to Article 5 (2)(b) of the InfoSoc Directive

and copyright enforcement issues, emphasizing remunerated access over exclusivity, a core principle of ACS.[285]

Generally, the exclusive right grants rightsholders control over how their works are used. However, this right is not absolute and can be limited by law or contract.[286] These limitations can affect the exercise of the right without altering its nature.[287] In this regard, the concept of copyright 'elasticity' forms the essence of the Alternative Compensation System (ACS).[288] Such elasticity does not aim to alter the essence of the right; rather, it spans from unrestricted individual exercise to gradually constraining collective rights management.[289]

Within the framework of ACS, various models impose differing levels of restrictions on the utilization of exclusive rights, thereby impacting the implementation and enforcement of copyright law.[290] Voluntary Collective Licensing and Extended Collective Licensing (ECL) are models where rightsholders voluntarily engage with Collective Management Organizations to oversee their rights collectively.[291] Despite delegating certain management functions to collective organizations, rightsholders still retain their exclusive rights.[292] Conversely, Mandatory Collective Management and Legal Licenses entail statutory limitations on rights exercise, compelling rightsholders to participate in collective management schemes or adhere to specific legal provisions. Although these restrictions may curtail the autonomy of individual rightsholders to some extent, they do not fundamentally alter the nature of their exclusive rights.[293] Furthermore, the State System differs somewhat from the previous models as it operates outside the traditional copyright framework.[294] In this model, rightsholders and creators are compensated through various tax or funding schemes designed to subsidize or reward them.[295] However, it may still involve a designated ACS to help manage these rewards. The idea of the exclusive right having elasticity refers to its ability to accommodate different levels of control and management without losing its inherent exclusivity.[296] This means that even as rightsholders opt into Collective Management Schemes or adhere to State Systems requirements, they still retain their exclusive authority over their works.[297] The exclusive right can stretch from

[285] Joao Pedro Quintais, *Copyright in the age of online access: Alternative compensation systems in EU copyright law* (n. 45) 84
[286] *Ibid* 93
[287] *Ibid*
[288] *Ibid*
[289] *Ibid*
[290] *Ibid* 91
[291] *Ibid*
[292] *Ibid*
[293] *Ibid*
[294] *Ibid* 87
[295] *Ibid*
[296] *Ibid* 89
[297] *Ibid* 90

individual management to more limiting collective rights management schemes, allowing for a flexible approach to copyright regulation while preserving the fundamental nature of exclusivity.

Different types of models have been proposed. However, in the context of AI, the first two—Voluntary Collective Licensing and Extended Collective Licensing (ECL)—are not ideal because rightsholders can opt-out at any time.[298] Although this opt-out option is a fundamental aspect of copyright law, it presents significant challenges for AI. As detailed in Chapter 3, especially in Subchapter 3.8, it is nearly impossible to "untrain" AI models once they have bene trained on certain data. This means that AI cannot simply be reverted to a prior state, becoming challenging for rightsholders to completely opt out their works from AI trained models. It would be more suitable to use a model that does not allow opting out, as it would create more stability in their business and might reduce the risk of legal disputes.

Regarding the Mandatory Collective Management Model, it is more restrictive compared to other models. It prevents rightsholders from directly exploiting their works by legally or contractually transferring the exercise of their rights to a Collective Management Organization (CMO), with no option to opt-out.[299] Furthermore, since participation in this system is not mandatory for users,[300] it is not a suitable model to consider in the context of AI, as it makes engagement difficult. Users are not compelled to register and participate as rightsholders, making it impossible to ensure fair remuneration since not all parties would be involved. Further, an obligation to license under Mandatory Collective Management does not differ from the practices of Voluntary ones.[301] Regarding the Legal License model, while it may be feasible for the law to determine the scope and subject matter, designate the Collective Management Organization (CMO), and identify the intermediary debtors of remuneration, and including setting tariffs under the licenses,[302] this thesis argues that in the current scenario of mass data usage in AI training, assessing individual licenses is impractical (chapter 4. 4.1 licensing copyright material for AI training).

5.3. ACS for fair compensation in AI:

The State System operates as a legal framework outside the traditional copyright framework, facilitating mass online use while ensuring some level of remuneration for rightsholders.[303] Instead of relying on direct negotiations between users and rightsholders, the State System entails government-managed compensation schemes, typically funded through the payment of fees. The aim is to create a revenue pool from consumers, which is then distributed

[298] *Ibid* 93, 100-113
[299] *Ibid* 113-115
[300] *Ibid* 122
[301] *Ibid*
[302] *Ibid* 126-129
[303] *Ibid* 132

among the rightsholders.[304] Moreover, whether the ACS is governmental or non-governmental, there would be a platform where rightsholders can register their works.[305] Unlike other alternative models such as Voluntary Collective Licensing and Extended Collective Licensing (ECL), Mandatory Collective Management, and Legal Licensing, the State System model operates independently of copyright protection and is governed by the State.[306] This means that the State must determine requirements regarding fair compensation.

Fair compensation for rightsholders could be determined based on the monetary value of the training data, meaning a fee would reflect the financial worth attributed to the use of copyrighted material in AI training. This could involve a pre-determined tax rate, where a percentage of the revenue earned from AI outputs that rely on copyrighted material is allocated to rightsholders. Compensation could also be determined based on extent of use. Similar to the royalty models in the digital music and video industries where rightsholders receive payments based on the revenue generated by these models,[307] AI companies could similarly aggregate revenues derived from AI-generated products and services that utilize copyrighted material in their training datasets. Given that providers are now required to disclose information about the content used in their training datasets,[308] a compensation model could be implemented that calculates payments based on the frequency or extent of use of copyrighted materials within these datasets.[309] The total revenue would be accumulated based on the revenue generated by AI models that have utilized copyrighted materials during their training process, and AI companies would contribute to a centralized revenue pool managed by a CMO. Nevertheless, while the proposed compensation fees for the use of copyrighted material in AI training suggest potential frameworks based on the extent of use and revenue managed by a CMO, further research is needed to assess the viability and fairness of these approaches, which is beyond the scope of this thesis.

[304] Katherine L. McDaniel, 'Accounting for Taste: An Analysis of Tax-and-Reward Alternative Compensation Schemes' (2007) 9, Tulane Journal of Technology and Intellectual Property.
[305] *Ibid* 255
[306] Joao Pedro Quintais, *Copyright in the age of online access: Alternative compensation systems in EU copyright law* (n. 45) 90
[307] On digital platforms like Spotify and YouTube, royalties are calculated based on the overall revenue generated and how frequently content is accessed. Instead of paying directly for each individual song or video, these platforms collect all the revenue they earn from users—such as subscription fees and ad revenue—into a large pool. This pool represents the total amount of money that will be shared among copyright holders. Revenue is then divided among these copyright holders based on how often their content was accessed within each pool. By using this model, platforms can efficiently manage and distribute large amounts of revenue across a vast number of works, ensuring that each rightsholder receives a fair share based on the popularity of their content. Junwei Deng, 'Computational Copyright: Towards A Royalty Model for Music Generative AI' (University of Illinois Urbana-Champaign, 2024) < https://ar5iv.labs.arxiv.org/html/2312.06646> accessed 6 August 2024.
[308] Artificial Intelligence Act, Article 53 (1) (d).
[309] Deng (n. 307)

With advances in technology and the widespread use of copyrighted materials, this model could serve as a robust mechanism in the AI context, compelling companies to adhere to state-mandated fair compensation requirements. This system is applicable to both commercial and non-commercial usage of copyright works.[310] Rightsholders would need to opt in and register their works with a government agency or Collective Management Organization (CMO), without the possibility of opting out. Funding for compensating rightsholders would be generated through fees or revenues derived from AI, and distributed by the agency through periodic pool or annual payments.[311] Afterwards, the creative images generated by AI could fall under the public domain, which would remove the commercial characteristic of commercial usage.[312]

Systems similar to the proposed approach have been successfully implemented in countries such as Norway, Spain, and Finland, where compensation for harm is funded through the State Budget.[313] In Norway, creators receive fair compensation based on usage studies that determine the extent of the harm.[314] Annual grants are allocated by the government, funded from the state budget[315] and distributed to rights holders through an umbrella 'Norwaco'.[316] Finland introduced a similar financing system in 2015,[317] while Spain adopted such a model in December 2012, with compensation amounts determined by the Ministry of Culture and paid annually to competent organization from the State budget.[318]

State systems could offer advantages in the context of AI. As AI relies on large datasets that include copyright works, such an alternative model would ensure that creators are fairly compensated for using their works in AI training. By providing a structured way to address the harm caused by the unauthorized use of copyrighted materials, these systems help to balance the interests of rightsholders on the one hand and the innovation and development of AI on the other. Additionally, this approach could be more effective than current individual licensing models since it streamlines the process for users and rightsholders. Using this alternative system for compensation simplifies access to copyrighted works and reduces the administrative burden associated with negotiating and managing multiple individual licenses.

[310] *Ibid* 136-138

[311] *Ibid*

[312] *Ibid*

[313] Digital Europe, 'Private Copying: Assessing Actual Harm and Implementing Alternative Systems to Device-Alternative Systems to Device-Based Copyright Levies' (Brussels, June 2015) 6 <https://cdn.digitaleurope.org/uploads/2019/01/Private%20Copying%20Assessing%20harm%20and%20implementing%20alternatives%20to%20copyright%20levies.pdf> accessed 3rd July 2024.

[314] *Ibid*

[315] *Ibid*

[316] Norwaco oversees the copyrights for TV, film, and music content for individuals in Norway as well as for foreign licensees. <https://norwaco.no/en/about-norwaco> accessed 3rd July 2024.

[317] Digital Europe (n. 313) 6

[318] *Ibid*

While the ACS offers many advantages, questions may arise regarding its implementation. For instance, how will the system operate across different countries with varying copyright laws? or how will fair compensation be calculated and distributed to copyright holders? Despite these challenges, the ACS encourages broader compliance, as AI companies would be compelled to adhere to state-mandate requirements, and rightsholders would be more likely to adhere to a straightforward, centralized system, thereby increasing the likelihood of providing proper compensation to creators addressing the financial harm caused by widespread unauthorized use. Moreover, this model can effectively mitigate uncertainty and minimizes the likelihood of disputes by ensuring fair compensation for the use of works in AI training, regardless of the scale or resources of the involved AI companies.

6. CONCLUSION

The success of generative AI relies on large datasets, advanced algorithms, and robust infrastructure. Generative AI, mainly through Stable Diffusion’s lens, illustrates both advancements and challenges in AI technology. While generative models demonstrate remarkable capabilities in generating high-quality content, they also raise copyright concerns due to their reliance on extensive and sourced datasets.

The training processes of AI directly affect the exclusive rights of reproduction. This thesis mainly focuses on the input phase of generative AI models, addressing two main issues: sourcing and utilizing copyrighted material within datasets and making temporary copies during AI training. Generative AI models, such as Stable Diffusion, heavily rely on vast datasets often scraped from the internet, which significantly risks infringing on right of reproduction. This challenges the traditional model of copyright enforcement, as AI-driven processes often rely on large datasets, including copyrighted material, for training and development. Issues arise when copyrighted content is inadvertently included, as seen in lawsuits by Getty Images and artists against companies like Stability AI. These lawsuits highlight the difficulty in distinguishing and proving the use of copyrighted material in AI-generated content.

The European copyright framework, including the InfoSoc and CDSM Directives, attempts to address these issues by providing specific exceptions and limitations for activities like Text and Data Mining. However, the effectiveness of these provisions is still limited by the complexities of AI training and the massive volumes of data involved. The introduction of the AI Act seeks to enhance transparency and ensure compliance with copyright rules. Yet, the feasibility of 'unlearning' data by AI systems remains a formidable challenge, making it difficult for right-holders to opt-out. Moreover, securing licenses for large-scale data use in AI training underscores the need for a more streamlined approach to copyright management. The CJEU ruling on the ‘private copying’ exception in *Austro-Mechana* vs. *Strato AG* marks a significant development in adapting copyright law to the digital age. The exception for right of reproduction under in Article 5(2)(b) of the InfoSoc Directive, which is based on fair compensation, aims to protect the economic interests of rightsholders, ensuring they are not disadvantaged by technological advancements. Despite these provisions, resolving copyright issues in AI training data remains complex and requires further legal exploration, particularly within the European Union's framework.

The need for significant transformation in copyright law has become increasingly apparent. The evolution of copyright perspectives has accelerated due to the pervasive use of online data, challenging the traditional notion of copyright as a mechanism for rightsholders to closely monitor and monetize each instance of their works' usage. The mass use of protected works for private purposes, followed by the exception under Article 5(2)(b) of the InfoSoc

Directive, serves as a strong example of balancing the interests of rightsholders with the potential for the use of copyright material, provided fair compensation is ensured.

The advent of AI introduces a new paradigm, as AI is increasingly recognized as a tool to serve the public good, foster innovation and creativity, and facilitate the dissemination of knowledge. As a result, copyright law may need to adapt to these shifts. Instead of solely focusing on strict enforcement of individual rights, there is a growing recognition of the need to balance the interests of rightsholders and follow technology development. This thesis explores alternative approaches to copyright management, including frameworks that allow the use of copyrighted materials for AI training while ensuring fair compensation for creators. One such approach is the implementation of an Alternative Compensation System (ACS) in the context of AI, offering an alternative to traditional copyright licensing schemes. Given that AI relies on large datasets, often including copyrighted works, such model would ensure that creators are fairly compensated for the use of their works in AI training. By providing a structured way to address the harm caused by unauthorized use of copyrighted materials, ACS can balance the interests of rightsholders with the innovation and development of AI. An ACS would regulate access to copyrighted digital works by requiring compensation, which would be managed by a Collective Rights Management organization (CMO) or specialized agencies. This system aims to recover profits lost due to the use of copyrighted materials in AI training and ensure that these profits are fairly distributed to rightsholders.

7. BIBLIOGRAPHY

Primary sources

Legislation

Berne Convention for the Protection of Literary and Artistic Works (adopted 9 September 1886, entered into force 5 December 1887, as revised at Paris 24 July 1971, and amended in 1979) 1161 UNTS 3

Charter of Fundamental Rights of the European Union [2012] C326/02

Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L 167/10

Directive (EU) 2019/790 of 17 April 2019 on Copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] OJ L 130/92

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 [2024] OJ L 1689/44

Case law

The United States

Getty Images (US) INC v Stability AI INC [2023] US District Court for the District of Delaware Case 1:23-cv-00135-UNA 1-36

Sarah Andersen, Kelly McKernan, Karla Ortiz vs Stability AI LTD, Stability AI INC, Midjourney INC, DeviantArt INC [2023] United States District Court Northern District of California San Francisco Division Case 3:23-cv-00201

European Union

Austro-Mechana vs. Strato AG [2022] Court of Justice of the European Union Case C-433/20 ECLI:EU:C:2022:217

Infopaq International A/S v Danske Dagblades Forening [2009] Court of Justice of the European Union C-5/08 EU:C:2009:465

Infopaq International A/S v Danske Dagblades Forening ('Infopaq II', 2012) Court of Justice of the European Union C-302/10 ECLI:EU:C:2012:16

Padawan SL v Sociedad General de Autores y Editores de España (SGAE) Court of Justice of the European Union C-467/08 ECLI:EU:C:2010:620

Secondary sources

Abduljawad M, Alsalmani A, 'Towards Creating Exotic Remote Sensing Datasets Using Image Generative AI' (IEE, 2022)

Agbaji, D., Alhassan, J., Galley, J., Kousari, R., Lund, B. D., Mannuru, N. R., Ogbadu-Oladapo, L., Pohboon, C. O., Saurav, S. K., Shahriar, S., Srivastava, A., Teel, Z. A., Tijani, S., Tummuru, S. P., Uppala, S., Vaidya, P., and Wang, T., 'Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development. Information Development' (2023) 0/0 SageJournals

Allen G, 'Understanding AI Technology' (2020) 7 <https://www.ai.mil/docs/Understanding%20AI%20Technology.pdf>

Alzubaidi, L., Zhang, J., Humaidi, A.J. et al., 'Review of deep learning: concepts, CNN architectures, challenges, applications, future directions.' (2021) 8/53 Journal of Big Data

Anderson A, Rainie L, 'Solutions to address the AI's anticipated negative impacts' (2018) <https://www.pewresearch.org/internet/2018/12/10/solutions-to-address-ais-anticipated-negative-impacts/>

Aydin O, Karaarslan E, 'Is ChatGPT leading Generative AI? What is Beyond Expectations?' (2023) 11(3) Academic Platform Journal of Engineering and Smart Systems

Balazs B, Handke C, Quintais P, 'The Economics of Copyright Compensation systems for Digital Use' (SERCI Annual Congress 2013, Paris, 8th and 9th July 2013) <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f963447a040555b50c5081020f35dd1e381f0fff>

Balle, B., Carlini, N., Hayes, J., Ippolito, D., Jagielski, M., Nasr, M., Sehwag, V., Tramèr, F., and Wallace, E., 'Extracting Training Data from Diffusion Models' (2023) https://arxiv.org/abs/2301.13188

Beaumont, R., Cherti, M., Coombes, T., Crowson, K., Gordon, C., Jitsev, J., Kaczmarczyk, R., Katta, A., Kundurthy, S., Mullis, C., Schuhmann, C., Schmidt, L., Schramowski, P., Vencu, R., Wightman, R., and Wortsman, M., 'LAION-5B: An open large-scale dataset for training next generation image-text models' (2022) <https://proceedings.neurips.cc/paper_files/paper/2022/file/a1859debfb3b59d094f3504d5ebb6c25-Paper-Datasets_and_Benchmarks.pdf>

Blázquez F, 'Private Copying at the Crossroads' on Who Pays for private Copying? (Audiovisual Observatory, 2011)

Bonadio E, Dinev P, McDonagh L, 'Can artificial intelligence infringe copyright? Some reflections' in Ryan Abbott (ed), Research Handbook on Intellectual Property and Artificial Intelligence (Edward Elgar Publishing Limited 2022)

Borghi M, Karapapa S, 'Copyright and Mass Digitization' (Oxford University Press, online edn, 2013)

Bourtoule L, Chandrasekaran V, Choquette-Choo CA, Jia H, Lie D, Papernot N, Travers A, and Zhang B., 'Machine Unlearning' (2021) <doi: 10.1109/SP40001.2021.00019>

Brizuela R, Garrido-Merchan E, 'ChatGPT is not all you need. A State of the Art Review of large Generative AI models' (2023) <https://arxiv.org/abs/2301.04655>

Bodo B, Handke C, Quintais J, 'Truce in the Copyright War? The Pros and Cons of Copyright Compensation Systems for Digital Use' (2018) 15/2 Review of Economic Research on Copyright Issues

Bulayenko O, Frosio G, Geiger C, 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU' in Concepción Saiz García and Raquel Evangelio Llorca (eds), Propiedad intelectual y mercado único digital europeo (Valencia, Tirant lo blanch, 2019), Centre for International Intellectual Property Studies (CEIPI) Research Paper No. 2019-08

Butler, B., Carroll, M., Contreras, J., Craig, C., Flynn, S., Guibault, L., Jaszi, P., Jütte, B., Katz, A., Margoni, T., Quintais, J., Sag, M., Samberg, R., Sasson, O., Schirru, L., Senftleben, M., Souza, A., and Tur-Sinai, O., 'Legal reform to enhance global text and data mining research: Outdated copyright laws around the world hinder research' (2022) 378/6623 Science

Chen L, Cai J, Nah F, Siau K, Zheng R, 'Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration' (2023) 25 (3) Journal of Information Technology Case and Application Research

Chau, D., Hoover, B., Lee, S., Li, K., Park, H., Peng, S., Strobelt, H., Wang, Z., Wright, A., and Yang, H., 'Diffusion Explainer: Visual Explanation for Text-to-image Stable Diffusion' (2023) <https://arxiv.org/abs/2305.03509>

Cellary W, Walczak K, 'Challenges for higher education in the era of widespread access to Generative AI' (2023) 9(2) Economics and Business Review

Committee on Legal Affairs, resolution on intellectual property rights for the development of artificial intelligence technologies (2020) Report 2020/2015(INI)

Committee of Legal Affairs, 'Report on intellectual property rights for the development on artificial intelligence technologies' (Report - A9-0176/2020)

'Common Crawl' <https://commoncrawl.org/>

Contractor D, Haines J, Hecht B, Hines C, Lee J, Li H, McDuff D, Vincent N., 'Behavioral Use Licensing for Responsible AI' in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22) (Association for Computing Machinery, New York 2022)

Chave L, Fisher M, Jacob R, Guidebook to intellectual property (7th edition, 2022)

‘DALL-E‘ <https://openai.com/research/dall-e>

Dehouchea N, Dehoucheb K, ‘What’s in a text-to-image prompt? The potential of stable diffusion in visual arts education’ (2023) 9/el6757 Heliyon

Digital Europe, ‘Private Copying: Assessing Actual Harm and Implementing Alternative Systems to Device-Based Copyright Levies’ (Brussels, June 2015) <https://cdn.digitaleurope.org/uploads/2019/01/Private%20Copying%20Assessing%20harm%20and%20implementing%20alternatives%20to%20copyright%20levies.pdf>

Deng J, ‘Computational Copyright: Towards A Royalty Model for Music Generative AI’ (University of Illinois Urbana-Champaign, 2024) <https://ar5iv.labs.arxiv.org/html/2312.06646>

Ducato E, Strowel A, ‘Ensuring Text and Data Mining: Remaining issues With the EU Copyright Exceptions and Possible Ways Out’ (2021) Intellectual Property Review

Electronic Frontier Foundation, ‘The Three-Step Test’ (study) <https://www.eff.org/files/filenode/three-step_test_fnl.pdf>

European Commission, Directorate-General for Communications Networks, Content and Technology, ‘Study on copyright and new technologies: copyright data management, and artificial intelligence’ (2012) Publications Office of the European Union

European Commission, Directorate-General for Communications Networks, Content and Technology, ‘Study on copyright and new technologies: copyright data management and artificial intelligence’ Publications Office of the European Union (2022)

European Commission ‘Artificial Intelligence for Europe’ (Communication) COM(2018) 237 final.

European Digital Rights, ‘Private Copying levies’ (Edris) <https://www.edri.org/files/privatecopyinglevies.pdf>

European Parliament, ‘Resolution of 20 October 2020 on Intellectual Property Rights for the Development of Artificial Intelligence Technologies’ (2020/2015(INI))

European Parliament, ‘The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspect’ (PE 604.942, 2018)

European Parliament (PT) ‘Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI’ COM (20231206IPR15699)

Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P., ‘Generative AI’ (2024) 66 Business Information System Engineering <https://doi-org.tilburguniversity.idm.oclc.org/10.1007/s12599-023-00834-7>

Geiping, J., Goldblum, M., Goldstein, T., Singla, V., and Somepalli, G., ‘Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models’ (2023)

<https://openaccess.thecvf.com/content/CVPR2023/papers/Somepalli_Diffusion_Art_or_Digital_Forgery_Investigating_Data_Replication_in_Diffusion_CVPR_2023_paper.pdf>

Global Partnership on Artificial Intelligence, 'GPAI Compromise Proposal' (2023) <https://www.openfuture.eu/wp-content/uploads/2023/12/231206GPAI_Compromise_proposalv4.pdf>

Ginart AA, Guan MY, Valiant G, and Zou J., 'Making AI Forget You: Data Deletion in Machine Learning' (33rd Conference on Neural Information Processing Systems NeurIPS, 2019) <https://proceedings.neurips.cc/paper_files/paper/2019/file/cb79f8fa58b91d3af6c9c991f63962d3-Paper.pdf>

Hakimshafaei M, 'Survey of Generative AI in Architecture and Design' (2023) 16 <https://escholarship.org/uc/item/47x6k9j8>

Handke C, 'Compensation Systems for Online Use' (2020) <https://doi.org/10.1007/978-3-030-44850-9_15>

Häuselmann A, 'Disciplines of AI: An overview of approaches and Techniques' (2021) 48 in Bart Custers and Edward Fischer Villaronga (ed), Law and Artificial Intelligence: Regulating AI and Applying in Legal practice (T.M.C Asser Press, 2022)

Hilty R, Hoffmann J, Scheuerer S, 'Intellectual Property Justification for Artificial Intelligence' (2020) Draft chapter. Forthcoming in: J.-A. Lee, K.-C. Liu, R. M. Hilty (eds.), Artificial Intelligence & Intellectual Property (Oxford University Press, 2020), Max Planck Institute for Innovation & Competition Research Paper No. 20-02

Hammon C, James, Rajendra A, and Trigg R, 'Generative AI: can intellectual property infringements in training data be avoided?' (2023) Lexology

Höppner, T, Hornschuh M, Uppenbrink K, 'Our call for safeguards Around Generative AI' (Initiative Irheberrecht, 2023) <https://urheber.info/media/pages/diskurs/call-for-safeguards-around-generative-ai/c93a5ab197-1681904353/final-version_authors-and-performers-call-for-safeguards-around-generative-ai_19.4.2023_12-50.pdf>

Hoang T, Liu Y, Lu Q, Staples M, Xia B, Xing Z, Xu X, Zhang D, and Zhu L., 'Tag your Fish in the Broken Net: A Responsible Web Framework for Protecting Online Privacy and Copyright' (2013) <https://doi.org/10.48550/arXiv.2310.07915>

Intellectual Property Office, Consultation outcome Artificial intelligence and intellectual property: call for views (UK, 2021) <https://www.gov.uk/government/consultations/artificial-intelligence-and-intellectual-property-call-for-views>

IRIS plus 2011-4, Who Pays for Private Copying? in Susanne Nikoltchev (ed), European Audiovisual Observatory (Strasbourg, 2011)

Jarrahi M, 'Artificial intelligence and the future of work: Human-AI symbiosis in organization decision making' (2018) Science Direct 61/4

Johansen D, Juliussen B, and Rui J., 'Algorithms that forget: Machine unlearning and the right to erasure' (2023) 51/105885 Computer Law & Security Review

Keller P, 'Protecting creatives or impeding progress? Machine learning and the EU copyright framework' (Institute for Information Law (IViR), 20 February 2023)
—— 'Generative AI and copyright: convergence of opt-outs?' (Kluwer Copyright Blog, 2023)

Kretschmer M, 'Private Copying and fair compensation: a comparative study of copyright levies in Europe' (An Independent Report Commissioned by the UK Intellectual Property Office, (2011) <https://microsites.bournemouth.ac.uk/cippm/files/2014/11/levy-final-report-jun-2011.pdf>

Kaur N, Panda S, 'Exploring the viability of ChatGPT as an alternative to traditional chatbot systems in library and information centers' (2023) 40(3) Library Hi Tech News

Kweon In, Zhang C, Zhang C, Zhang M, 'Text-to-image Diffusion Models in Generative AI: A Survey' (2023) <https://arxiv.org/abs/2303.07909>
'LAION-5B' <https://laion.ai/blog/laion-5b/>

Mariani M, 'Generative Artificial Intelligence and Innovation: Conceptual Foundations' (2022)

Matulionyte R, Selvadurai N, 'Reconsidering creativity: copyright protection for works generated using artificial intelligence' (2020) 15/7 Journal of Intellectual Property Law & Practice

Margoni T, 'Artificial Intelligence, Machine learning and EU copyright law: who owns AI?' (2018) CREATe Working Paper

McDaniel K, 'Accounting for Taste: An Analysis of Tax-and-Reward Alternative Compensation Schemes' (2007) 9, Tulane Journal of Technology and Intellectual Property

'Midjourney' <https://www.midjourney.com/explore>

National Research Council, et al. The digital dilemma: Intellectual property in the information age. (National Academies Press, 2000)

'Open Ai' <https://openai.com/blog/chatgpt>

Pila J and Torremans P, European Intellectual Property Law (2nd edition, Oxford University Press 2019)

Quang J, 'Does training AI violate copyright law? (2021) 36/4 Berkeley Technology Law Journal

Quintais P, Alternative Compensation Models for Large-Scale Non-Commercial Online Uses' (2015). (ALAI International Congress, Bonn, June 2015)

—— Copyright in the age of online access: Alternative compensation systems in EU copyright law (40, Law International BV, 2017) 3

—— 'Generative AI, Copyright and the AI Act' (Kluwer Copyright Blog, 9 May 2023)

Rosati E, ‘Copyright and the Court of Justice of European Union’ (Oxford, 2019)

—— ‘Copyright as an obstacle or an enabler? A European perspective on text and data mining and its role in the development of AI creativity’ (2019) 27(2) Asia Pacific Law Review

Roos J, ‘Artificial intelligence: Copyright & Consequences’ (2023) 20-22 <https://studenttheses.uu.nl/handle/20.500.12932/44493>

Sag M, ‘Copyright safety for Generative AI’ (2023), Forthcoming in the Houston Law Review 8

Sætra H, ‘Generative AI: here to stay, but for good?’ (2023) 75/102372 Technology in Society

Saadat M, Shuaib M, ‘Advancements in Deep Learning Theory and Applications: Perspective in 2020 and beyond’ in Marco Antonio Aceves-Fernandez (ed) Advances and Applications in Deep Learning (IntechOpen, 2020)

Simopoulou A, ‘Text and Data Mining under EU Copyright law’ (2020) <https://repository.ihu.edu.gr/xmlui/bitstream/handle/11544/29743/Text%20and%20Data%20Mining%20under%20EU%20Copyright%20Law.pdf?sequence=1>

‘Stable Diffusion’ <https://stablediffusionweb.com/>

‘Spawning.ai’ <https://spawning.ai/>

Torres I, ‘Copyright implications of the use of generative AI’ (2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4531912>

US Patent and Trademark Office (USPTO), ‘Public Views on Artificial intelligence and Intellectual Property (2020) AI-Report_2020-10-07 <https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf>

Vesala J, ‘Developing Artificial Intelligence-Based Content Creation: Are EU Copyright and Antitrust Law Fit for Purpose?’ (2023) 54 IIC <https://doi.org/10.1007/s40319-023-01301-2>

Wiggers K, ‘Spawning lays out plans for letting creators opt out of generative AI training’ (TechCrunch, 3 May 2023)

Yin L, ‘Copyright Infringement in AI-Generated Artworks’ (2024) <https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9158262&fileOId=9158279>

Zirpoli C, ‘Generative Artificial Intelligence and Copyright Law’ (2023) Congressional Research Service (CRS) <https://crsreports.congress.gov/product/pdf/LSB/LSB10922>