Training of Generative AI Systems and Copyright Law: A U.S. and European Perspective

By Flore Brunetti

Introduction

The surge of artificial intelligence (AI) in the past year has given rise to heated debates, fueled by the Writers Guild of America (WGA) and the Screen Actors Guild's (SAG-AF-TRA) strikes, as to the protection of artists in the age of generative AI. Defined as "an application of AI used to generate outputs in the form of expressive material such as text, images, audio, or video,"[1] generative AI relies on the collection by machine learning algorithms (e.g., large language models, or LLMs) of vast amounts of preexisting data, at least partly scraped from the Internet.[2] The generative AI system processes this data to "learn patterns inherent in human-generated data and then use those patterns to synthesize similar data."[3] Thus, to generate high-quality texts, images, and sounds, generative AI systems need to be trained on high-quality data,[4] which includes copyrighted works.[5]

With increasingly efficient generative AI systems becoming available on the market worldwide, the Authors Guild "call[ed] attention to the inherent injustice of building lucrative generative AI technologies using copyrighted works"[6] without any consent, credit, or compensation for the copyright owners. Similarly, artists guilds within the European Union are also demanding better protection of copyright owners and increased transparency regarding the works used to train AI.[7] Several AI market leaders are concurrently being sued in the U.S.,[8] in an attempt to curb their practice of "ingesting everything on the internet, with no concern for copyright."[9]

The coming courts' decisions and eventual new laws will be paramount in the regulation of generative AI training, at both the local and international levels. Given the pivotal role of the U.S. and the EU in AI regulation and the need for a "more unified international approach [. . .] for trustworthy AI development,"[10] a comparative analysis of the U.S. and the EU's takes on generative AI training is essential in identifying the challenges faced while awaiting further developments. This article focuses on assessing whether (1) the processing of copyrighted works for purposes of training a generative AI system falls within the scope of copyright law, and, if so, (2) whether AI developers benefit from exceptions to copyright owners' exclusive rights. The question as to if the AI-generated output is infringing is not covered by this analysis.

Applicability of Copyright Law to Generative AI Training

Under both U.S.[11] and EU[12] law, copyright owners are granted an exclusive reproduction right, which is the right most likely to be involved in the generative AI training process.[13] Determining whether the training of generative AI on copyrighted works implicates the reproduction right leads to the question as to whether the deconstruction of the work for training purposes qualifies as a copy and whether this copy is sufficiently permanent or stable.

Deconstruction of the Copyrighted Work and Notion of Copy

To implicate the reproduction right under U.S. law, there must be a copy "fixed by any method now known or later developed."[14] Similarly, EU law refers to "reproduction by any means and in any form, in whole or in part."[15] From a technical standpoint, the copyrighted works used for generative AI training are turned "into a machine-readable format compatible with the technology to be deployed for [text and data mining]"[16] prior to the data being extracted and recombined to identify informational value and patterns to generate new data.[17]

Despite the broad scope of the reproduction right, some commentators rely on the above-described deconstruction process to argue that generative AI training does not involve a reproduction for copyright law purposes, as the algorithms merely capture the informational value of the broken-down works, rather than their protected expression.[18] However, as highlighted by the U.S. Copyright Office (USCO) regarding random access memory (RAM) copying, "the reproduction right is not limited to copies of an entire work,"[19] and material broken-down into packets representing only portions of a work can constitute copies.[20] From an EU perspective, the Court of Justice of the European Union (CJEU) held, in a case involving the electronic processing of newspaper articles using an automated digital process, that nothing indicates that parts of a work "are to be treated any differently from the work as a whole,"[21] especially as "the cumulative effect of those extracts may lead to the reconstitution of lengthy fragments [. . .] reflect[ing] the originality of the work in question."[22] Furthermore, the existence of an exception to the reproduction right for purposes of text and data mining in the EU[23] supports the argument that this process involves a reproduction.

The fact that the training data is destined for a machine and is not necessarily communicated to the end user in the generated output does not change this analysis under U.S. law[24] as a copy "can be perceived [. . .] with the aid of a machine or device."[25] From an EU perspective, although the applicable directive does not expressly refer to machine use, its broad language likely encompasses copies destined for a machine.[26] However, this analysis must be nuanced in light of a CJEU decision holding that while a very short sound sample usually qualifies as a reproduction,[27] the reproduction right does not apply if the sample is used "in a modified form unrecognizable to the ear, in a new work."[28] If transposed to AI-generated works, this decision could limit the application of the reproduction right in the training process, "so as to permit some uses that do not lead to an output in which the protected elements contained in the input are visible or audible."[29] Nonetheless, the impact of this decision must be limited, as (i) it pertains to the phonogram producer's related rights rather than copyrights, and (ii) it merges the notions of reproduction and communication to the public, even though the reproduction right is a standalone right that can be infringed in and of itself. Thus, while worth considering, "sampling should be considered as a particular hypothesis, mixing the two acts of use and which as such differs in nature from data mining, which only concerns the act of reproduction."[30]

In light of the above, it is likely that at least some uses of copyrighted works in the generative AI training process implicate the reproduction right under both U.S. and EU law, provided the copies are sufficiently permanent or stable.

Necessity of a Sufficiently Permanent or Stable Copy

Under U.S. law, a copy must be "sufficiently permanent or stable to permit it to be perceived [. . .] for a period of more than transitory duration"[31] to implicate the reproduction right. In the EU, while the "sufficiently permanent or stable" requirement is not embedded in the notion of reproduction itself, an exception to the reproduction right allows some temporary acts of reproduction.[32] From a technical standpoint, the generative AI training process "necessarily involves first making copies of the data to be analyzed."[33] While these copies are not necessarily stored,[34] copyright owners argue that "if a model can accurately summarize a work, or [. . .] generate an image that appears to contain a mutilated watermark from plaintiff's work, then a copy exists somewhere in the back end of these generative AI models or, at a minimum, [. . .] infringement occurred during the training process."[35]

Whether said copies are sufficiently permanent or stable to implicate the reproduction right under U.S. law will be determined by the courts on a case-by-case basis in the absence of "a general rule defining how long a reproduction must endure to be "fixed."[36] However, the USCO highlighted that:

> The definition of 'fixed' does not require that a copy be permanent or that it last for any specified period of time. [. . .] These copies need exist only long enough to be [. . .] communicated (e.g., to a computer's processing unit) in order for their economic value to be realized. [. . .] Commercial exploitation in a network environment can be said to be based on selling a right to perceive temporary reproductions of works.[37]

Therefore, the generative AI training process likely involves sufficiently permanent or stable copies for the reproduction right to apply under U.S. law. Indeed, provided the copies exist long enough for the generative AI system to perceive their informational value, they would arguably have realized their economic value by contributing to improving the system.

The analysis differs under EU law, where:

> Temporary acts of reproduction [. . .] 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 [. . .], and which have no independent economic significance, shall be exempted from the reproduction right."[38]

As these conditions are cumulative and must be interpreted strictly,[39] it is unlikely that this exception would apply to the unauthorized use of copyrighted works for purposes of generative AI training. First, for the act to be transient, the technological process must automatically delete the copy upon completion of said process, "without human intervention."[40] Second, whether the use is lawful is the focal point of the debate, and the 'sole purpose' condition will likely not be fulfilled. Furthermore, the absence of independent economic significance requires that the economic advantage derived from the temporary act of reproduction is not "distinct or separable from the economic advantage derived from the lawful use of the work concerned."[41]

In light of the above, while arguably "there are extraction techniques that would reproduce parts of the work so minimal to fall below the threshold of copyright infringement,"[42] it is likely that at least some uses of copyrighted works for purposes of generative AI training involve the making of copies sufficiently permanent or stable to implicate the reproduction right under both U.S. and EU law.

Generative AI Training and Copyright Law Exceptions

Provided that the use of copyrighted works in the generative AI training process implicates the reproduction right, such use would constitute copyright infringement, absent an applicable exception. In the U.S., fair use is at the forefront of the conversation. In the EU, practitioners operate with a list of limited exceptions, in the absence of an equivalent to the fair use doctrine.

Generative AI and Fair Use under U.S. Law

In the U.S., whether the use of copyrighted works to train generative AI systems is considered to be fair will be determined by the courts considering (1) the purpose and character of the use, (2) the nature of the copyrighted work, (3) the amount and substantiality of the portion of the work used, and (4) the effect of the use upon the potential market for the copyrighted work.[43]

While all four factors must be considered, most AI developers rely heavily on the transformative nature of the use, a subpart of the purpose and character of the use, to argue in favor of a finding of fair use. As an illustration, OpenAI claims that:

> Training of AI systems is clearly highly transformative. Works in training corpora were meant primarily for human consumption for their standalone entertainment value. [. . .] Intermediate copying of works in training AI systems is, by contrast, 'non-expressive': the copying helps computer programs learn the patterns inherent in human-generated media [. . .] [to create] a useful generative AI system.[44]

However, it can be argued that copyrighted works are, in fact, used for their standalone entertainment value as "the goal of copying and showing a protected work to an AI system is for the analog of 'direct human consumption of the author's expression' in order for the system to learn about that expression."[45] Furthermore, as the Supreme Court stated in *Warhol Foundation v. Goldsmith*, the transformative nature of the use "must be weighed against other considerations, like commercialism,"[46] which "weighs against finding fair use."[47] However, the impact of *Warhol* must be nuanced, as a district court recently "decline[d] to overread one decision"[48] and "focus[ed] on transformativeness"[49] before concluding that whether the use of copyrighted works to train generative AI systems is transformative is "a material question of fact that the jury needs to decide."[50]

Another central factor in the finding of fair use is the effect of the use on the potential market for the copyrighted work. Copyright owners argue in this regard that the unauthorized use of copyrighted works for generative AI training "destroys copyright owners' licensing markets"[51] and "threatens to damage [artists'] profession by flooding the market with mediocre [. . .] [machine generated content] based on [artists'] work."[52] By contrast, Google argues that "the fact that a use 'enables [a user of an AI tool] to enter the market for works of the same type as the copied work' [. . .] is not the kind of market harm that is relevant to fair use."[53] OpenAI further claims that, since the copyrighted works used to train generative AI "are consumed by machines, not humans, the authors should lose no potential audience due to the use of their works."[54] However, as an exception to the reproduction right, fair use must not "unreasonably prejudice the legitimate interests of the author"[55] under the international Berne Convention. Given the high risk of loss of income and market dilution caused by generative AI, the application of fair use to the unauthorized training of for-profit generative AI systems on copyrighted works would arguably unreasonably prejudice the legitimate interests of artists worldwide.

The above factors, along with the nature and the portion of the copyrighted works used, will be considered by the courts on a case-by-case basis. Depending on the circumstances of each case, some generative AI training processes might be considered fair use, while others would be deemed infringing. In the context of this legal uncertainty, responses to the USCO's notice of inquiry might help establish some guidance while awaiting further developments.

Generative AI and TDM Exception Under EU Law

EU law sets forth a limited list of exceptions to copyright owners' monopoly of use. While most of the current exceptions do not adequately grasp the issues at play in the context of generative AI training,[56] the ongoing debate has shed light on (i) the exception for transient or incidental acts of reproduction[57] and (ii) exceptions for data mining.[58] As further detailed in Section II.B. above, even though the transient or incidental reproduction exception might apply depending on the actual technology used, it is unlikely to justify most uses of copyrighted works in generative AI training. However, the exception for text and data mining (TDM), defined as "any automated analytical technique aimed at analyzing text and data in digital form in order to generate information,"[59] is broad enough to cover the ingestion of copyrighted works for purposes of generative AI training, including for commercial purposes.[60]

A distinction must be made here between the TDM exception for scientific research, which essentially benefits "institutions that provide a cultural or public service in the interest of society on a non-for-profit basis,"[61] and the general TDM exception. The latter allows the "reproductions and extractions of lawfully accessible works [. . .] for the purposes

of text and data mining,"[62] and covers "any TDM activity, whether non-profit or for profit, as long as it falls under the definition of TDM,"[63] with no limitation as to its beneficiaries. However, the broad scope of the general TDM exception is limited in two ways: First, the works reproduced for TDM purposes must be lawfully accessible, meaning "based on an open access policy [. . .] contractual arrangements [. . .] other lawful means" or "freely available online,"[64] and second, the copyright owners can exercise their opt-out right[65] and "expressly reserve [. . .] [the right to use the work for TDM purposes] in an appropriate manner, such as machine-readable means in the case of content made publicly available online."[66] Furthermore, as with fair use in the U.S., the breadth of this exception must be analyzed under the light of the Berne Convention.

However, these existing legal guardrails are criticized by copyright owners for their lack of practical efficiency, heightened by AI developers' lack of transparency.[67] This need for transparency has recently been addressed by the EU Artificial Intelligence Act (AI Act), formally adopted by the European Parliament on March 13, 2024. While copyright is not the focus of the AI Act, its most up-to-date version requires AI developers "that place general-purpose AI models on the Union market"[68] to "put in place a policy to comply with Union copyright law, and in particular to identify and comply with, including through state of the art technologies, [copyright owners' opt-out]"[69] and to "draw up and make publicly available a sufficiently detailed summary about the content used for training [. . .], according to a template provided by the AI Office."[70] Furthermore, "[w]hile 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."[71]

In light of the above, while the general TDM exception applies to the generative AI training process, AI developers will arguably be required to obtain authorizations from copyright owners in most cases, given the lawful access requirement and copyright owners' right to opt-out. Pending entry into force of the AI Act, both copyright owners and AI developers are coming up with proposed opt-out mechanisms, the effectiveness of which remains to be seen.

Conclusion

While outcomes may vary depending on the technology used, the generative AI training process likely implicates the reproduction right in both the U.S. and the EU. Thus, unless an exception such as the U.S. fair use doctrine or the EU general TDM exception applies, AI developers will likely be required to obtain authorizations from the copyright owners to use their work as training data.

Even though some AI developers recognize the need to give "creators greater control over how their content is used in AI model training"[72] and started entering into agreements with copyright owners,[73] the implementation of an effective authorization system faces several challenges. First, the current lack of transparency from AI developers on the works used as training data, coupled with the existence of pirated copies of copyrighted works on the Internet, renders the effective exercise and monitoring of an opt-out right virtually impossible for copyright owners. Second, considering the volume of data needed to develop an efficient AI system, obtaining authorizations from each copyright owner would be too burdensome for AI developers, especially small market players, leading to a "risk of concentration [where] only a few very well-funded AI publishers [would be able to] access cultural data at a high price."[74] Furthermore, to avoid excessive reduction of the training data set, which could lead to biased algorithms and "cultural submersion [. . .] if AIs are trained on [less diverse] data,"[75] "the negotiated fees must be substantial enough for authors and publishers to find it worthwhile to participate but cannot be unreasonably expensive for AI companies."[76]

New laws and bills are emerging, both in the EU (AI Act) and in the U.S. (AI Foundation Model Transparency Act[77] and Generative AI Copyright Disclosure Act[78] proposals) to address the lack of transparency issue. Furthermore, to implement a more scalable authorization system, solutions proposed on both sides of the Atlantic lean towards a sector-specific collective license system, where copyright owners would be represented by collective management organizations (CMOs) to negotiate deals on their behalves.[79] This solution should soon be illustrated in France, where the Society of Authors, Composers, and Publishers of Music (SACEM) recently opted out of generative AI training to "restore the exclusive rights of creators by making data-mining operations subject to prior authorization."[80] However, the complexity of the negotiation process is illustrated by the *New York Times'* lawsuit against OpenAI and Microsoft, which occurred after months of negotiations.

In light of the above, while both U.S. and EU traditional copyright law principles cover the use of copyrighted work to train generative AI, several challenges still need to be addressed for the effective implementation of regulations that would protect copyright owners' rights and foster innovation. This complex issue will continue to evolve as court rulings and legislative debates unfold and is paradigmatic of the need for copyright law to evolve in line with technological developments.

Flore Brunetti is a French and New York-trained attorney with dual membership in the New York and Paris bars, whose practice focuses on intellectual property and entertainment law. After graduating from New York University School of Law, she practiced in France for a few years, where her work focused on international productions for a U.S. clientele. Her expertise in navigating cross-border matters and her passion for the entertainment industry led her to recently relocate to the United States to pursue the next steps of her legal career. Flore is a member of EASL's Motion Pictures Committee and of the European Affairs Committee of the New York City Bar Association.

Endnotes

1.   U.S. Copyright Office, Artificial Intelligence and Copyright Notice of Inquiry and Request for Comments, 88 Fed. Reg. 59942, 59948-49 (2023).

2.   Open AI LP, Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation, PTO-C-2019-0038 at 2 (2019).

3.   *Id.*

4.   Adobe Inc., Comments on the U.S. Copyright Office Notice of Inquiry and Request for Comments on Artificial Intelligence and Copyright at 2 (2023).

5.   OpenAI LP Comments, *supra* note 2 at 1-2.

6.   The Authors Guild, *Open Letter to Generative AI Leaders*, (Jul. 18, 2023), *available at* https://actionnetwork.org/petitions/authors-guild-open-letter-to-generative-ai-leaders/.

7.   Le Monde, *Intelligence artificielle : « Construisons dès aujourd'hui une IA de rang mondial respectueuse de la propriété littéraire et artistique »*, (Sept. 29, 2023), *available at* https://www.lemonde.fr/idees/article/2023/09/29/intelligence-artificielle-construisons-des-aujourd-hui-une-ia-de-rang-mondial-respectueuse-de-la-propriete-litteraire-et-artistique_6191655_3232.html.

8.   Rachel Kim, *AI and Copyright in 2023: In the Courts*, Copyright Alliance, (Jan. 4, 2024), *available at* https://copyrightalliance.org/ai-copyright-courts/.

9.   Thomas H. Davenport and Randy Bean, *The Impact of Generative AI on Hollywood and Entertainment*, MIT Sloan Management Review, (Jun. 19, 2023), *available at* https://sloanreview.mit.edu/article/the-impact-of-generative-ai-on-hollywood-and-entertainment/.

10. Alex Engler, *The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment*, The Brookings Institution, (Apr. 25, 2023), *available at* https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/.

11. 17 U.S.C. § 106(1).

12. Directive 2001/29/EC on the harmonization of certain aspects of copyright and related rights in the information society dated May 22, 2001, § 2.

13. Alexandra Bensamoun and Joëlle Farchy, Final Report of the Artificial Intelligence and Culture Mission, French Higher Council of Literary and Artistic Property (CSPLA) at 48 (2020).

14. 17 U.S.C. § 101.

15. Directive 2001/29/EC, *supra* note 12 at § 2.

16. In-Depth Analysis for the JURI Committee, The exception for text and data mining (TDM) in the proposed Directive on Copyright in the Digital Single Market at 5 (2018).

17. *Id.*

18. CSPLA Report, *supra* note 13 at 47-49.

19. United States Copyright Office, DMCA Section 104 Report at 123 (2001).

20. *Id.* at 122-23.

21. CJEU, *Infopaq International A/S v. Danske Dagblades Forening*, C-5/08, July 16, 2009, § 38.

22. *Id.* at § 50.

23. Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market dated April 17, 2019, § 4.

24. Jenny Quang, Does Training AI Violate Copyright Law? Berkeley Technol. Law J. Vol 36:1407, 1407, 1413 (2021).

25. 17 U.S.C. § 101.

26. Directive 2001/29/EC, *supra* note 12.

27. CJUE, *Pelham*, C-476/17, July 29, 2019, § 29.

28. *Id.* at § 31.

29. European Commission, Study on copyright and new technologies: Copyright data management and artificial intelligence, at 21 (2022).

30. CSPLA Report, *supra* note 13 at 49.

31. 17 U.S.C. § 101.

32. Directive 2001/29/EC, *supra* note 12 at § 5.1.

33. Open AI LP Comments, *supra* note 2 at 2.

34. Stability AI, Response to United States Copyright Office Inquiry into Artificial Intelligence and Copyright at 13 (2023).

35. *Samuel v. Eichner*, Matt Grillo and Josh Graham, The Copyright Implications of Generative AI, AIPLA Annual Meeting 2023, at 10-12.

36. USCO DMCA Report, *supra* note 19 at 111.

37. *Id.* at 110-12.

38. Directive 2001/29/EC, *supra* note 12 at § 5.1.

39. CJEU, *Infopaq International A/S v. Danske Dagblades Forening*, C-302/10, January 17, 2012, § 26-27.

40. CJEU C-5/08, *supra* note 21 § 64.

41. *Id.* at § 50.

42. In Depth Analysis for the JURI Committee, *supra* note 16 at 6.

43. 17 U.S.C. § 107.

44. Open AI LP Comments, *supra* note 2 at 5.

45. Kristian Stout, Geoffrey A. Manne, and Emily Corbeille, ICLE Comments on Artificial Intelligence and Copyright, at 11 (2023).

46. U.S. Supreme Court, *Andy Warhol Foundation for the Visual Arts Inc. v. Goldsmith et Al.*, May 18, 2023, at 12.

47. District Court of Delaware, *Thomson Reuters Enterprise Center GmbH and West Publishing Corp. v. Ross Intelligence Inc.*, September 25, 2023, at 16.

48. *Id.* at 16-17.

49. *Id.*

50. *Id.* at 20.

51. Cala Coffman, *Does the Use of Copyrighted Works to Train AI Qualify as a Fair Use*, Copyright Alliance, (Apr. 11, 2023), *available at* https://copyrightalliance.org/copyrighted-works-training-ai-fair-use/.

52. Authors Guild Open Letter, *supra* note 6.

53. Google Inc., Artificial Intelligence and Copyright, COLC-2023-0006 at 10-11 (2023).

54. Open AI LP Comments, *supra* note 2 at 8.

55. Berne Convention for the protection of literary and artistic works dated September 9, 1886, § 9(2).

56. CSPLA Report, *supra* note 13, at 50.

57. Directive 2001/29/EC, *supra* note 12 at § 5.1.

58. Directive (EU) 2019/790, *supra* note 23 § 4.1.

59. *Id.* at § 2(2).

60. E.U. Commission Study, *supra* note 29 at 195.

61. *Id.* at 193.

62. Directive (EU) 2019/790, *supra* note 23 § 4.1.

63. E.U. Commission Study, *supra* note 29 at 195.

64. Directive (EU) 2019/790, *supra* note 23 Recital 14. *See also* E.U. Commission Study, *supra* note 29 at 195.

65. This opt-out right does not apply to the TDM exception for scientific research.

66. Directive (EU) 2019/790, *supra* note 23 § 4.3.

67. Le Monde, *supra* note 7.

68. European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, Recital 106.

69. *Id.* at § 53.1.(c).

70. *Id.* at § 53.1.(d).

71. *Id.* at Recital 107.

72. Adobe Comments, *supra* note 4, at 4.

73. Le Monde, *Le Monde and Open AI sign partnership agreement on artificial intelligence*, (Mar. 13, 2024), *available at* https://www.lemonde.fr/en/about-us/article/2024/03/13/le-monde-signs-artificial-intelligence-partnership-agreement-with-open-ai_6615418_115.html.

74. French Artificial Intelligence Commission, *Our AI: Our Ambition for France*, at 106 (2024).

75. *Id.*

76. The Authors Guild, Comments on Artificial Intelligence and Copyright; n°2023-6 at 14 (2023).

77. Emilia David, *AI companies would be required to disclose copyrighted training data under new bill*, The Verge (Dec. 22, 2023) https://www.theverge.com/2023/12/22/24012757/ai-foundation-model-transparency-act-bill-copyright-regulation.

78. The Authors Guild, Authors Guild Supports Groundbreaking Bill to Create AI Transparency Between Creators and Companies (Apr. 12, 2024) https://authorsguild.org/news/ag-supports-generative-ai-copyright-disclosure-act/.

79. The Authors Guild, *supra* note 76 at 9-10.

80. SACEM, In Favour of Virtuous, Transparent, and Fair AI, Exercises Its Right to Opt-Out, (Oct. 12, 2023), https://societe.sacem.fr/en/news/our-society/sacem-favour-virtuous-transparent-and-fair-ai-exercises-its-right-opt-out.