

Copyright Infringement & AI: A Case Study of Authors Guild v. OpenAI and Microsoft

Maria Alexandra Mărginean

Disclaimer: Any opinions expressed in this article are those of the authors and not necessarily those of 4iP Council.

1. Introduction	2
2. Authors Guild V OpenAI	4
3. Copyright Infringement at the Input Level – US law	9
4. Copyright infringement at the input level – EU law	13
5. Conclusion	20

1. Introduction

Just five days after its release in late 2022, ChatGPT-3.5 achieved 1 million users.¹ Since then OpenAI—the U.S. based artificial intelligence (AI) research and deployment company that developed ChatGPT—has experienced exponential growth, situating the company amongst the most rapidly expanding technology companies in history.² The impressive success of the creator of ChatGPT is based on the research and development of generative AI models. Such models are known for creating content or data similar to that which is used to train them. The produced output is based on the input (the prompt³) of a user.⁴ Outputs can be e.g., code (like OpenAI Codex), images (such as, DALL-E, Midjourney, Stable Diffusion, DeviantArt), audio (for example, AudioCraft, Soundraw), synthetic data (MOSTLY.AI, MDClone), or text (like OpenAI GPT (Generative Pre-trained Transformer), BloombergGPT, LLaMa, Bard).

ChatGPT is a large language model (LLM) which is trained to produce human-like text.⁵ To generate new content mimicking human creativity, this AI algorithm must be trained on large data sets.⁶ The content produced are known as outputs. ChatGPT outputs are often plausible, however over half of its answers are false or untruthful. This is because ChatGPT was “trained to predict the next word on a large dataset of Internet text, rather than to safely perform the language task that the user wants.”⁷

Following its success obtained by having the fastest-growing user base, questions began to circulate regarding the type of data used to train generative AI models. Such questions arose because the training sets used to train LLMs, including ChatGPT, consist of a wide range of

¹ ‘100+ Incredible ChatGPT Statistics & Facts in 2024 | Notta’ <<https://www.notta.ai/en/blog/chatgpt-statistics>> accessed 29 February 2024.

² ‘OpenAI on Track to Hit \$2bn Revenue Milestone as Growth Rockets’ <<https://www.ft.com/content/81ac0e78-5b9b-43c2-b135-d11c47480119>> accessed 29 February 2024.

³ Prompts are instructions given to an LLM for generating content. See Daniel Gervais, ‘Humans as Prompt Engineers’ (*Kluwer Copyright Blog*, 14 June 2023) <<https://copyrightblog.kluweriplaw.com/2023/06/14/humans-as-prompt-engineers/>> accessed 10 January 2024.

⁴ Andrés Guadamuz, ‘A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs’ (26 February 2023), p. 5 <<https://papers.ssrn.com/abstract=4371204>> accessed 8 November 2023.

⁵ ‘What Is ChatGPT? | OpenAI Help Center’ <<https://help.openai.com/en/articles/6783457-what-is-chatgpt>> accessed 8 December 2023.

⁶ Datasets are defined as collections of data, which are used in training and testing the algorithms. See ‘Datasets Definition | Encord’ <<https://encord.com/glossary/datasets-definition/>> accessed 11 March 2024.

⁷ ‘Aligning Language Models to Follow Instructions’ <<https://openai.com/research/instruction-following>> accessed 8 December 2023.

internet text, books, and other sources. In particular, concerns were raised that these models were infringing copyright by using copyrighted works in their training material without the owner's authorisation. These questions were initially the subject of policy, industry and academic debate but eventually found their way into the courts.

One of these cases is *Authors Guild v OpenAI*,⁸ filed at the end of September 2023 before the court for the Southern District Court of New York. The plaintiffs are 17 fiction writers together with the Authors Guild — a professional organization for published writers with over 14,000 members. At the heart of the case are claims of copyright infringement, with OpenAI being accused of using copyrighted works without permission for the training of its GPT models.

To understand the potential legal challenges presented by AI training data in Europe, this paper will analyse the *Authors Guild v. OpenAI* dispute through the lens of EU law. Specifically, the analysis will focus on potential copyright infringement at the input level. Following this introduction, Chapter 2 delves into a factual analysis of the *Authors Guild v OpenAI* case, specifically the claims made by the plaintiffs in the US. Chapter 3 presents a short analysis of US law, providing a brief examination of the fair use doctrine. This doctrine is relevant since it permits use of copyrighted material without requiring permission from the rights owners if specific circumstances are met. The analysis of the US law serves to understand how US courts handle copyright questions regarding AI, and offers a parallel for the subsequent analysis of EU law. To address the lack of clear legal precedent in the EU regarding text and data mining and copyright, Chapter 4 explores relevant EU legislation. Chapter 4 will focus on legislation pertaining to the right of reproduction, moral rights, and how the text and data mining exceptions in these legal instruments might apply in the context of AI training data. These exceptions allow for the legitimate use of copyrighted material in support of technological progress, research, and the digital economy. Chapter 5 provides some conclusions.

⁸ *Authors Guild v OpenAI Inc* SDNY 1:23-cv-08292.

2. Authors Guild V OpenAI

The OpenAI AI tool ChatGPT has sparked one of the most heated debates in human history, ranging from philosophical to technical to legal aspects.⁹ The end of August 2023 saw a significant increase in the number of articles¹⁰ discussing the tension between writers and tech companies. Particular attention was paid to OpenAI, which was accused of using the dataset Books3 to train their LLMs. The dataset can be freely downloaded and allegedly contains over 183,000 fiction and non-fiction books covered by copyright without the corresponding licenses.¹¹

2.1. Introduction

As discussed, *Authors Guild v OpenAI*¹² is just one case before courts around the world concerning copyright and AI. Filed Under 17 U.S. Code (U.S.C.) section 501,¹³ the Authors Guild¹⁴ and 17 authors claimed that OpenAI's ChatGPT was trained on unauthorised

⁹ Some view AI as “an intellectual revolution”, while others are more sceptical and perceive it with a critical eye. Alain Strowel, ‘ChatGPT and Generative AI Tools: Theft of Intellectual Labor?’ (2023) 54 IIC - International Review of Intellectual Property and Competition Law 491.

¹⁰ Alex Reisner, ‘Revealed: The Authors Whose Pirated Books Are Powering Generative AI’ (*The Atlantic*, 19 August 2023) <<https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/>> accessed 31 October 2023. See also Alex Reisner, ‘These 183,000 Books Are Fueling the Biggest Fight in Publishing and Tech - The Atlantic’ (25 September 2023) <<https://www.theatlantic.com/technology/archive/2023/09/books3-database-generative-ai-training-copyright-infringement/675363/>> accessed 31 October 2023; Alex Reisner, ‘What I Found in a Database Meta Uses to Train Generative AI’ (*The Atlantic*, 25 September 2023) <<https://www.theatlantic.com/technology/archive/2023/09/books3-ai-training-meta-copyright-infringement-lawsuit/675411/>> accessed 31 October 2023.

¹¹ Alex Reisner, ‘These 183,000 Books Are Fueling the Biggest Fight in Publishing and Tech - The Atlantic’ (25 September 2023) <<https://www.theatlantic.com/technology/archive/2023/09/books3-database-generative-ai-training-copyright-infringement/675363/>> accessed 31 October 2023. ; Richard Flanagan, the 2014 Booker Prize winner, considers the inclusion of such a vast number of copyrighted works to be “the biggest act of copyright theft in history.” Kelly Burke, “‘Biggest Act of Copyright Theft in History’: Thousands of Australian Books Allegedly Used to Train AI Model’ *The Guardian* (28 September 2023) <<https://www.theguardian.com/australia-news/2023/sep/28/australian-books-training-ai-books3-stolen-pirated>> accessed 8 December 2023.

¹² *Authors Guild v OpenAI Inc* SDNY (n 11).

¹³ Title 17 U.S.C outlines U.S. copyright laws. According to section 501 “(a) Anyone who violates any of the exclusive rights of the copyright owner as provided by sections 106 through 122 or of the author as provided in section 106A(a) (...) is an infringer of the copyright or right of the author, as the case may be.” Copyright Act 1976 (17 USC). <<https://uscode.house.gov/browse/prelim@title17&edition=prelim>>

¹⁴ ‘About the Guild’ (*The Authors Guild*) <<https://authorsguild.org/about/>> accessed 16 January 2024.

copyright-protected works.¹⁵ They also submitted that the LLM poses a threat to their careers and livelihoods.¹⁶ The reasons cited were that ChatGPT can quickly and efficiently generate content at little to no cost, something for which authors would ordinarily be paid.¹⁷ Furthermore, they argued that ChatGPT can produce works that might have copied and/or imitated their original works.¹⁸ In addition, the plaintiffs referred to statements made to the US Congress by OpenAI's CEO Sam Altman on 22 June 2023.¹⁹ On that occasion, Mr Altman expressed that OpenAI prioritises the growth of the creator economy.²⁰ According to plaintiffs, OpenAI also acted in contradiction to its own CEO's statements, who had (i) made assurances that OpenAI's aim was not to replace creators, rather to support and empower their creativity;²¹ and (ii) advocated in favour of creators having control over the use of their works, benefiting from AI technology. OpenAI, the plaintiffs argued, did not act in accordance with these assurances, and instead chose to infringe the applicants' copyright.²²

In December 2023 Microsoft Corporation (Microsoft) was added to the list of defendants²³ because of its alleged strong financial ties with OpenAI. First, it was argued, the LLM developed by OpenAI is deeply integrated into Microsoft services and products.²⁴ Secondly, the training of the model of ChatGPT could not have occurred without the tech giant's²⁵

¹⁵ 'Authors Guild v. OpenAI Inc. (Complaint Filed on 19 September 2023) No. 1:23-Cv-08292' (CourtListener), para. 4<<https://www.courtlistener.com/docket/67810584/40/authors-guild-v-openai-inc/>> accessed 8 December 2023.

¹⁶ *ibid*, para 2.

¹⁷ *ibid*, para 2.

¹⁸ *ibid*, para 2.

¹⁹ *ibid*, para 5.

²⁰ Creator economy refers to individuals who leverage the Internet to produce content, grow an audience and monetize these efforts by generating revenues. See Renana Peres and others, 'The Creator Economy: An Introduction and a Call for Scholarly Research' (1 June 2023), p.2 <<https://papers.ssrn.com/abstract=4663506>> accessed 16 January 2024.

²¹ *Ibid*.

²² 'Authors Guild v. OpenAI Inc. (Complaint Filed on 19 September 2023) No. 1:23-Cv-08292' (n 19), para 5.

²³ 'Authors Guild v. OpenAI Inc., 1:23-Cv-08292, (S.D.N.Y. Dec 05, 2023) ECF No. 40' <<https://www.courtlistener.com/docket/67810584/40/authors-guild-v-openai-inc/>>.

²⁴ 'The Inside Story of Microsoft's Partnership with OpenAI | The New Yorker' <https://www.newyorker.com/magazine/2023/12/11/the-inside-story-of-microsofts-partnership-with-openai> accessed 8 December 2023.

²⁵ 'Microsoft Added As Defendant In Writers' OpenAI IP Suit - Law360' <<https://www.law360.com/articles/1773329/microsoft-added-as-defendant-in-writers-openai-ip-suit>> accessed 9 December 2023.

“financial and technical support.”²⁶ Furthermore, since November 2023, Microsoft holds a ‘non-voting’ observer seat on the board of OpenAI, granting it increased insight and potential influence over the company’s internal decisions.²⁷

The plaintiffs²⁸ contended that OpenAI’s LLM must be fed with large amounts of data to be trained. To support this claim, they cited a 2020 report published by the U.S. Patent and Trademark Office (“USPTO”). The report observed that, at the input level, the training of an LLM entails the partial or entire reproduction of works.²⁹ Moreover, the plaintiffs maintained that books are known to be a rich source for the development of generative AI, necessary for creating high-quality LLMs.³⁰

Evidence may indicate that OpenAI used copyrighted data without authorisation, in ChatGPT-3 (free) and ChatGPT-4 (paid) versions.³¹ This can be seen in e.g. OpenAI’s response to the USPTO’s Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation.³² In its comments, OpenAI confirmed that its LLMs are being trained on ample datasets accessible to the public that incorporate copyrighted works.³³ Thus, the company declared that the use of large amounts of copyrighted data is necessary for the LLM to provide qualitative outputs. Moreover, they admitted that copies must be made to

²⁶ OpenAI was established as a non-profit research entity in 2015 and added a profit division in 2019, which enabled partners to hold equity in the startup and recover their investments.; Microsoft has invested \$13 billion in OpenAI and currently holds a 49% share in OpenAI’s “for-profit arm.” See Alex Heath, ‘Microsoft Joins OpenAI’s Board with Sam Altman Officially Back as CEO’ (*The Verge*, 30 November 2023) <<https://www.theverge.com/2023/11/29/23981848/sam-altman-back-open-ai-ceo-microsoft-board>> accessed 16 January 2024.

²⁷ Alex Heath, ‘Microsoft Joins OpenAI’s Board with Sam Altman Officially Back as CEO’ (*The Verge*, 30 November 2023) <<https://www.theverge.com/2023/11/29/23981848/sam-altman-back-open-ai-ceo-microsoft-board>> accessed 16 January 2024.

²⁸ ‘Authors Guild to Add Microsoft in Its Lawsuit against OpenAI’ <<https://www.courthousenews.com/authors-guild-to-add-microsoft-in-its-lawsuit-against-openai/>> accessed 8 December 2023.

²⁹ USPTO, ‘Public Views on Artificial Intelligence and Intellectual Property Policy’ (2020), p. 23-24 <<https://www.uspto.gov/about-us/news-updates/uspto-releases-report-artificial-intelligence-and-intellectual-property>> accessed 9 December 2023.

³⁰ ‘Authors Guild v. OpenAI Inc. (Complaint Filed on 19 September 2023) No. 1:23-Cv-08292’ (n 19), para 67-68.

³¹ In March 2023, OpenAI released ChatGPT-4 the underlying program that fuels ChatGPT Plus, a subscription-based chatbot with advanced capabilities, *Ibid*, §80-81.

³² ‘Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation’ (*Federal Register*, 30 October 2019) <<https://www.federalregister.gov/documents/2019/10/30/2019-23638/request-for-comments-on-intellectual-property-protection-for-artificial-intelligence-innovation>> accessed 22 March 2024.

³³ OpenAI, ‘Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation’ (US Patent and Trademark Office Dkt 2019) No. PTO-C-2019-0038 <https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf>.

analyse this data.³⁴ Still, uncertainty remains regarding the data that has been used in training ChatGPT-4 since OpenAI has refused to share or publish the source of their datasets.³⁵

OpenAI claims in their ChatGPT-3 paper³⁶ that their LLMs were trained with “Common Crawl,”³⁷ “Books1” and “Books2.” Common Crawl ‘crawls’/searches the Internet and collects large amounts of web data. Books1 and Books2, on the other hand, are datasets developed by OpenAI for training their LLMs. Books1 is known to include books that were part of Project Gutenberg, which is an online publisher that provides access to books which have expired copyrights, or which have a license for non-commercial distribution.³⁸ If true, at first glance this would benefit OpenAI’s defence since there seems to be no copyright protection over the content in the Books1 dataset. However, while the content of Books2 has not been disclosed, some believe it includes pirated copyrighted materials from Library Genesis, Z-Library and Bibliotik.³⁹ According to the plaintiffs, there is a strong suggestion that the similarities between Books2 and Books3 are connected to OpenAI’s LLMs qualitative development. Additionally, they claim that the limited availability of copyrighted sources is directly connected to OpenAI’s LLM’s qualitative development from GPT-3 to GPT-4.⁴⁰

2.2. Claims of copyright infringement against OpenAI

Considering what has been presented above, the plaintiffs brought three claims of copyright infringement. The first two concerned OpenAI exclusively, while the third one included Microsoft in the list of defendants.

³⁴ ‘Authors Guild v. OpenAI Inc. (Complaint Filed on 19 September 2023) No. 1:23-Cv-08292’ (n 19).

³⁵ ‘OpenAI Co-Founder on Company’s Past Approach to Openly Sharing Research: “We Were Wrong” - The Verge’ <<https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview>> accessed 16 January 2024.

³⁶ Tom B Brown and others, ‘Language Models Are Few-Shot Learners’ (arXiv, 22 July 2020) <<http://arxiv.org/abs/2005.14165>> accessed 10 December 2023.

³⁷ ‘Common Crawl - Mission’ <<https://commoncrawl.org/mission>> accessed 10 December 2023.

³⁸ Alex Reisner, ‘Revealed: The Authors Whose Pirated Books Are Powering Generative AI’ (*The Atlantic*, 19 August 2023) <<https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/>> accessed 31 October 2023.

³⁹ *ibid.* See also ‘Amended Complaint – #40 in Authors Guild v. OpenAI Inc. (S.D.N.Y., 1:23-Cv-08292) – CourtListener.Com’ (n 8), § 105.

⁴⁰ ‘Authors Guild v. OpenAI Inc. (Complaint Filed on 19 September 2023) No. 1:23-Cv-08292’ (n 19).

The initial claim pertained to a *direct copyright infringement* under title 17 U.S.C. section 501.⁴¹ The literary works’ owners submitted that OpenAI OpCo LLC had wilfully infringed their exclusive rights.⁴² The plaintiffs asserted that this infringement occurred “by reproducing their copyrighted works in copies for “training” OpenAI OpCo LLMs and ChatGPT.”⁴³ Furthermore, it was argued, that the defendant was aware that the datasets used for the training consisted of copyright-protected works.⁴⁴

The remaining two claims are of a secondary liability nature: vicarious and contributory liability. The secondary liability provisions of copyright law are solely developed by judges, lacking an open-ended statutory basis similar to fair use jurisprudence under section 107 of the Copyright Act.⁴⁵ The second claim is one for *vicarious copyright infringement*. Vicarious copyright liability applies to parties which indirectly benefit from the direct copyright infringement of other entities in their control.⁴⁶ In order to be vicariously liable, the party must have an economic interest and choose not to stop or restrict the infringement.⁴⁷ Accordingly, even if OpenAI Inc. and OpenAI GP LLC were not directly infringing, they would still be held liable. This however would only be the case if it can be shown that they provided the means for such infringement and directly derived financial benefit from it.⁴⁸

Lastly, a claim of *contributory copyright infringement* was asserted, contending that the “other OpenAI defendants”⁴⁹ and Microsoft had played a contributing role in the direct infringement. According to the plaintiffs, the defendants contributed to and assisted in the direct infringement by providing funding, capital, technology, personnel, and other resources. They also controlled

⁴¹ *ibid*, para 334-341.

⁴² Copyright Act 1976 (17 USC), §106(1).

⁴³ ‘Authors Guild v. OpenAI Inc. (Complaint Filed on 19 September 2023) No. 1:23-Cv-08292’ (n 27), §340.

⁴⁴ *Ibid*, §334-341.

⁴⁵ James Boyle and Jennifer Jenkins, *Intellectual Property: Law & Information Society*, p.515

<<https://web.law.duke.edu/cspd/papers/epubs/IPCasebook2014-Ch14.pdf>>.

⁴⁶ Robert M Hirning, ‘Contributory and Vicarious Copyright Infringement in Computer Software’ [2006] Chicago-Kent Journal of Intellectual Property, p. 11-12.

⁴⁷ *ibid*.

⁴⁸ *Ibid*, §342-345.

⁴⁹ Other OpenAI Defendants” refers to OpenAI LLC, OpenAI Global LLC, OAI Corporation LLC, OpenAI Holdings LLC, OpenAI Startup Fund I LP, OpenAI Startup Fund GP I LLC, and OpenAI Startup Fund Management LLC. See ‘Authors Guild v. OpenAI Inc. (Complaint Filed on 19 September 2023) No. 1:23-Cv-08292’ (n 19).

or managed the assets used for the infringement and offered business, legal, strategic, or operational guidance to achieve it.⁵⁰

Considering the above, it is evident that the plaintiffs saw copyright infringement at the input level in all three claims. The ruling of this case will clarify copyright boundaries in the AI era. Furthermore, the court's decision has the potential to challenge the established precedents that concern the intersection between copyright law and AI.

3. Copyright Infringement at the Input Level – US law

This chapter explores the fair use doctrine, which the courts must analyse to determine its applicability as a defence against copyright infringement. To do so, the purpose of copyright protection and the right of reproduction will be briefly scrutinized. The chapter will conclude with an analysis of the fair use factors and their applicability to the Authors Guild v Open AI and Microsoft. The analysis of the fair use doctrine is necessary to address whether there is a fair use of copyrighted works, as it clarifies whether such usage falls within the boundaries established by copyright law.

3.1 Purpose of Copyright Protection

According to U.S. Court of Appeals for the Second Circuit, the ultimate scope of copyright law is to enrich the knowledge and understanding of the public by enabling access to copyrighted works.⁵¹ While authors play and enjoy a significant role in the copyright realm, the main beneficiary of copyright law ought to be the public.⁵² Therefore, the purpose of copyright is to provide incentives for authors to progress the public's access to knowledge.⁵³

⁵⁰ Ibid, §346-349.

⁵¹ *Authors Guild v Google Inc.*, No. 13-4829 (2d Cir 2015), p. 13.

⁵²“The ultimate goal of copyright is to expand public knowledge and understanding, which copyright seeks to achieve by giving potential creators exclusive control over copying of their works, thus giving them a financial incentive to create informative, intellectually enriching works for public consumption. This objective is clearly reflected in the Constitution's empowerment of Congress “*To promote the Progress of Science ... by securing for limited Times to Authors ... the exclusive Right to their respective Writings.*” U.S. Const., Art. I, § 8, cl. 8 (emphasis added). Thus, while authors are undoubtedly important intended beneficiaries of copyright, the ultimate, primary intended beneficiary is the public, whose access to knowledge copyright seeks to advance by providing rewards for authorship.” *Authors Guild v Google, Inc* No. 13-4829 (2d Cir 2015), p. 13.

⁵³ *ibid*.

3.1. The Right of Reproduction

17 U.S.C. section 106 lays down the five exclusive rights for copyrighted works.⁵⁴ One of the most significant rights granted under this section is the right of reproduction.⁵⁵ Accordingly, a copyright owner can exclusively deal with and authorise the reproduction of copies of its work. According to the applicants, this right was breached by OpenAI when reproducing copyrighted works for the training of LLMs without the proper licenses.⁵⁶

3.2. The Fair Use Doctrine

The exclusive rights established under 17 U.S.C. section 106 are subject to some limitations constituting the fair use doctrine, under section 107. Section 107 provides four factors, which serve as guidelines for courts to determine whether a particular use of copyrighted material is considered fair. If so, an exemption from the infringement would apply.

The fair use doctrine factors are interconnected and collectively evaluated. The importance of each of them can vary depending on the context of the use. The factors are the following:⁵⁷

- (I) *The purpose and character of the use.* Under this factor, the courts must assess whether (i) the use of the new work advances socially beneficial activities and (ii) the use is transformative.⁵⁸ This means that the new work must exhibit a new character or add a new element or a new scope to the original work. For example, a movie (new work) made based on a book (original copyrighted work) would likely not be considered transformative.

⁵⁴ The 5 rights under are 1) the right to reproduce the copyrighted work; 2) the right to prepare derivative works based upon the work; 3) the right to distribute copies of the work to the public; 4) the right to perform the copyrighted work publicly and 5) the right to display the copyrighted work publicly. See Copyright Act.

⁵⁵ 'Rights Granted Under Copyright Law (BitLaw)' <<https://www.bitlaw.com/copyright/scope.html>> accessed 27 February 2024.

⁵⁶ 'Authors Guild v. OpenAI Inc. (Complaint Filed on 19 September 2023) No. 1:23-Cv-08292' (n 19), para 63.

⁵⁷ 'The Meaning of the Four Fair Use Factors' <<https://copyright.byu.edu/the-meaning-of-the-four-fair-use-factors>> accessed 27 February 2024.

⁵⁸ *Campbell v Acuff-Rose Music Inc* (510 US).

The concept of transformative use was applied in cases such as *Authors Guild, Inc. v. HathiTrust*⁵⁹ and *Authors Guild, Inc. v. Google*.⁶⁰ In both cases, digitisation for creating search indexes was deemed transformative since it served a different purpose from the original works, and it enabled new types of research.⁶¹

In *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*,⁶² the Supreme Court stated that the new message, meaning or expression of a new work does not suffice on its own to fulfil the first criterion for fair use.⁶³ The level of transformation needed to fulfil the first criterion must exceed what is considered necessary to constitute a derivative work.⁶⁴ The Court pointed out, that the derivative work right⁶⁵ includes the concept of transformation, which encompasses “any other form in which a work may be recast, transformed or adapted.”⁶⁶ Therefore, the Supreme Court stated that transformative use hinders the copyright owner’s exclusive right to create derivative works if the inclusion of a new expression is decisive regarding the first factor.

Additionally, the Court clarified that although the new expression in a secondary work is relevant in assessing its transformative nature, it is not the sole determining factor. In the judgement, the commercial nature of a use was considered an extra aspect of the first criterion. The court noted the relevance of the works commercial nature while also clarifying that it is not the decisive factor. Allowing a similar use could potentially permit a variety of commercial works to be used for similar purposes as the original.

⁵⁹ *Authors Guild v HathiTrust*, 755 F3d 87 (2d Cir 2014).

⁶⁰ *Authors Guild v. Google, Inc.* (n 55).

⁶¹ Van Lindberg, ‘Building and Using Generative Models Under US Copyright Law’ (30 May 2023) <<https://papers.ssrn.com/abstract=4464001>> accessed 23 March 2024.

⁶² *Andy Warhol Found for the Visual Arts, Inc v Goldsmith* (143 S Ct 1258).

⁶³ *ibid.* p. 12.

⁶⁴ *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith* (n 65), p.3.

⁶⁵ Section 106(2), Copyright Act.

⁶⁶ “Such transformations may be substantial, like the adaptation of a book into a movie.” *Authors Guild v Google Inc.*, No. 13-4829 (2d Cir 2015), p.16.

In the present case, it can be argued that the use of copyright-protect books to train LLMs aids developers in creating qualitative language models.⁶⁷ The original books serve a creative and aesthetic purpose vastly distinct from the data-focused objectives the books serve in training the AI model. Hence, this fair use factor would be fulfilled. Similar to the search indexes in previously discussed Authors Guild cases, LLMs also appear to be transformative in nature. This transformative nature would stem from producing LLMs with distinct purposes and capabilities compared to their training inputs.⁶⁸

- (II) *The nature of the copyrighted work.* This factor focuses on the creativity of the copyrighted work.⁶⁹ The higher the creativity of the protected work the lower the chance that fair use is accepted by courts. Some examples of highly creative works are novels and poems, which might have been used in the training data set. In the same vein, the use of factual works (e.g., scientific papers and historical documents) would more likely qualify for a fair use defence.⁷⁰

In the current case, the applicants are fiction writers which implies a higher level of creativity in their works. Therefore, applying the above analysis, the use of their creative works in the LLMs' training, without prior authorisation, might not be justified. Thus, no fair use exception would apply based on this factor.

- (III) *The amount and the substantiality of the portion used in relation to the copyrighted work as a whole.* In this context, U.S. jurisprudence has clarified that copying works in their entirety can fall under the fair use exemption.⁷¹ In *Authors Guild v. Google, Inc* the U.S. Court of Appeals for the Second Circuit found Google's copy of copyright-protected texts a fair use exemption, as it enabled 'full-text search functions.' In other words, Google needed to copy and retain the works for the proper functioning of the database.⁷² Beneficial to Google's defence was the fact that it did not display digital books in their entirety. It rather showed fragments of

⁶⁷ Guadamuz (n 4).

⁶⁸ Lindberg (n 65), p 38.

⁶⁹ 'Copyright and Fair Use' (n 54).

⁷⁰ Harvard Law Review, 'Google LLC v. Oracle America, Inc.' (*Harvard Law Review*, 10 November 2021) <<https://harvardlawreview.org/print/vol-135/google-llc-v-oracle-america-inc/>> accessed 23 January 2024.

⁷¹ *Authors Guild v. Google, Inc.* (n 55).

⁷² *Ibid.*

text, based on the term searched by the users.⁷³ The Court found that Google's goal was limited to displaying short segments of text in order to aid the user in deciding whether the book was of interest to their needs.⁷⁴

Thus, in the case of *Authors Guild v OpenAI*, this factor may be interpreted in favour of Open AI. Just as Google needed material for its 'full-text search functions,' Open AI LLMs require a vast number of works in their training process for the proper functioning of the models. Furthermore, Open AI did not display the training data sets at the input level. This is comparable to Google only showing small excerpts. Therefore, a substantial amount of copying for training purposes may be allowed in the present case, as it aligns with the justifications presented in the Google case.

- (IV) *The effect of the use upon the potential market for or value of the copyrighted work.* This last factor is frequently regarded as being key in assessing fair use.⁷⁵ The Supreme Court ruled in *Campbell v Acuff Rose* that the more transformative the purpose, the less likely it is to be a market substitute for the original.⁷⁶ Furthermore, in *Authors Guild v Google*, the Court evaluated the use of copyrighted works in the context of text and data mining (TDM).⁷⁷ In that case, the Second Circuit Court of Appeals concluded that the market and value impact of TDM did not violate copyright law. This finding was based on the fact that the flow of information made available by a protected work is not protected by copyright. Consequently, in the *Authors Guild v OpenAI* case, to fulfil this fair use exception factor, the LLMs at the input level (when the AI training process takes place) should not negatively impact the market or the value of the copyrighted work at its initial

⁷³ U.S. Copyright Office Fair Use Index, 'Authors Guild, Inc. v. Google Inc., No. 13-4829-Cv (2d Cir. Oct. 16, 2015)' <<https://www.copyright.gov/fair-use/summaries/authorsguild-google-2dcir2015.pdf>>.

⁷⁴ *Authors Guild v. Google, Inc.* (n 56).

⁷⁵ 'The Meaning of the Four Fair Use Factors' (n 61).

⁷⁶ *Campbell v Acuff-Rose Music Inc* (n 54), para 571.

⁷⁷ TDM is a process of quantitatively analysing various types of data, through automated techniques. 'Fair Use Week 2023: Looking Back at Google Books Eight Years Later | Authors Alliance' <<https://www.authorsalliance.org/2023/02/24/fair-use-week-2023-looking-back-at-google-books-eight-years-later/>> accessed 12 March 2024.

input stage.⁷⁸ In other words, the use of the books in the data training should not damage the market of the original works.⁷⁹

4. Copyright infringement at the input level – EU law

The landscape in Europe regarding AI case law is quite opaque because to date, there is only one ongoing lawsuit on the subject, *Robert Kneschke v Laion e.V.* This case will be heard in Germany by the Hamburg Regional Court at the end of April 2024.⁸⁰ In the absence of EU case law, one can only reflect on how the *Authors Guild v OpenAI* case would be determined by EU courts. The following analysis explores the legal implications for intellectual property rights and the legal responsibilities of companies developing and/or marketing generative AI models. Consequently, it will draw parallels from already existing EU legislation and case law that could fill the existing gap.

While there is no case law specifically addressing AI, the following analysis explores existing EU legislation and case law that might provide parallels for addressing intellectual property rights and the legal responsibilities of companies developing and marketing generative AI models. The analysis will focus on key copyright concepts—such as the right of reproduction, the TDM exceptions, and moral rights—to examine their potential application in the context of generative AI and their implications for protecting authors' creative works.

4.1. Right of Reproduction

According to article 2(a) of the Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society (InfoSoc Directive), authors have the exclusive right to authorise or prohibit reproductions of their works.⁸¹ TDM processes typically

⁷⁸ This principle arises from the understanding that the copies are strictly used as data for training the generative AI models.

⁷⁹ Stephen Wolfson, 'Fair Use: Training Generative AI' (*Creative Commons*, 17 February 2023) <<https://creativecommons.org/2023/02/17/fair-use-training-generative-ai/>> accessed 23 March 2024.

⁸⁰ The case concerns the LAION 5b dataset, which contains links to images under the photographer's Robert Kneschke copyright. Some of the images found were uploaded on Shutterstock and had watermarks. 'Gerichtstermin ist anberaumt - Klage gegen LAION e.V. vor dem Landgericht Hamburg' (*SLD IP RA GmbH*) <<https://www.sld-ip.com/blog/klage-gegen-laion-e-v-vor-dem-landgericht-hamburg/>> accessed 5 February 2024.

⁸¹ According to the reproduction right, "Member States shall provide for the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in

transform the copies of copyrighted material into a format that allows its analysis. Thus, TDM processes involve acts that require the creation of transient copies, which are temporary digital copies of a work created as part of a technical process. According to the Court of Justice of the EU in the *Infopaq* case,⁸² if a work meets the originality criteria, both the entire work and parts of it enjoy copyright protection. Hence, reproducing them would lead to the infringement of article 2(a) InfoSoc.⁸³ Nevertheless, article 5(1) InfoSoc provides an exception for transient copies that facilitates such technological processes.⁸⁴ To benefit from such exception, five cumulative criteria must be met:⁸⁵

- I. *The act is temporary.* Training of OpenAI models is conducted by feeding the model vast quantities of information. According to OpenAI, such information is not memorised but analysed in a search for patterns and transforming those into mathematical representations.⁸⁶ Therefore, the OpenAI tool would not access the training data when required to generate an output. Instead, “it is working from a leaner and much more concise set of abstract concepts.”⁸⁷ Therefore, this criterion would be fulfilled since datasets would no longer be needed after the training of the AI model has concluded.⁸⁸
- II. *The act is transient or incidental.* Transient copies refer to copies considered necessary for completing the technological process, with the condition that they are automatically deleted upon completing their task.⁸⁹ Incidental copies are those which neither exist nor have a purpose beyond the

part”; See Article 2, Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

⁸² Case C-5/08, *Infopaq International A/S v Danske Dagblades Forening*.

⁸³ Thomas Margoni and Martin Kretschmer, ‘A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology’ (2022) 71 GRUR International 685.

⁸⁴ Sean Flynn and others, ‘Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action’ [2020] SSRN Electronic Journal <<https://www.ssrn.com/abstract=3578819>> accessed 12 February 2024.

⁸⁵ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* (n 84), para 54; Kalpana Tyagi, ‘The Copyright, Text & Data Mining and the Innovation Dimension of Generative AI’ [2023] SSRN Electronic Journal <<https://www.ssrn.com/abstract=4615929>> accessed 12 February 2024.

⁸⁶ Isaac Sandiumenge, ‘Copyright Implications of the Use of Generative AI’ [2023] SSRN Electronic Journal <<https://www.ssrn.com/abstract=4531912>> accessed 12 March 2024.

⁸⁷ Guadamuz (n 4).

⁸⁸ *ibid.*

⁸⁹ Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* (n 87).

technological process in which they are involved.⁹⁰ EU courts might find this criterion to be fulfilled, in the context of *AuthorsGuild v Open AI*. Training AI requires that copies of the information in the data set be made. This is necessary since the data needs to be accessed, read, stored, and analysed to process and extract information and generate outputs. Furthermore, while the generative AI models of OpenAI do not need to retain copies, there is a possibility that transient copies might be generated.⁹¹ In that case, this criterion would be satisfied.

- III. *The act is integral or essential part of the technological process.* Any copying or reproduction of content must be conducted as part of the technological process. These actions ought to be essential for the proper and efficient functioning of the procedure.⁹² Hence, the third criterion may be fulfilled in the case of OpenAI since the use of the datasets is an integral and essential part of AI training.⁹³
- IV. *The sole purpose is to enable transmission in a network between third parties by an intermediary, or a lawful use of a work or protected subject-matter.* The use is considered lawful if the rightsholder has authorised it or the use is not prohibited by law.⁹⁴ In the case of OpenAI, the use may be considered lawful since it is an integral part of the training and functioning of AI models. However, the use may also be found unlawful in circumstances where the rightsholder restricts the process contractually, or through domestic law.⁹⁵
- V. *No independent economic significance.”* According to the Court of Justice of the European Union,⁹⁶ any temporary copies should not possess

⁹⁰ *Case C-360/13 Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd and Others (Meltwater)*.

⁹¹ Guadamuz (n 4).

⁹² *Case C-360/13 Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd and Others (Meltwater)* (n 95).

⁹³ Guadamuz (n 4).

⁹⁴ Recital 33, Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (Infosoc Directive).

⁹⁵ Thomas Margoni, ‘Artificial Intelligence, Machine Learning and EU Copyright Law: Who Owns AI?’ [2018] SSRN Electronic Journal <<https://www.ssrn.com/abstract=3299523>> accessed 12 March 2024.

⁹⁶ *Football Association Premier League and Others*, Joined Cases C-403/08 and C-429/08, 4 October 2011, ECLI:EU:C:2011:631 (FAPL), *Case C-302/10 Infopaq International A/S v Danske Dagblades Forening*.

independent economic value. In the case of generative AI, the resulting model has an economic significance, whereas the temporary reproductions do not.⁹⁷ At the same time, the value of the generative AI model stems from the large number of copies used in training. Thus, it may be assumed that there is economic significance in the overall model rather than in specific individual copies used for training. Hence, this criterion would not be fulfilled.

To conclude, OpenAI would likely infringe the right of reproduction. If so, EU courts would need to undertake a technical examination the training model to determine whether these copies fall under the transient copy exception.⁹⁸ The criteria under the exception of article 5 InfoSoc have a cumulative character, and it is strictly interpreted. Consequently, the exception laid down under article 5 InfoSoc may not apply to TDM activities.⁹⁹ Considering this, OpenAI may need to rely on another defence. One possibility would be to advocate in favour of TDM exceptions found in the Digital Single Market (DSM) Directive,¹⁰⁰ which will be presented below.

4.2. TDM Exceptions

a) Current legal framework

Under EU law, TDM is defined as “any automated analytical technique aimed at analysing text and data in digital form to generate information which includes but is not limited to patterns, trends and correlations.”¹⁰¹ The value of TDM has exponentially increased due to the growing

⁹⁷ Guadamuz (n 4).

⁹⁸ Ibid.

⁹⁹ Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, ‘Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/Eu’ [2019] SSRN Electronic Journal <<https://www.ssrn.com/abstract=3470653>> . in Kalpana Tyagi, ‘The Copyright, Text & Data Mining and the Innovation Dimension of Generative AI’ [2023] SSRN Electronic Journal <<https://www.ssrn.com/abstract=4615929>> accessed 12 February 2024.

¹⁰⁰ Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC <<https://eur-lex.europa.eu/eli/dir/2019/790/oj>>

¹⁰¹ Article 2(2), ‘Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market’ <<https://eur-lex.europa.eu/eli/dir/2019/790/oj>> accessed 5 February 2024.

capabilities of computer processing and the availability of vast amounts of information on the Internet.¹⁰²

The DSM Directive includes two TDM exceptions in articles 3 and 4. The first exception¹⁰³ (article 3 DSM) applies to acts of reproduction and extraction by research organisations¹⁰⁴ and cultural heritage institutions.¹⁰⁵ This means that the beneficiaries of this exception are institutions dedicated to offering cultural or public services that do not operate for profit. Nevertheless, private parties may benefit from this exception when acting within the framework of a public-private partnership. However, this benefit is limited, depending on the duration and scope of the collaborative project with the public institution.¹⁰⁶ According to Recital 14 of the DSM Directive, TDM beneficiaries under article 3 need to have lawful access to the work they are accessing. This includes open access, contractual agreements, or freely available online content. This would likely not be the case for OpenAI.

The second exception is limited in its extent, although it exempts a larger range of users. Article 4 permits the reproduction and extraction of legally accessible works for TDM purposes.¹⁰⁷ The works can be used as long as necessary for TDM purposes.¹⁰⁸ Nevertheless, rightsholders are allowed to opt-out of such TDM activities. Meaning they can exclude their copyrighted work from being included in a lawful collection of data. According to article 4(3) DSM Directive, “the exception or limitation shall apply on the condition that the use of works [...] has not been expressly reserved by their rightsholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online”. Recital 18 of the DSM Directive recognises various methods for rightsholders to impede TDM when opting out, such as metadata and terms and conditions of a website or a service. Other viable options

¹⁰² P Bernt Hugenholtz, ‘The New Copyright Directive: Text and Data Mining (Articles 3 and 4)’ (*Kluwer Copyright Blog*, 24 July 2019) <<https://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/>> accessed 5 February 2024.

¹⁰³ Article 3, ‘Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market’ (n 53).

¹⁰⁴ Article 2(1), *ibid.*

¹⁰⁵ Article 2(3), *ibid.*

¹⁰⁶ Recital 11, *ibid.*

¹⁰⁷ Article 4(1), *ibid.*

¹⁰⁸ Article 4(2), *ibid.*

are contractual agreements or unilateral declarations when these options are more practical or effective.¹⁰⁹

In light of the above, it can be seen that OpenAI does not fall under article 3 DSM exception due to the company's commercial activity. Instead, OpenAI may want to invoke the TDM exception under article 4 DSM. However, in the case that OpenAI has used Books3 dataset (and other pirated works) in their training, this exception would not be applicable. This is because OpenAI has not reproduced and extracted "legally accessible works." Additionally, OpenAI will have to fulfil the obligation of respecting the rightsholders' opt-out.

b) Future legal framework

In a study published in February 2022,¹¹⁰ several rightsholders expressed concerns regarding these TDM exceptions. They complained about the difficulty in verifying if their opt-out requests have been honoured by AI companies. This difficulty arises since their material is not necessarily identifiable in the AI-generated output.

These concerns and doubts are discussed to some extent in the text of the EU AI Act¹¹¹ The novel legislation focuses on rules for generative-purpose AI (GPAI) systems and models. At the same time, it comprises two provisions which concern copyright law.

The first requirement, explicitly states that the TDM exception under article 4 DSM, applies to the use of copyrighted works for training AI models. Hence, using copyrighted works requires permission from their rightsholders. Additionally, providers of GPAI models must obtain authorisation from rightsholders for TDM activities if the latter have exercised their right to opt-out.¹¹²

The second requirement compels providers of GPAI models to make detailed summaries of the content used in generative AI training available. The summaries ought to

¹⁰⁹ Crowell&Moring and others, *Study on Copyright and New Technologies: Copyright Data Management and Artificial Intelligence* (Publications Office of the European Union 2022) <<https://data.europa.eu/doi/10.2759/570559>> accessed 7 February 2024.

¹¹⁰ Ibid.

¹¹¹ 'A First Look at the Copyright Relevant Parts in the Final AI Act Compromise - Kluwer Copyright Blog' <https://copyrightblog.kluweriplaw.com/2023/12/11/a-first-look-at-the-copyright-relevant-parts-in-the-final-ai-act-compromise/#_ftn1> accessed 12 March 2024.

¹¹² Ibid.

include the main data sources used, such as public and private databases. The aim is to aid third parties in assisting if the models were lawfully trained as required by article 4(1) DSM Directive.¹¹³

Consequently, OpenAI will be required under the EU AI Act to reveal the sources of its training data.

4.3. Moral Rights

This last section examines moral rights which, in comparison to economic rights, tend to receive minimum attention in legal analysis.¹¹⁴ Under EU copyright law, moral rights remain unharmonized.¹¹⁵ The moral rights recognised at the EU level are the rights of integrity and paternity.¹¹⁶ The EU sets a minimum standard for moral rights, allowing Member States to grant additional protection.¹¹⁷ The InfoSoc Directive excludes moral rights from its scope (Recital 19). However, they ought to be exercised in line with Member States' legislation, the Berne Convention,¹¹⁸ the WIPO Copyright Treaty¹¹⁹ and the Performances and Phonograms Treaty.¹²⁰ Following article 6bis of the Berne Convention, moral rights are upheld until the termination of economic rights (life of the author + 70 years). One important factor is that while economic rights may be transferred, moral rights remain with the author. This can create further legal issues when it comes to licensing authorisation of works for AI training. For example, the copyright owners may consent to the use of the work in the training data set based on the economic right of reproduction. Nevertheless, the AI developer is bound to get the

¹¹³ *ibid.*

¹¹⁴ Rita Matulionyte, 'The (Forgotten) Moral Rights in the Age of AI' (*Kluwer Copyright Blog*, 7 February 2022) <<https://copyrightblog.kluweriplaw.com/2022/02/07/the-forgotten-moral-rights-in-the-age-of-ai/>> accessed 13 February 2024.

¹¹⁵ Josef Drexl and others, 'Artificial Intelligence and Intellectual Property Law - Position Statement of the Max Planck Institute for Innovation and Competition of 9 April 2021 on the Current Debate' [2021] SSRN Electronic Journal <<https://www.ssrn.com/abstract=3822924>> accessed 13 February 2024.

¹¹⁶ 'EU Copyright - EUR-Lex' <<https://eur-lex.europa.eu/EN/legal-content/glossary/eu-copyright.html>> accessed 13 February 2024; See also Article 6bis of the Berne Convention: "Independently of the author's economic rights, and even after the transfer of the said rights, the author shall have the right to claim authorship of the work and to object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his honor or reputation."

¹¹⁷ Drexl and others (n 103).

¹¹⁸ *Ibid.*

¹¹⁹ WIPO Copyright Treaty, Dec. 20, 1996.

¹²⁰ WIPO Performances and Phonograms Treaty, Dec. 20, 1996.

authorisation of the authors who hold the moral right to process their protected works. This will add to the complexity of the pre-collection of data for the training of the AI.

Furthermore, the processing of works might lead to the infringement of the right of integrity. Moral rights owners may object to their work being used in training AI and perceive it as a degradation of their work.¹²¹ In the case of training generative AI, the Books3 dataset, for example, contained no explicit label, title, names or metadata of the authors of potential copyrighted works.¹²² The input data consisted only of the text necessary for training AI models. The removal of the author's information can, however, constitute an infringement of the paternity right since the attribution of the author is not preserved. On the other hand, the removal was conducted by the dataset's owner and not directly by OpenAI. Whether OpenAI could be held liable for using copies that have not maintained the authors' attribution, is a matter yet to be determined.

5. Conclusion

OpenAI has pioneered advancements in AI research and development, obtaining worldwide recognition with ChatGPT. The increased use of generative AI, however, has raised concerns regarding the data used for its training.

This has led a group of copyright owners to file lawsuits in US courts, alleging copyright infringement. One such ongoing case is *Authors Guild v OpenAI*. Whether US courts find OpenAI practices in the training of its LLM a fair use remains to be seen. While US courts are dealing with an ever-expanding number of lawsuits against AI companies, similar claims may arise before EU courts.

In Europe, courts may consider that OpenAI infringes the right of reproduction by using copies of books to train the ChatGPT models without the corresponding licences. The DSM Directive provides two TDM exemptions in their articles 3 and 4. The first 'research' exemption under article 3 DSM, is not applicable, since OpenAI has a commercial activity. However, according to the AI Act, the second TDM exception outlined in Article 4 DSM could apply if OpenAI obtains the training data from legally accessible works and respects writers' right to opt out. If OpenAI has utilized the Books3 dataset or any type of pirated works in their training, this exception would not apply. Furthermore, moral rights are an additional element

¹²¹ Ibid.

¹²² Alex Reisner, (n 14).

that the courts of the Member States ought to consider. Firstly, the companies training AI may be infringing on the right to integrity. This stems from the authors' perception that such processing might be considered a derogatory treatment of their work. Secondly, the paternity right is infringed by the alterations made to the copyrighted material, with the removal of watermarks and the author's information. While this infringement is conducted by the data set owner, the infringement might extend to and implicate companies utilising their datasets for AI training.