Copyright Conundrums: Rethinking Authorship Rights Against AI in Open Access

Abstract

The present article reviews the current regulatory framework concerning established statutory exceptions enabling generative AI to derive protected works in training data. It advocates the extended application of such prominent exceptions in the United States and the European Union in light of the novelty of generative AI. Nonetheless, acknowledging that existing regulations primarily focus on rightsholders without addressing impacts on the greater data commons, the author highlights the issue of governing open access commons. Beyond evaluating individual rights, the article highlights the broader issue of governing the open-access commons in the context of generative AI. By building on current open-access licensing frameworks, the author proposes a participatory dialogue approach to establish contractual norms that promote ethical attribution practices and preserve the quality of shared data resources.

I. Introduction

Our data-centric environment presents a paradox—while enveloped in an abundance of information, many face considerable obstacles to accessing it. Though attributable to driving societal advancement, our data repository is predominantly held by private interests. The development of large language models (LLMs), however, appears to put a spanner in the works, offering new avenues for data access while exacerbating inequality for those unable to leverage it. Amidst the initial anxiety of educational institutions for Artificial intelligence (AI) as a 'cheating tool', generative AI has and will have larger implications for the knowledge economy, analogous to the Industrial Revolution.[1] Although generative AI stands at the forefront of innovation enabling access such as through conversational interfaces, concerns regarding data governance are warranted due to its capacity to repurpose data on a massive scale.

This 'existential threat' underlying the rapid development of AI thus raises fundamental quandaries of regulation to ensure fair and equitable use.[2] Though the notion of derivative works has existed in intellectual property law, the scale of creation and dissemination

---

[1] Hannah Devlin, 'AI "Could Be as Transformative as Industrial Revolution"' *The Guardian* (3 May 2023) <https://www.theguardian.com/technology/2023/may/03/ai-could-be-as-transformative-as-industrial-revolution-patrick-vallance>

[2] 'AI Poses "Risk of Extinction," Industry Leaders Warn - The New York Times' <https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html>

exercised by developments in AI technology is unprecedented;[3] its function to mirror human expression has characterised debates pertaining to the 'death of the author', engendering concern for those engaging in intellectual work and critical inquiry, including the present author. Ultimately, generative AI raises concern for equitable access and distribution of the Internet data commons undermining open access and creative common initiatives.

In light of the transformative impact of generative AI on the knowledge economy, numerous jurisdictions have implemented and extended legal frameworks to ensure transparency *vis-a-vis* data collection and intellectual property rights. Nonetheless, contention regarding the sufficiency of such policies has dominated recent scholarship. It is apparent that a fundamental tension resides between ensuring compensation to authors of original works and granting generative AI sufficient corpus to offer its innovative services.

These works however have tended to restrict their scope to copyright law lacking consideration of its associated risks against the greater data commons. The unprecedented reproduction of data is notably prone to proliferate unverified information. This study thus uniquely leverages the analysis of the data ecosystem to situate the academic discourse towards the sustainability of the open-access data commons. The following Section of the article will survey existing intellectual property provisions across major jurisdictions such as the United States and the European Union, as listed as applicable to LLMs. The author endeavours to assess the legal protections against copyright infringement of AI, with particular scrutiny on the corpus of data in training LLMs. Section III will evaluate the regulation of generative AI in relation to its complementary benefits in open-access models.

II. Legal Framework of Intellectual Property Regulations

Analogous to human inspiration, AI systems are inherently dependent on prior works upon which their programming has been trained. This process inadvertently involves the reproduction of existing works when the digitalisation of materials is involved. Generative AI is therefore prone to copyright infringement given its reliance upon protected materials. The corpus of training data has thus been a significant avenue of litigation for which the legality of web scraping technology in the 'text and data mining' (TDM) practices of LLMs remains dubious.[4]

---

[3] Simon Chesterman, 'Good Models Borrow, Great Models Steal: Intellectual Property Rights and Generative AI' (11 October 2023) <https://papers.ssrn.com/abstract=4590006>

[4] For a comprehensive compilation of ongoing litigation, see <https://www.bakerlaw.com/services/artificial-intelligence-ai/case-tracker-artificial-intelligence-copyrights-and-class-actions/>

The objection against the training avenues for LLMs is exacerbated by the inclusion of the pirated collections of books and journals within the 'internet-based books corpora' entitled 'Books3' suspected to originate from shadow libraries such as 'Library Genesis, Z-Library, and Bibliotik that circulate via the BitTorrent file-sharing network.'[5]

OpenAI has particularly acknowledged that its algorithms undergo training using extensive datasets that encompass copyrighted materials and that this training procedure necessitates the initial creation of data copies for analysis. While these corporations have recently enabled an 'opt-out' service that enables creators to remove their protected works,[6] the default effect of this service nonetheless implicates a lack of proper authorization that potentially amounts to copyright infringement.

a) The United States and the 'Fair Use' Doctrine

In the United States, the use of protected material to train algorithms arguably violates the exclusive right to reproduction of a copyright owner under 17 U.S.C. §106(1). The provision however stipulates an exception to such under the fair use doctrine. The statutory instrument particularly involves a four-pronged consideration constituting fair use:

> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
> (2) the nature of the copyrighted work;
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
> (4) the effect of the use upon the potential market for or value of the copyrighted work.[7]

Regarding the first prong, *Campbell v. Acuff-Rose Music*[8] specifies that the "transformative" nature of the new creation reduces the importance of commercial incentives that could otherwise argue against a determination of fair use. The threshold is thus whether the work

---

[5] 'Erotica, Atwood, and "For Dummies": The Books Behind Meta's Generative AI - The Atlantic' <https://www.theatlantic.com/technology/archive/2023/09/books3-ai-training-meta-copyright-infringement-lawsuit/675411/>

[6] Kali Hays, 'OpenAI Offers a Way for Creators to Opt out of AI Training Data. It's so Onerous That One Artist Called It "Enraging."' (Business Insider) <https://www.businessinsider.com/openai-dalle-opt-out-process-artists-enraging-2023-9>

[7] 17 U.S.C. §107

[8] 510 U.S. 569 (1994)

has “superseded” the original creation ‘with [a] further purpose or different character…that has [a] new expression, meaning, or message’.

Pertaining to this notion, OpenAI has particularly submitted that its practices of training algorithms are “transformative” on two grounds 1) its practices are “non-expressive” and 2) its output is different.[9] The first ground presumes that protected works are created for consumption and appreciation of the author’s expression, which copying such works for training AI systems is “fundamentally different”- the object of the process of creating a useful AI system is submitted to be different than the original intent of the creator. The second contention posits that the output generated by OpenAI's AI system exhibits substantial distinction from original works, given the algorithm’s utilisation of an extensive collection of works that result in generative outputs that possess only limited commonalities with each respective work.

However, it is observed that the works of generative AI systems are ultimately made for “consumption and appreciation” of the expression of AI agents. The first ground comes into scrutiny when considering that the algorithm uses protected works as training data with the programmed intent of creating or replicating an expressive piece.

The second prong is given little consideration.[10] The third prong suggests that ‘amount and substantiality’ are not considered in the context of making a copy but its use for “being made accessible to a public for which it may serve as a competing substitute.”[11] Drawing upon the analogous facts in *Google* where copying protected books for the search engine to display excerpts of books constitutes fair use, OpenAI asserts that its practices are opaque. This fact indeed favours a finding of fair use. Moreover, for the fourth prong, OpenAI distances the effects of generative AI in undervaluing original works as a broader concern about “the relationship between automation, labor, and economic growth”, which distributive concerns can be addressed via more efficient policies.[12] Authors nor the market for copyrighted works is harmed due to including original works in the corpus. The fourth prong underscores a normative attribute that would be more appropriately addressed by economic analysis. AI being transformative and opaque, likely, *a fortiori,* has a stronger claim. Thus fair use is

[9] https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf
[10] The Authors Guild Inc., et al. v. Google, Inc. 804 F.3d 202
[11] See Google, 804 F.3d at 220.
[12] Louis Kaplow & Steven Shavell, Why the Legal System Is Less Efficient than the Income Tax in Redistributing Income, 23 J. Legal Stud. 667 (1994)

considerably a flexible doctrine tending to enable copyrighted works to be used reasonably in the context of AI.

It is notable to consider that owners of copyrighted works have actionable claims if the AI program both "(1) had access to their works and (2) created 'substantially similar' outputs".[13] According to *Sturdza v. United Arab Emirates*[14], the plaintiff must prove the alleged infringer has "actually copied" including circumstantially by evidence that they "had access to the work". For (1), it is well known that the AI program was trained by copyrighted work. The question thus lies in (2) where the dispute arises as to whether generative AI creates "substantially similar" output to original works. *Shaw v. Lindheim*[15] and *Boisson v. American County Quilts and Linens, Inc.*[16] respectively articulate similar but different thresholds for "a substantially similar total concept and feel" or the failure of an "ordinary reasonable person…to differentiate between the two works" respectively. In the context of AI, the "substantial similarity" test is factually based on the nature of the output in relation to the copyrighted work. Thus, the American system has yet to address copyright infringement during training explicitly but has adequate guidelines for cases of direct infringement.

b) The European Union Artificial Intelligence Act 2024

The European Union's response to the concerns of training datasets consisting of protected works is contextualised within the recent authoritative guidelines of the Artificial Intelligence Act.[17] Nonetheless, the AI Act itself does not impose additional conditions to infringement particular to LLMs but reiterates the notion that "any use of copyright protected content requires the authorisation of the rightsholder concerned unless [existing] relevant copyright exceptions and limitations apply"[18]. Thus, the AI Act on a general note simply reinforces the preexisting copyright regime such that AI must obtain permission from rightsholders to include data in the training corpus. This is however rather difficult given the scale of datasets used for training; the sheer number of rightsholders involved means great difficulty in mandating AI to obtain consent explicit license from all stakeholders. The issue is exacerbated by the automated web-scraping technology leveraged by LLMs to extract data on an unprecedented scale, rendering regulation and monitoring arduous.

[13] https://crsreports.congress.gov/product/pdf/LSB/LSB10922
[14] 281 F. 3d 1287
[15] 919 F. 2d 1353
[16] 273 F. 3d 262
[17] 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 ("Artificial Intelligence Act").
[18] Recital no. 105 (n 17).

A solution is to extend the applicability of the European Union derivative works regime to generative AI. Enshrining the Berne Convention where "[any] alterations of a literary or artistic work shall be protected as original works without prejudice to the copyright in the original work"[19], EU copyright law institutes two forms of legal protection afforded to databases. The first, stipulated by Directive 96/9/EC, consists of general copyright protection for databases "constitut[ing] the author's own intellectual creation"[20] with the exclusive right "to authorise temporary or permanent reproduction…any form of distribution…and any display"[21] among others. The second establishes *sui generis* database rights that entitle the creator of a database to "prevent extraction and/or re-utilization of the whole or of a substantial part…of the contents of that database" when the creator has made a "substantial investment…to prevent extraction and/or re-utilization"[22]. This principle is reflected in Article 2 of the InfoSoc Directive 2001/29/EC, where the author is granted an "exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form"[23]. Article 15 of Directive 2019/790/EC most recently further specifies the protection of press publications from online infringement.[24] In the additional circumstance that the database is not eligible for both forms of protection, consider *Ryanair Ltd v. PR Aviation BV*[25] where owners are entitled to introduce contractual restrictions constituting a similar *sui generis* copyright via terms and conditions. In the context of generative AI, authors of a database may very well include contractual clauses that prohibit web scrapping even when its content is not *per se* protected by established protections provided by Directive 96/9/EC.

Considering the vast repository of rights, the Act leverages the text and data mining exceptions in the Directive on Copyright in the Digital Single Market ("DSMD") to enhance access. Articles 3(1) and 4(1) particularly make explicit reference to its exception from the existing copyright regime noted above. Particularly, when TDM is carried out for the

---

[19] Article 2(3) Berne Convention for the Protection of Literary and Artistic Works (adopted 14 July 1967, entered into force 29 January 1970) 828 UNTS 221.
[20] Article 3(1) Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases OJ L 77, 27.3.1996, p. 20–28.
[21] Article 5 (n 20).
[22] Article 7(1) (n 20).
[23] 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 OJ L 167, 22.6.2001, p. 10–19.
[24] 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 (Text with EEA relevance.) PE/51/2019/REV/1 OJ L 130, 17.5.2019, p. 92–125.
[25] Case C-30/14 *Ryanair Ltd v PR Aviation BV* [2015]

purposes of scientific research or a non-commercial agenda, Article 3(1) provides an exception irrespective of its infringing capabilities. Secondly, for commercial TDM activities, Article 4(3) introduces an additional exception if the protected materials have 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"[26]. This 'opt-out provision' thus enables rightsholders to reserve the content of their works for TDM purposes expressly—to negotiate licenses with AI developers or otherwise. However, this also means that copyright content not actively managed by rightsholders and made publicly available lawfully is available for TDM processes to train generative AI.

Recital 18 of the directive attempts to clarify this 'opt-out regime' such that reserving rights are most appropriate via machine-readable means such as "metadata and terms and conditions of a website or a service"[27]. Nonetheless, the relative vagueness of 'machine-readable' raises significant interpretative questions to be dealt with in future litigation regarding training LLMs. Generally, it should be understood that the language of the opt-out should be unambiguous insofar as rights reserved by the rightsholder. The provisions further lack sufficient guidance on the particular timeframe the rightsholder is required to express reservation and effectuate the exception; it is unclear whether rightsholders should opt out during the training phase or only when data is publicly available. This condition thus underscores the proposed approach from Novelli et al for heightened obligation for LLMs to autonomously analyse terms of online databases and websites to differentiate between content that has been expressly reserved and those freely accessible and extractable for the data corpus.[28] It is however worth considering the lack of standardised protocols for machine-readable opt-outs. An avenue would be to leverage protocols where rightsholders are able to assert their reservations across multiple websites; particularly via unit-based identifiers such as watermarking or metadata that do not prejudice the nature of the opt-out regardless of the location of content.[29] The proactive 'opt-out' measure from OpenAI is thus possibly attributable to a threshold of care.

Although the AI Act defers to the DSMD on the question of exempting LLMs, its drafters have contemplated the practicalities of verifying compliance. Article 53(1)(a) comprises an

---

[26] Article 4(3) (n 24).
[27] Recital no. 18 (n 24).
[28] Claudio Novelli and others, 'Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity' [2024] SSRN Electronic Journal.
[29] Kalpana Tyagi, 'Copyright, Text & Data Mining and the Innovation Dimension of Generative AI' (2024) 19 Journal of Intellectual Property Law & Practice 557.

obligation to provide "technical documentation of the model…upon request, to the AI Office and the national competent authorities" pertaining to the "information on the data used for training, testing and validation"[30]. The Act further stipulates two specific requirements for compliance with the copyright regime. Article 53(1)(c) notes that AI models must "through state-of-the-art technologies, [comply with] a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790"[31]. This implies that the proactive 'opt-out' measure is likely attributable to the threshold of care required for developers to systematically adapt modern technology to respect authors' reservation of rights. Article 53(1)(d) finally imposes a transparency provision to "make publicly available a sufficiently detailed summary about the content used for training"[32] according to templates from the AI Office. The threshold for such a 'sufficiently detailed summary' is however ambiguous and subject to the AI Office. Recital 107 provides a preliminary portrayal of the standard where the "generally comprehensive" summary must list "the main data collections or sets that went into training the model"[33]. Provided that the Recital purports to "increase transparency on the data that is used in the pre-training and training", the summary must encompass various types of data used throughout regardless of the technicality of whether such is utilised to directly influence the model during its training stage.

Nonetheless, Article 4(2) of the DSMD underscores the reproduction and extraction of content can only be "retained for as long as is necessary for the purposes of text and data mining."[34] A narrow interpretation of such a provision implies that LLMs must delete copyrighted content as soon as the training phase is completed.[35] Novelli thus notes the imperative for a "broad normative interpretation" of the TDM process.[36] Moreover, the demarcation between Research and commercial TDM is insufficient in the context of AI. Consider the legality of licensing the use of a generative AI model trained on protected data. Though developed using data from research exemptions, the LLM is nonetheless independently commercialisable with significant prejudice to the rightsholder.[37]

Additionally, Article 7(2) applies the three-step test from Article 5(5) of the InfoSoc Directive. The exception and limitations to Articles 3 and 4 are only applicable in "(1) certain special

[30] Annex XI (n 17).
[31] Article 53(1)(c) (n 17).
[32] Article 53(1)(d) (n 17).
[33] Recital no. 107 (n 17).
[34] Article 4(2) (n 24).
[35] Giuseppe B Abbamonte, 'The Application of the Copyright TDM Exceptions and Transparency Requirements in the AI Act to the Training of Generative AI' 46 European Intellectual Property Review 479.
[36] Novelli and others (n 23) 16.
[37] Abbamonte (n 24).

cases which (2) do not conflict with a normal exploitation of the work or other subject-matter and (3) do not unreasonably prejudice the legitimate interests of the rightholder."[38] The Convention is thus noted to prescribe a cumulative test for a restrictive application of the doctrine.[39] This suggests that the TDM exception is strictly unable to legitimise reproductions that substitute or economically compete with the protected material used for AI training, which is noticeably the concern associated with generative AI.[40]

Such a narrow provision is however problematic considering the multiplicity and novelty of rights associated with data mining. Divergent from the 'fair use' provision with "comparable abstract criteria"[41], the EU conception of the test is inherently narrow thereby restricting courts from synthesising additional permissions on an *ad hoc* basis. Senftleben critiques that the provision neither offers sufficient legal certainty for users of copyrighted material given the relative vagueness of 'normal exploitation' and 'legitimate interest' nor provides sufficient flexibility to accommodate emerging cultural, social and economic needs for data mining.[42] In *Pelham v Hütter and Schneider-Esleben*[43], the Court of Justice of the European Union (CJEU) has barred the external balancing of competing fundamental rights beyond the copyright *acquis* in the three-step test. The Court notably problematically alleges that all considerable rights and interests, in particular freedom of artistic expression, are to be found within the Directive itself, restricting all possible avenues of discretion within the rigid structure of the precisely defined test. In *Pelham*, the Court with such reasoning held that Member States are not entitled to further exceptions other than those provided for in the InfoSoc Directive; the German 'free use' provision or an open-ended norm is thus incompatible with EU law.

Furthermore, Courts have problematically conducted the balancing exercise within the constraints of the Directive in the presumption that not 'conflict[ing] with a normal exploitation' and not 'unreasonably prejudic[ing] legitimate interests' offered sufficient consideration for freedom of information and expression among others. Note that the

---

[38] Article 7(2) (n 23).
[39] Eleonora Rosati, 'No Step-Free Copyright Exceptions: The Role of the Three-Step in Defining Permitted Uses of Protected Content (Including TDM for AI-Training Purposes)' (10 November 2023) <https://papers.ssrn.com/abstract=4629528>
[40] Novelli and others (n 23) 16.
[41] Martin R.F. Senftleben, "Comparative Approaches to Fair Use: An Important Impulse for Reforms in EU Copyright Law", in: Graeme B. Dinwoodie (ed.), Methods and Perspectives in Intellectual Property, Cheltenham: Edward Elgar 2013, 30–67.
[42] M Senftleben, *EU Copyright 20 Years After the InfoSoc Directive – Flexibility Needed More Than Ever* (Cheltenham Edward Elgar Publishing 2022) 187 <https://dare.uva.nl/search?identifier=ed140f95-68f3-4811-9acb-503c2d888ec6>
[43] Case C-476/17 *Pelham GmbH and Others v Ralf Hütter and Florian Schneider-Esleben* [2019] 624

three-step test only enables judges to further restrict narrow exception privileges; the Court is simply unable to conduct a holistic balancing exercise when simply construing a provision negatively. This is exacerbated by the mantra of strict interpretation established in *Infopaq International A/S v Danske Dagblades Forening*.[44] The defendant is not only responsible for the burden to prove compliance with the copyright limitation but also for adherence to the broader standards of the three-step test; this negative consideration against the socially beneficial TDM practices in fact undermines legal certainty for existing limitations and exceptions. AI developers are simply unable to rely on the statutory provisions of copyright exception granted that courts are able to deem generative AI practices infringing according to the flexible consideration of 'legitimate interests' especially in the context of the sceptic view[45] against AI development.

The development of case law pertaining to LLMs would thus lean towards the limits of the text and data mining exceptions of the DSMD, particularly on the nature of copying when extracting data for model training. The relatively restrictive EU doctrine would thus centre upon the demarcation between the transitory nature of copying for extracting non-protectable ideas from data or practices that have undermined the essence of the protected work.

III. Generative AI and Open Access Commons

The discourse on regulating the training process above has predominantly centred on the legitimate interests of rightsholders. In the context of the greater information environment, Generative AI poses significant systemic risks to the integrity of the information available on the internet. The proliferation of low-value content, biased material and general disinformation is attributed to its speed of generation, low cost and high accessibility for internet users. Particularly, content filters are determined in a 'black box' with limited human intervention to calibrate input.[46] Generated information thus becomes more homogenous as LLMs provide outputs based on the occurrence of the data irrespective of its falsity, causing the duplication of content from identical sources. Outputs are furthermore difficult to evaluate as quality assurance of available data generated by AI is difficult without developing a centralised auditing framework.[47] Prior research has suggested that

[44] CJEU, 16 July 2009, case C-5/08, Infopaq, paras 56–57 ("provisions of a directive which derogate from a general principle established by that directive must be interpreted strictly")
[45] Arne Bewersdorff and others, 'Myths, Mis- and Preconceptions of Artificial Intelligence: A Review of the Literature' (2023) 4 Computers and Education: Artificial Intelligence 100143.
[46] Saffron Huang and Divya Siddarth, 'Generative AI and the Digital Commons' (arXiv, 20 March 2023) <http://arxiv.org/abs/2303.11074>.
[47] ibid.

text-to-image generation has notably exacerbated demographic stereotypes[48] and replicated discriminative biases[49].

This engenders a negative feedback loop where the publishing of poor-quality and often inaccurate data and information becomes future datasets for training LLMs, resulting in worsening output as AI perpetuates a self-replicating practice of leveraging upon easily accessible bot-generated content. The crisis impacting the information environment is exacerbated by authors being disincentivised from contributing to the digital commons; authors are less likely to endorse open-access models provided that their work is susceptible to exploitation in an ambiguous regulatory landscape. As low-value output joins the commons and becomes datasets for newer LLMs, there is a significant risk of the corruption of the information commons. Such have greater implications for democratic discourse and voter turnout.

The exploitation of open-access authorial works in training datasets is therefore noted to resemble Garrett Hardin's 'tragedy of the commons' where online data is rivalrous but non-excludable, enabling developers to free-ride and corrupt the value of shared online resources for the collective.[50] The data ecosystem will be overused and underinvested, leading to the eventual depletion of quality information replaceable by generated content. In response to such tragedy, scholars of the Ostrom school have proposed approaches in the collectively determined rules for managing shared resources for the right to exclude non-members.[51] This is however hard to achieve in the data ecosystem given the accessibility of the internet and the difficulty of restricting data information. Alternatively, the school of the open commons suggested the management of informational resources on the basis of 'symmetric use privileges' available to all rather than relying on exclusive collective or individual proprietary rights.[52]

One may refer to the existing Creative Commons licensing regime (i.e. Copyleft licensing and ShareAlike clauses) that provides a standardised manner of granting public permission to

---

[48] See Federico Bianchi and others, 'Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale', 2023 ACM Conference on Fairness, Accountability, and Transparency (2023) <http://arxiv.org/abs/2211.03759>.
[49] See Vered Shwartz and Yejin Choi. Do neural language models overcome reporting bias? In Proceedings of the 28th International Conference on Computational Linguistics, pp. 6863–6870, 2020.
[50] See Garrett Hardin, 'The Tragedy of the Commons' (1968) 162 Science 1243.
[51] See Elinor Ostrom, *Governing the commons: The evolution of institutions for collective action* (Cambridge University Press 1990).
[52] See Yochai Benkler, 'Open-Access and Information Commons' in Francesco Parisi (ed), Yochai Benkler, *The Oxford Handbook of Law and Economics* (Oxford University Press 2017)

use creative work. The licence, in general, serves three purposes: *firstly* to stipulate permitted uses of creative works; *secondly* to require users who would otherwise infringe copyright to give attribution to the copyright owner[53], and *thirdly* to impose a reciprocity condition on users that modify and distribute the data to license their derivative work in open access. Consequently, one may observe that the regime requires 'replenishing' the commons as the license stipulates a condition of synthesising quality output derived from protected works.[54] For instance, the Attribution-ShareAlike (CC BY-SA 4.0) licence enables the user to remix the work even for commercial purposes as long as they provide credit and license their derivative works under identical conditions.[55]

Nevertheless, as noted above, it is disproportionately difficult for AI to recognize all the works incorporated within its corpus, given that the capacity to generate text is predicated on having processed billions of existing works.[56] Generative AI faces challenges in determining the source of its output given the nature of information symmetry across internet sources.[57] Certain authors thus have analogised data training with the practice of organic learning, where the assimilation of information is viewed as a natural process of acquiring knowledge from the commons.[58] Such claims reflect the greater discourse regarding the idea/expression dichotomy, particularly when generative AI output is confined to general concepts from an array of protected works that, on their own, do not possess copyright protection, provided the training data does not directly attribute to the original creative choices of the authors.[59] This practice is reinforced by the international copyright regime, particularly via Article 9(2) of the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) and Article 2 of the WIPO Copyright Treaty. Therefore, in cases where output is unattributable to the unique quality of a singular author, copyright law does not offer a basis for enabling open-access licensing obligations.

---

[53] Michael W Carroll, 'Creative Commons and the Openness of Open Access' (2013) 368 New England Journal of Medicine 789.
[54] Beatriz Botero Arcila, 'Who Owns Generative AI Training Data? Mapping the Issue and a Way Forward' (10 January 2024) <https://papers.ssrn.com/abstract=4744202>
[55] (n 51).
[56] See Zeynep Ülkü Kahveci, 'Attribution Problem of Generative AI: A View from US Copyright Law' (2023) 18 Journal of Intellectual Property Law & Practice 796.
[57] See Ekin Akyurek and others, 'Towards Tracing Knowledge in Language Models Back to the Training Data' in Yoav Goldberg, Zornitsa Kozareva and Yue Zhang (eds), *Findings of the Association for Computational Linguistics: EMNLP 2022* (Association for Computational Linguistics 2022) <https://aclanthology.org/2022.findings-emnlp.180>.
[58] See Andres Guadamuz, 'Creative Commons and AI Training' (*TechnoLlama*, 19 November 2023) <https://www.technollama.co.uk/creative-commons-and-ai-training>; Beth Montague-Hellen, 'Empowering Knowledge through AI: Open Scholarship Proactively Supporting Well Trained Generative AI' (2024) 37 Insights <https://insights.uksg.org/articles/10.1629/uksg.649>.
[59] For an elucidation of the idea/expression dichotomy in the context of generative AI, see Mark A Lemley and Bryan Casey, 'Fair Learning' (2021) 99 Texas Law Review 743.

Szkalej and Senfleben thus propose a contractual approach to oblige AI developers using CC-licensed works to accept their associated obligations.[60] In the EU context, the authors suggest leveraging the 'opt-out regime' per Article 4(3) DSMD to effectuate the strategic reservation of copyright. Following such an approach, creators are engendered to extend contractual SA obligations relying on copyright infringement as leverage for negotiation. This would involve the establishment of a 'contractual chain-binding model' where developers are incentivised to embed CC-licensed obligations in the contractual terms of using generative AI systems; for instance, simply refusing internet users to use the conversational interface unless agreeing to be bound by the license. This tailor-made contractual license thus enables AI developers broad freedom to use open access resources for training purposes while adhering to the 'symmetric use privileges' to 'replenish' the commons. On its merits, the license obliges the developer to make available the final trained model under licensing conditions beneficial to the rightsholder irrespective of whether the output itself contains copyright-protected material.[61]

However, the copyright status (i.e. nature of protection afforded to ideas) is decisive for further downstream use. In the likely scenario that the AI output lacks copyrightable elements, it is impossible to enforce CC licensing conditions when there is no longer any original work to protect.[62] Simply put, the idea/expression dichotomy remains unmitigated for users who are not obligated to adhere to the original licensing terms. In this regard, there is perhaps merit in reimagining consent in the context of 'social licenses' to supplement current attempts to extend open access licenses. Verhulst, Sandor and Stamm propose the participatory engagement of stakeholders and the broader public to engender trustworthy industry practices for democratic governance of data reuse.[63] Developing industry practice whereby downstream users cite works according to the mandatory guidelines of generative AI platforms, in dialogue with authors, would mitigate the propensity of users to circumvent licensing objectives in maintaining the integrity of the open access data commons.

---

[60] Kacper Szkalej and Martin Senftleben, 'Generative AI and Creative Commons Licences: The Application of Share Alike Obligations to Trained Models, Curated Datasets and AI Output' (20 June 2024) <https://papers.ssrn.com/abstract=4872366>.
[61] ibid.
[62] ibid.
[63] Stefaan Verhulst, Laura Sandor and Julia Stamm, 'The Urgent Need to Reimagine Data Consent' [2023] Stanford Social Innovation Review <https://ssir.org/articles/entry/the_urgent_need_to_reimagine_data_consent>.

IV. Conclusion

The article has surveyed the current regulatory landscape pertaining to established statutory exceptions permitting generative AI to utilise protected works. It suggests that generative AI is nonetheless subject to existing copyright regimes. In the United States, generative AI output is likely considered 'transformative' to effectuate the fair use exception. In contrast, the European Union imposes greater obligations in light of the 'opt-out provision' supplemented by novel transparency provisions in the AI Act.

Nonetheless, the existing regimes tend to focus on protecting rightsholders lacking subsequent examination of the harm of generative AI upon the internet data ecosystem. Beyond the adjudication of individual entitlements, the article addresses the greater concern for the governance of the open-access commons in light of generative AI. Leveraging upon existing open-access licensing regimes, the author suggests the framework of participatory dialogue to establish contractual norms of adhering to ethical attribution practices and maintaining the data quality of commons.

V. Bibliography

Abbamonte GB, 'The Application of the Copyright TDM Exceptions and Transparency Requirements in the AI Act to the Training of Generative AI' 46 European Intellectual Property Review 479

'AI Poses "Risk of Extinction," Industry Leaders Warn - The New York Times' <https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html>

Akyurek E and others, 'Towards Tracing Knowledge in Language Models Back to the Training Data' in Yoav Goldberg, Zornitsa Kozareva and Yue Zhang (eds), *Findings of the Association for Computational Linguistics: EMNLP 2022* (Association for Computational Linguistics 2022) <https://aclanthology.org/2022.findings-emnlp.180>

Benkler Y, 'Open-Access and Information Commons' in Francesco Parisi (ed), Yochai Benkler, *The Oxford Handbook of Law and Economics* (Oxford University Press 2017) <https://academic.oup.com/edited-volume/28361/chapter/215223447>

Bewersdorff A and others, 'Myths, Mis- and Preconceptions of Artificial Intelligence: A Review of the Literature' (2023) 4 Computers and Education: Artificial Intelligence 100143

Bianchi F and others, 'Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale', *2023 ACM Conference on Fairness, Accountability, and Transparency* (2023) <http://arxiv.org/abs/2211.03759>

Botero Arcila B, 'Who Owns Generative AI Training Data? Mapping the Issue and a Way Forward' (10 January 2024) <https://papers.ssrn.com/abstract=4744202>

Carroll MW, 'Creative Commons and the Openness of Open Access' (2013) 368 New England Journal of Medicine 789

Carroll MW, 'Copyright and the Progress of Science: Why Text and Data Mining Is Lawful' (2019) 53 UC Davis Law Review 893

'Case Tracker: Artificial Intelligence, Copyrights and Class Actions' (*BakerHostetler*) <https://www.bakerlaw.com/services/artificial-intelligence-ai/case-tracker-artificial-intelligence-copyrights-and-class-actions/>

Chesterman S, 'Good Models Borrow, Great Models Steal: Intellectual Property Rights and Generative AI' (11 October 2023) <https://papers.ssrn.com/abstract=4590006>

de la Durantaye K, 'Garbage In, Garbage Out. Regulating Generative AI Through Copyright Law' (28 August 2023) <https://papers.ssrn.com/abstract=4572952>

Devlin H, 'AI "Could Be as Transformative as Industrial Revolution"' *The Guardian* (3 May 2023) <https://www.theguardian.com/technology/2023/may/03/ai-could-be-as-transformative-as-industrial-revolution-patrick-vallance>

'EDPB: "Consent or Pay" Models Should Offer Real Choice | European Data Protection Board' <https://www.edpb.europa.eu/news/news/2024/edpb-consent-or-pay-models-should-offer-real-choice_en>

'Erotica, Atwood, and "For Dummies": The Books Behind Meta's Generative AI - The Atlantic' <https://www.theatlantic.com/technology/archive/2023/09/books3-ai-training-meta-copyright-infringement-lawsuit/675411/>

Gans JS, 'Copyright Policy Options for Generative Artificial Intelligence' (11 April 2024) <https://papers.ssrn.com/abstract=4707911>

González-Solar L and Fernández-Marcial V, 'Sci-Hub, a Challenge for Academic and Research Libraries' (2019) 28 El Profesional de la Información <https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/epi.2019.ene.12>

Guadamuz A, 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs' (26 February 2023) <https://papers.ssrn.com/abstract=4371204>

Guadamuz A, 'Creative Commons and AI Training' (*TechnoLlama*, 19 November 2023) <https://www.technollama.co.uk/creative-commons-and-ai-training>

Hardin G, 'The Tragedy of the Commons' (1968) 162 Science 1243

Hays K, 'OpenAI Offers a Way for Creators to Opt out of AI Training Data. It's so Onerous That One Artist Called It "Enraging."' (*Business Insider*) <https://www.businessinsider.com/openai-dalle-opt-out-process-artists-enraging-2023-9>

'House of Lords - Large Language Models and Generative AI - Communications and Digital Committee' <https://publications.parliament.uk/pa/ld5804/ldselect/ldcomm/54/5411.htm#footnote-046>

Huang S and Siddarth D, 'Generative AI and the Digital Commons' (arXiv, 20 March 2023) <http://arxiv.org/abs/2303.11074>

Kahveci ZÜ, 'Attribution Problem of Generative AI: A View from US Copyright Law' (2023) 18 Journal of Intellectual Property Law & Practice 796

'Large Language Doddle? Generative AI and UK Copyright Law Explained' (*Clifford Chance*) <https://www.cliffordchance.com/content/cliffordchance/expertise/services/intellectual-property/global-ip-updates/2023/q2/large-language-doddle-generative-ai-and-uk-copyright-law-explained.html>

Lemley MA and Casey B, 'Fair Learning' (2021) 99 Texas Law Review 743

Maddi A and Sapinho D, 'On the Culture of Open Access: The Sci-Hub Paradox' (2023) 128 Scientometrics 5647

Montague-Hellen B, 'Empowering Knowledge through AI: Open Scholarship Proactively Supporting Well Trained Generative AI' (2024) 37 Insights <https://insights.uksg.org/articles/10.1629/uksg.649>

Morel V and others, 'Your Consent Is Worth 75 Euros A Year -- Measurement and Lawfulness of Cookie Paywalls', *Proceedings of the 21st Workshop on Privacy in the Electronic Society* (2022) <http://arxiv.org/abs/2209.09946>

Novelli C and others, 'Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity' (arXiv, 14 January 2024) <http://arxiv.org/abs/2401.07348>

Purtova N and van Maanen G, 'Data as an Economic Good, Data as a Commons, and Data Governance' (2024) 16 Law, Innovation and Technology 1

Quang, Jenny;, 'Does Training AI Violate Copyright Law?' <https://lawcat.berkeley.edu/record/1253129>

Quintais JP, 'Generative AI, Copyright and the AI Act' (1 August 2024) <https://papers.ssrn.com/abstract=4912701>

Rosati E, 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspects' <www.europarl.europa.eu/supportinganalyses>

Rosati E, 'No Step-Free Copyright Exceptions: The Role of the Three-Step in Defining Permitted Uses of Protected Content (Including TDM for AI-Training Purposes)' (10 November 2023) <https://papers.ssrn.com/abstract=4629528>

Senftleben M, *EU Copyright 20 Years After the InfoSoc Directive – Flexibility Needed More Than Ever* (CheltenhamEdward Elgar Publishing 2022) <https://dare.uva.nl/search?identifier=ed140f95-68f3-4811-9acb-503c2d888ec6>

Szkalej K and Senftleben M, 'Generative AI and Creative Commons Licences: The Application of Share Alike Obligations to Trained Models, Curated Datasets and AI Output' (20 June 2024) <https://papers.ssrn.com/abstract=4872366>

Tyagi K, 'Copyright, Text & Data Mining and the Innovation Dimension of Generative AI' (2024) 19 Journal of Intellectual Property Law & Practice 557

Verhulst S, Sandor L and Stamm J, 'The Urgent Need to Reimagine Data Consent' [2023] Stanford Social Innovation Review <https://ssir.org/articles/entry/the_urgent_need_to_reimagine_data_consent>