Βιοηθικά

Τόμ. 10, Αρ. 2 (2024)

Bioethica

Εκπαίδευση τεχνητής νοημοσύνης και πνευματική ιδιοκτησία: Θα πρέπει το δίκαιο πνευματικής ιδιοκτησίας να επιτρέπει στις μηχανές να μαθαίνουν;

Pedro Martins Fernandes

doi: 10.12681/bioeth.39041

Βιβλιογραφική αναφορά:

Fernandes, P. M. (2024). Εκπαίδευση τεχνητής νοημοσύνης και πνευματική ιδιοκτησία: Θα πρέπει το δίκαιο πνευματικής ιδιοκτησίας να επιτρέπει στις μηχανές να μαθαίνουν;. *Βιοηθικά*, *10*(2), 8–21. https://doi.org/10.12681/bioeth.39041

AI Training and Copyright: Should Intellectual Property Law Allow Machines to Learn?

Pedro Martins Fernandes[1,2]

[1] University of Lisbon, Lisbon, Portugal.
[2] Intern, National Commission for Bioethics & Technoethics, Greece.

pd.martins.fernandes@gmail.com

Abstract

This article examines the intricate legal landscape surrounding the use of copyrighted materials in the development of artificial intelligence (AI). It explores the rise of AI and its reliance on data, emphasizing the importance of data availability for machine learning (ML) systems. The article analyzes current relevant legislation across the European Union, United States, and Japan, highlighting the legal ambiguities and constraints posed by IP rights, particularly copyright. It discusses possible new solutions, referencing the World Intellectual Property Organization's (WIPO) call for discussions on AI and IP policy. The conclusion stresses the need to balance the interests of AI developers and IP rights holders to promote technological advancement while safeguarding creativity and originality.

Keywords: Artificial Intelligence; copyright law; legal challenges; text and data mining; fair use.

Εκπαίδευση τεχνητής νοημοσύνης και πνευματική ιδιοκτησία: Θα πρέπει το δίκαιο πνευματικής ιδιοκτησίας να επιτρέπει στις μηχανές να μαθαίνουν;

Pedro Martins Fernandes[1,2]

[1] Πανεπιστήμιο Λισαβόνας, Λισαβόνα, Πορτογαλία.
[2] Ασκούμενος, Εθνική Επιτροπή Βιοηθικής και Τεχνοηθικής, Ελλάδα.

Περίληψη

Το άρθρο εξετάζει το σύνθετο νομικό τοπίο για τη χρήση υλικού τεχνητής νοημοσύνης (ΤΝ) που προστατεύεται από πνευματικά δικαιώματα. Διερευνά την ανάπτυξη της ΤΝ και τη σημασία της διαθεσιμότητας δεδομένων για τα συστήματα μηχανικής μάθησης (ML). Αναλύεται η ισχύουσα σχετική νομοθεσία στην Ευρωπαϊκή Ένωση, τις Ηνωμένες Πολιτείες και την Ιαπωνία, με έμφαση στις νομικές ασάφειες και τους περιορισμούς που θέτουν τα δικαιώματα πνευματικής ιδιοκτησίας. Διερευνώνται πιθανές νέες λύσεις, στο πνεύμα της πρόσκλησης του Παγκόσμιου Οργανισμού Διανοητικής Ιδιοκτησίας (WIPO) για την σχέση των προϊόντων ΤΝ και της πολιτικής για τη διανοητική ιδιοκτησία. Το συμπέρασμα τονίζει την ανάγκη εξισορρόπησης των συμφερόντων των προγραμματιστών ΤΝ και των κατόχων δικαιωμάτων διανοητικής ιδιοκτησίας για την προώθηση της τεχνολογικής προόδου με παράλληλη διασφάλιση της δημιουργικότητας και της πρωτοτυπίας.

Λέξεις κλειδιά: Τεχνητή Νοημοσύνη, δίκαιο πνευματικής ιδιοκτησίας, νομικά προβλήματα, εξόρυξη δεδομένων και κειμένου, δίκαιη χρήση.

1. Introduction to the problematic

The rise of high-performance Artificial Intelligence (AI), perceived as an ongoing revolution, has led several nations to develop AI strategies to capitalize on its significant benefits. Machine Learning (ML), a key subset of AI, drives this enthusiasm by enabling computers to autonomously improve their behavior and predictive capabilities, resulting in notable efficiency and advancement across various sectors.

Data, the digital representation of information, is essential for developing ML-based systems. These systems process large amounts of data to identify relationships and patterns, allowing algorithms to learn and make predictions or decisions based on new, unseen data. AI performance is directly proportional to the quantity and quality of data, making data availability crucial for AI development.

Generally, data is freely usable and transferable, not subject to ownership rights.[1] The EU has reinforced the importance of open data in the digital economy through several regulations,[2] aiming to make more data available, supporting the growth and innovation of data-driven technologies.

Despite the apparent accessibility of data, significant legal constraints, such as trade secrets, personal data rights, and state secrets, exist to safeguard other socially significant values. One of the most pronounced and litigation-prone restrictions in AI training is the protection of works provided by intellectual property (IP) rights, particularly copyright, which monopolizes the use of original creative works for a limited time to incentivize creativity and originality.

The presence of IP-protected works in AI training datasets introduces considerable legal ambiguity, posing challenges for AI developers in utilizing important publicly available data while risking numerous lawsuits, undermining the advancement of this technology and its social benefits.

Meanwhile, intellectual property owners also face obstacles. Despite holding, in principle, the rights to protect their creations, they often don't have the resources to effectively safeguard their intellectual property rights once their works have been processed into the algorithms along with large amounts of other data, making difficult to prove that their work was used in AI training.

Furthermore, as the accuracy of AI models heavily depends on data availability, copyright law can either enhance AI quality or disrupt it by causing biased decisions. While big tech companies can afford to produce their own data or pay for licenses, smaller AI entrepreneurs, fearing copyright infringement, often resort to less reliable sources such as “biased, low-friction data”, outdated public domain works, and potentially distorted data from Creative Commons (CC) licensed works from Wikipedia.[3] This reliance on "low quality" data jeopardizes the ethical integrity of AI systems, undermines essential social

[1] Property law is a closed system in civil law, which means that the law limits the number of real property rights. Since data is not legislated as an object of property, nor even unanimously qualified as “res”, there is no legal ownership of data.

[2] These include the Regulation (EU) 2018/1807 on the free flow of non-personal data, the Data Governance Act (Regulation (EU) 2022/868) to facilitate data sharing across sectors and EU countries, and the Directive (EU) 2019/1024 on open data and the re-use of public sector information.

[3] Levendowski A. How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem, 93 Wash. L. Rev. 579 (2018). 602 - 619. Available at: https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2

values, and affects the overall quality of AI tools, even for large companies.[4]

2. Current relevant legislation on IP protected data

Worldwide, there are few rules that provide legal certainty about the issues raised by the use of IP-protected works for AI training. It is therefore necessary to rely on the interpretation of established norms and case law in order to work out, on a case-by-case basis, the solution that a given legal system can provide to the matter.

2.1 European Union law

The European Parliament (EP) released a resolution on intellectual property rights for AI development (2020/2015(INI)), a non-binding guide. It recognizes the issues with tracing protected works used in AI, which hinders fair remuneration for authors, and suggests that auditable data records could improve protection for right-holders.

Making the European Union (EU) the world leader in AI technologies is referred to as a goal, requiring an effective intellectual property system suited for the digital age, removing legal barriers, and unlocking AI's potential in the data economy. It stresses the importance of balanced IP rights protection to ensure legal certainty, build trust, and encourage investment, while also protecting human creators and adhering to ethical principles.

Finally, The EP emphasizes that the lawful use of copyrighted works and data in AI must be assessed under existing copyright limitations and exceptions, such as the text and data mining exception in the Directive on copyright in the Digital Single Market.

2.1.1 Copyright and database sui generis protection

Copyright protects the "rights of the author in their literary and artistic work"[5] rather than ownership of the work. In Europe, this protection is automatic, requiring no registration, following the Berne Convention.

Originality is traditionally a condition to the establishment of copyright among continental states, following the French doctrine of‘ Droit d’Auteur’. The EU's Software, Term, and Database Directives describe it as "the author’s own intellectual creation,"[6] a concept extended by the Court of Justice of the European Union (CJEU) to all subject matters in the Infopaq decision.[7] This notion reflects the author's personality, interpreted by the CJEU as the ability to make free and creative choices[8], imprinting the work with a personal touch.[9]

According to CJEU case law, the measure of originality required for the work to be protected can be very modest. In Infopaq I, for instance, the Court of Justice stated that while individual words are not protectable, their combination and selection can be done in a way that express the author’s creativity in an original manner, con-

[4] The inclusion of data derived from additional copyrighted works increases the overall size of the dataset, which can reduce the relative importance of low-quality, free-use data.

[5] Art. 1 of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979).

[6] See respectively article 1/3 of the Software Directive, article 3/1 of the Database Directive and article 6 of the Term of Protection Directive.

[7] Case C-05/08 Infopaq International, ECLI:EU:C:2009:465.

[8] Case C-604/10 Football Dataco, at 39.

[9] Case C-145/10 Painer, ECLI:EU:C:2011:798.

cluding that even eleven consecutive words can potentially express the author's own intellectual creation.[10]

In addition to copyright[11], the EU recognizes in a pioneering way a legal protection of databases, defined as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means" (art. 1º/2 Database Directive), a concept that embody both the protected and non-protected works that constitute the database.

The Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 (Database Directive) established a dual protection regime, a copyright, not for the content of the database, but for the arrangement or selection of the content that "constitute the author's own intellectual creation" (art. 3º) and a sui generis right for the maker of the database that limit the extraction of the database's content. (art. 7º)

The sui generis right for the database maker is a related right of copyright created to protect the investment deployed in the obtaining, verification or presentation of the contents by prohibiting the extraction and reutilization of the whole or of a substantial part of the contents of that database, while extracting and reutilizing insubstantial parts of it that results from normal exploitation of the database is permitted (art. 8º). According to the CJEU jurisprudence, the extraction and reutilization of the database content will be prohibited only when such actions risk depreciating the protected investment, reducing considerably the scope of database content protection.[12]

Regarding that, in Europe, there is no requirement for registration of copyrighted material, the low originality criteria for a production to be considered protected and even the limitation of the use of non-protected work within databases, the possibility of IP-protected work to integrate the data used in AI training is enormous. Consequently, the development of ML models would be constantly under the threat of illegality when if no exceptions apply.

2.1.2 Text and data mining exception

The 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 adopted an exception to the prohibition of unauthorized reproductions and extractions of protected works for the purposes of text and data mining (TDM).

TDM, defined as an "automated analytical technique aimed at analyzing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations," represents most of what AI developers do when training AI systems and could facilitate the use of IP-protected data, but the scope of the exceptions is limited.

This permitted use of protected work was originally created for research purposes. The European legislation, recognizing the importance of the exploitation of all kinds of data to gain knowledge and promote innovation, provided a mandatory exception to the exclusive right of reproduction and to the right to prevent extrac-

[10] ECJ, Case C-5/08 Infopaq International, para. 48.

[11] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society.

[12] Sousa e Silva N. 'Inteligência Artificial e Propriedade Intelectual: Está tudo bem?' I Congresso de Inteligência Artificial e Direito, Edições Almedina (2023), 201-220.

tion from a database "by research organizations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access". (art. 3º/1)

Stakeholders that have different purposes than exclusively research, including commercial, are as well beneficiaries of the exception to encourage innovation also in the private sector. However, there is one extra requirement, the right-holders of the IP-protected work can't have expressly reserve the rights to make reproductions and extractions for text and data mining (art. 4º/3). It represents a presumed license (opt-out) applicable to IP-protected works that have to be expressively denied by the right-holder to prevent or monetize the use of his/her work by TDM.

Despite the directive's aim to promote innovation through lawful data analysis essential for data-driven technologies, the opt-out provision for text and data mining (TDM) has led to a general contractual ban on TDM in the terms and conditions of much publicly available content. This ban is often reinforced by technical measures that prevent crawling and indexing necessary for TDM.[13] Consequently, the TDM exception has been effectively obstructed when right-holders opt-out, making the prohibition of TDM a standard practice in terms and conditions.

2.1.3 EU AI Act

The European regulation on AI (AI Act), a pioneering piece of legislation on AI regulation, is currently in its final stages of implementation. Although this legal document does not affect the enforcement of copyright rules as provided for under Union law, it embodies important statements and rules regarding the use of IP-protected works in AI development.

Following the mentioned resolution of the European Parliament, recital 105 of the AI Act confirms the EP position that the use of copyright, and related rights, protected content requires the authorization of the rightsholder concerned unless relevant copyright exceptions and limitations apply. Article 53/1/c of the regulation goes further regarding the application of Directive (EU) 2019/790 in AI training. It implements the obligation for providers of general-purpose AI models[14] to put in place a policy to identify and comply with the expressed reservations of copyrights and related rights (the opt-out). All the providers should comply with this obligation, regardless of the jurisdiction in which the copyright-relevant acts used in the training of those general-purpose AI models take place (Recital 106).

The AI Act establishes another important provision about the content used to power general-purpose AI models, the obligation for its providers to draw up and make public available "a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office" (Art. 53/1/d). The summary have to take into account the need to protect trade secrets

[13] Ducato R, Strowel A. "Limitations to Text and Data Mining and Consumer Empowerment Making the Case for a Right to "Machine Legibility". CRIDES Working Paper Series, 31 October 2018.

[14]" AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are released on the market". (Article 3/63 of the AI Act).

and confidential business information and be generally comprehensive in its scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law (Recital 107).[15]

Compliance with the obligations applicable to the providers of general-purpose AI models should be proportionate to the type and size of model provider, excluding the need for compliance for persons who develop or use models for non-professional or scientific research purposes, and should allow simplified ways of compliance for SMEs, including start-ups, that should not represent an excessive cost and not discourage the use of such models (Recital 109).

It is important to have in mind that the obligations emerged from the EU AI Act are not restricted to the AI models developed within the European Union's territory. This legislation has a territorial scope extended to all providers that place on the market both AI systems or general-purpose AI models in the Union and if the output produced by the AI system is used in the Union, irrespective of whether those providers are established or located within the Union or in a third country (art. 2/1/a and art. 2/1/c). Such significant extraterritorial effect obliges all the AI developers and providers interested in the expressive European market to comply with the requirements of the AI Act, transforming this activity in a potentially worldwide way.

2.2 United States legislation and case law

Copyright in the United States, unlike the French 'Droit d'Auteur,' aims to promote artistic progress for public intellectual enrichment by allowing authors to benefit from their creative labor. This utilitarian approach is enshrined in the US Constitution, which empowers Congress to secure exclusive rights for authors and inventors for limited times to promote progress in science and useful arts.[16] To guarantee that the established objective of copyright isn't disturbed by its right holders, three judicial doctrines have been established: copyright protects the form of expression, not ideas; facts are not protected by copyright regardless of discovery effort; and the fair use doctrine, which legitimizes secondary creativity.[17]

2.2.1 Fair use doctrine

The fair use doctrine is an exception from copyright formalized by Title 17 of the US Code §107, allowing the use of copyrighted materials without the owner's consent. The main idea is that the copy serves a different function from the original work and doesn't create a substitution, also known as transformative use. In the words of Judge Pierre Leval, who articulated the concept:

"The use must be productive and must employ the quoted matter in a different manner or for a different purpose from the original.... If... the secondary use adds value to the original -if the quoted matter is used as raw material, transformed in the creation of new information, new aesthetics, new insights and understandings- this is the very type of activity that the fair use doctrine intends to protect for the enrichment of society."[18]

[15] The norms of Articles 53/1/c and 53/1/d are also applied to general-purpose AI models under free and open source license. (Recital 104 and Art. 53/2 of the AI Act).

[16] Constitution of the United States. art. I, § 8, cl. 8.

[17] Leval PN. Commentary, Toward a Fair Use Standard,103 HARV. L.REV (1990). 1105, 1111.

[18] Ibidem.

Fair use is a mixed question of law and fact, which means that the finding of whether something constitutes fair use is case-specific considering (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.[19]

In Authors Guild, Inc. v. Google, Inc., in 2015, the court decided that copy a work to extract information not protected by copyright is lawful according to fair use. This understating could cover also Machine Learning uses, where the data extracted from copyrighted works for pattern analysis aren't explicitly covered by copyright rules.

2.2.2 Case Law

The advent of generative AI systems based on Machine Learning promoted a series of lawsuits concerning the alleged use of copyrighted work to train AI systems without the authorization or license of the right holder, the plaintiffs claim that such use is an infringement of the monopoly right of exploring their work.

In Getty Images v. Stability AI, filed in February 2023 in Delaware, Getty Images alleged that Stability AI used over 12 million of its images to train Stable Diffusion, violating Getty's terms of use. The court rejected the defendants ' motion to dismiss in January 2024. Another lawsuit involves visual artists Sarah Andersen, Kelly McKernan, and Karla Ortiz, who filed a class action in January 2023 in California against Stability AI, Midjourney, and DeviantArt, claiming these companies used their copyrighted works to train various AI models, resulting in outputs that are "indistinguishable" from theirs. In October 2024, the court allowed Andersen's claims regarding her registered works to proceed but dismissed other claims. OpenAI also faces a lawsuit from authors Paul Tremblay, Sarah Silverman, Christopher Golden, and Richard Kadrey, who allege that their copyrighted books were used to train ChatGPT. The court dismissed most claims against OpenAI, except for direct copyright infringement, but no merits decision had been taken.

In Thomson Reuters v. ROSS, the issue of fair use in AI training was addressed for the first time. ROSS was accused of using Thomson Reuters 'proprietary information from the Westlaw platform to enhance its AI-powered legal platform, leading to claims of copyright infringement and tortious interference with contract. The court denied ROSS's motions to dismiss and for summary judgment, emphasizing that the plaintiffs 'claims warranted a jury trial. The court highlighted the four factors of fair use under Title 17 of the US Code §107: whether ROSS's AI merely analyzed language patterns or directly replicated copyrighted content, the nature of the copyrighted work and its protection, the extent and necessity of copying for transformation, and the potential market impact and public benefit, all of which required a jury's assessment.

Finally, In December 2023, The New York Times filed a lawsuit against OpenAI and its major financial backer, Microsoft, alleging unauthorized use of millions of its articles to train chatbots. The Times claims this constitutes "free-riding" on its significant investment in journalism and creating a substitute for the newspaper, seeking "billions of dollars in statutory and actual damages." Additionally, the lawsuit demands the deletion of all chatbot models and training data containing copyrighted material from The Times. This case is significant as The Times has a history of defending its journalistic expression through litigation, potentially resulting in substantial monetary penalties under the statutory damages clause of the Copyright Act and the destruction of GPT-based products if The Times wins, it could also establish new fair use prece-

[19] Copyright Law of the United States and Related Laws Contained in Title 17 of the United States Code, pp. 20.

dents, as the defense is based on Section 107 of the Copyright Act.

2.2.3 Proposed bill for the "Generative AI Copyright Disclosure Act of 2024"

Many cases struggle with the lack of evidence regarding the use of copyrighted material for AI training, as AI outputs are influenced by datasets but typically do not reproduce the works entirely, leaving copyright owners to base lawsuits on detected similarities in AI outputs as indirect proof. To address this, Article 53/1/d of the EU AI Act requires AI developers to disclose all training data in a clear summary without compromising trade secrets or confidential commercial information.

In the United States, a similar bill for the "Generative AI Copyright Disclosure Act of 2024," was introduced by Congressman Adam Schiff. This proposed legislation requires a detailed summary of all copyrighted works used in generative AI systems, with a civil penalty of at least $5,000 for non-compliance. Unlike the EU provision, this bill has a retroactive effect, giving companies with existing AI systems 30 days to submit the summary, and new systems must comply 30 days before public release. Supported by numerous entertainment industry organizations and unions, this legislation would enhance transparency in AI development but leaves the determination of fair use applicability to the courts.

2.3 Japanese legislation and data analyses exception

The Japanese legal system has one of the most permissive legislations worldwide regarding the use of copyrighted training data for AI development. An amendment to the Copyright Act of Japan in 2018 introduced Article 30-4, which establishes an exception to copyright protection applicable to AI training. This allows providers to conduct machine learning relatively free of legal issues.

According to Article 30-4, the use of copyrighted material without the permission of the copyright holder is permitted to the necessary extent if the purpose is not for oneself or others to enjoy the thoughts and sentiments expressed in the work. The provision includes examples where the purpose is not human enjoyment, such as "information analysis," listed in item 2. AI training typically falls within this category since it uses the work as data to extract information rather than to create enjoyment from the ideas or feelings expressed in the work.

However, this exception does not apply when the use creates new works that evoke essential characteristics or the creative expression of the original.[20] Additionally, the provision is not applicable if the action unreasonably prejudices the interests of the copyright owner, determined on a case-by-case basis by considering if it conflicts with the market of the copyright holder's works or prejudices potential future markets.[21]

The Japanese Copyright Act does not clarify if using data from a website as training data for algorithms is permissible if the website's Terms of Use prohibit such use. This creates legal uncertainty regarding the acceptance of data use in violation of terms of use or contracts. Another concern is the jurisdiction of Japanese law, particularly in cases where AI developers need to determine the legality of their actions. Generally, copyright infringement is regulated by the laws of the country where the infringement occurred. The location of the server providing the AI model is crucial in determining jurisdiction, potentially affecting the application of Japan's copyright exception when foreign service providers use training data on servers located abroad, even

[20] Fukuoka, Shinnosuke; Murata, Tomonobu; Mizuguchi, Atsuki. Legal Issues in Generative AI under Japanese Law - Copyright. Robotics / Artificial Intelligence Newsletter, 2023

[21] Basic ideas on flexible rights limitation provisions in response to the development of digitization and networking (related to Articles 30-4, 47-4 and 47-5 of the Japanese Copyright Act), Japan Copyright Office.

if the users are in Japan. Conversely, service providers developing AI in Japan with users abroad would presumably be subject to Article 30-4 of Japan's Copyright Act.

3. Possible new solutions

Globally, the issues arising from the impact of AI on IP remain unsettled, leading the World Intellectual Property Organization (WIPO) to release a 2019 document addressing these concerns.[22] Section 13 focuses on copyright issues related to AI training data that may include creative works subject to copyright. The document outlines key issues for discussion to form a shared understanding but does not provide conclusions or recommendations. WIPO's IP global forum aims to clarify existing law interpretations, guide stakeholders, and facilitate international norms. Key inquiries include whether using copyrighted data without authorization for machine learning constitutes infringement, and if explicit exceptions should be made under copyright law.

In addition to the different existing jurisdictions that may present a solution to this emerging issue, different approaches have been supported by experts in recently published doctrine. Among them are the creation of a more permissive TDM exception, the establishment of an online clearinghouse for ML training and the interpretation of the American fair use doctrine taking into account the fair learning principle.[23]

3.1 Broader Text and Data Mining exception

The Joint Comment to WIPO on Copyright and AI, endorsed by 16 members of the Global Expert Network on Copyright User Rights, aims to stimulate discussion on the implications of freedom to use training corpora for commercial or scientific purposes, without presenting an ultimate solution. It distinguishes between two processes involving protected works and text and data mining (TDM) for AI training, questioning if existing law should allow these processes.

The first TDM-relevant activity involves applying computational processes to copyrighted works to derive data, such as conducting internet searches or querying databases like Google Books. The authors argue that although this involves using data derived from copyrighted works without authorization, it often does not constitute a copyright infringement due to the fact/expression dichotomy in law. However, computational processes may require reproducing and storing copyrighted works, raising whether creating a database to be mined necessitates a copyright exception.

R. Ducato and A. Strowel assert that when reproductions are made for search and TDM, the work is not used as a work but merely as a tool to derive information, without public enjoyment of the expressive features. They argue that TDM should not be considered illicit, as it does not meet the 'use of the work as a work' condition for

[22] *Cfr*. WIPO, WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (Al), Draft Issues Paper on Intellectual Property and Artificial Intelligence, Second Session, WIPO Secretariat, available at: https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1.pdf (accessed on 23/04/2024).

[23] Kop M. Machine Learning & EU Data Sharing Practices (March 3, 2020). Stanford - Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust and IPR Developments, Stanford University, Issue No. 1/2020, Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3409712

copyright infringement.[24] The Joint Comment also highlights the potential negative impact on TDM research, machine learning, and AI development if these processes are deemed copyright infringements without an exception. Examples are the equity and ethical issues, such as transparency, accountability and algorithmic discrimination;[25] and the impacts of a globally fragmented legal system to the extent different national laws took different approaches to answering.

The text suggests that WIPO should also evaluate the purpose limitations of research exceptions, especially those limited to 'non-commercial' research,[26] considering their impact on public-private partnerships and socially beneficial commercial TDM products like internet search and language translation. Ducato and Strowel critique the narrow scope of the European TDM exception, emphasizing that TDM should promote research innovation for both commercial and non-commercial purposes, as the boundary between these types of research is often blurred.[27]

3.2 Online Clearinghouse for machine learning training

Given the wide range of works and owners involved in machine learning training sets, licensing each individual piece of copyrighted material is impractical and would likely obstruct, rather than facilitate, the use of such data.[28] The WIPO Conversation on IP and AI explores alternatives for dealing with the unauthorized use of copyrighted data, including the feasibility of a collective rights society similar to a "one-stop shop" with a compulsory licensing system. This system would allow for the commercial and scientific use of data, while ensuring that rightsholders are compensated, thus reconciling the flow of data with the interests of creators who contribute to the development of AI.

However, implementing such a system poses significant challenges. The large volume of works and the diversity of their owners complicate licensing agreements, raising questions of jurisdictional boundaries and the regulatory basis for licensing non-expressive uses that do not compete in the original market. Questions also arise about who should benefit from such a system - authors, publishers or Collective Management Organizations - and concerns about over-licensing, particularly when non-expressive or functional elements of copyrighted works are used for data mining and machine learning purposes. These complexities highlight the need for careful analysis and possibly new legal frameworks to effectively manage licensing in the context of AI development.

3.3 Fair Learning

Obtaining legal protection through fair use of copyrighted works for AI training involves navigating a complex and unpredictable framework defined by four fact-specific factors. Professor Larry Lessig famously characterized fair use as simply the right to hire a lawyer due to its uncertainty. For AI training datasets, several fair use factors often weigh against its application, such as the wholesale copying of entire works without alteration, directly impacting the third statutory factor that assesses the amount of the work used.

Moreover, AI's capability to replicate outputs of creative professionals raises concerns about its competitive implications, potentially influencing how courts view the substitutive nature of a permissive fair use doctrine. The sheer volume of works involved further complicates matters, increasing the risk of litigation from numerous copyright holders, discouraging many AI companies from relying on fair use as a legal defense.

[24] *See* Ducato and Strowel, *supra* note 13.
[25] *See* Levendowski, *supra* note 3.
[26] Article 3/1 of the Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019.
[27] *See* also Ducato and Strowel, *supra* note 13.
[28] *See* Lemley and Casey, *infra* note 29.

In response to these challenges, Mark Lemley and Bryan Casey propose integrating a principle they term "fair learning" into the fair use analysis of AI training data.[29] The principle posits that uses aiming not to obtain or integrate copyrightable elements of a work but to access, learn, and utilize its unprotectable aspects should be deemed presumptively fair under the first fair use factor,[30] which assesses the purpose and character of the use. It suggests that only if such use significantly disrupts the plaintiff's core market should the fourth fair use factor,[31] outweigh a determination of fair learning under the first factor. This approach seeks to provide a structured framework that recognizes the transformative nature of AI applications while carefully balancing the rights of copyright holders.

The fair learning principle acknowledges that not all uses of copyrighted material by ML systems can be considered fair. Some AI applications specifically seek to incorporate the expressive elements of works, which are protected by copyright, into their training sets. This approach poses a risk of significant substitutive competition with the original work, potentially impacting its market. However, fair learning holds that learning from copyrighted material should generally be allowed, similar to the way people learn from cultural pieces for personal enrichment. Most ML systems aim to extract public domain factual or structural information from works, using this knowledge for practical applications rather than for consuming the protected expression itself. Recognizing this distinction as fair learning helps ensure that ML development can proceed without unjustified legal constraints.

The adoption of fair learning as a lawful purpose under the first factor would favor the idea that fair use is not constrained to the use that are transformative or that have no market consequence,[32] but rather applies when they serve valuable social purpose,[33] opening the way to a more pluralistic vision of fair use.

4. Conclusion

Considering both the objectives of the utilitarian American copyright law and the creativity protective droit d'auteur, the use of copyrighted (and neighboring rights protected) materials to collect information should not be considered illegitimate, since the technological process does not aim to use the work as a creative expression, but as a source of quality data necessary for the proper functioning of the machine. Furthermore, its mere use in AI training does not discourage the production of creative content, but instead stimulates it through new tools and exciting potential.

The real legitimate concern for authors of works used in the development of AI models is the possible use of these systems to generate content that is similar to their original work in a way that replaces or limits its market, which would also be considered an infringement of the author's copyright if it were carried out by a human without the use of tools based on AI.

[29] Lemley MA, Casey B. Fair Learning (January 30, 2020). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3528447.

[30] Ideas, facts, functions, methods, and stock literary are not protectable by copyright law.

[31] For example, withdrawing an entire training database directly affects the market, as its value lies in its use for ML, unlike the value of any individual copyrighted work.

[32] The fair use doctrine emphasizes transforming copyrighted works, but machine learning systems typically don't transform the databases they train on, often using them entirely for commercial purposes.

[33] *See* Levendowski, *supra* note 3.

One possible way to balance the legitimate interests involved in using IP-protected works for training AI could be, firstly, implementing a text and data mining exception for any use (both research and commercial), as seen in Japanese law and intended by European law.[34] Secondly, it could involve a policy that ensures transparency for the author, similar to European and American legislative initiatives,[35] while also protecting the creativity inherent in the works used for AI training.

Copyright, due to the central doctrine of "idea-expression dichotomy," does not support prohibiting the use of a creative work in order to remove relevant information that serves to the development of AI. Establishing a general exception for TDM with no opt-outs would provide the legal certainty that this promising technology needs, while also avoiding the risks of bias and monopolization that restricting the use of protected works potentially causes.[36]

Likewise, it is pertinent to protect the legitimate interest of authors by requiring the disclosure of works used in AI training, as it permits audibility and empowers authors to demonstrate when their work is unfairly prejudiced. Additionally, implementing a specific regime to prevent AI outputs from closely resembling original works is essential to protect authors from losing market share. This result can be pursued both by regulating the technology so that it does not allow such plagiarism to take place,[37] and by stipulating an appropriate sanction for users who, despite technological impediments, have used a usurped creative expression to limit or replace the market for the original work used to train the AI.

Bibliography

1. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance (Data Governance Act).
2. Levendowski A. How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem, 93 Wash. L. Rev. 579 (2018). pp. 602 - 619. Available at: https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2.
3. European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies (2020/2015(INI)).
4. Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979).
5. Case C-05/08 Infopaq International, ECLI:EU:C:2009:465.
6. Case C-604/10 Football Dataco, at 39.
7. Case C-145/10 Painer, ECLI:EU:C:2011:798.
8. Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001

[34] The EU's aim was to promote innovation by allowing lawful data analysis, which is essential for the development of data-driven technologies. However, the opt-out approach for TDM has resulted in generalized contractual prohibitions of TDM in the terms and conditions of publicly available content.

[35] Successively, the EU AI Act and the Bill for the Generative AI Copyright Disclosure Act.

[36] *See* Levendowski, *supra* note 3.

[37] This provision could be enforced by another AI-powered system that monitors the works used in the audited AI's training dataset through legally required summaries. This monitoring AI would compare the audited system's results with copyrighted works to detect infringements, though specific criteria for detection must be developed. Additionally, the monitoring AI could define the permissible purposes for using the AI output.

on the harmonization of certain aspects of copyright and related rights in the information society.
9. Sousa e Silva N. 'Inteligência Artificial e Propriedade Intelectual: Está tudo bem?' I Congresso de Inteligência Artificial e Direito, Edições Almedina (2023), pp. 201-220.
10. Ducato R, Strowel A. "Limitations to Text and Data Mining and Consumer Empowerment Making the Case for a Right to "Machine Legibility". CRIDES Working Paper Series, 31 October 2018.
11. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
12. Constitution of the United States.
13. Leval PN. Commentary, Toward a Fair Use Standard,103 HARV. L.REV. 1105, 1111 (1990).
14. Copyright Law of the United States and Related Laws Contained in Title 17 of the United States Code, pp. 20.
15. Authors Guild, Inc. v. Google Inc., No. 13-4829-cv (2d Cir. Oct. 16, 2015).
16. H.R.7913 - To require a notice be submitted to the Register of Copyrights with respect to copyrighted works used in building generative AI systems, and for other purposes. (Bill for Generative AI Copyright Disclosure Act)
17. Fukuoka S, Murata T, Mizuguchi A. Legal Issues in Generative AI under Japanese Law - Copyright. Robotics / Artificial Intelligence Newsletter, 2023 .
18. Basic ideas on flexible rights limitation provisions in response to the development of digitization and networking (related to Articles 30-4, 47-4 and 47-5 of the Japanese Copyright Act), Japan Copyright Office.
19. WIPO, WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (Al), Draft Issues Paper on Intellectual Property and Artificial Intelligence, Second Session, WIPO Secretariat, pp. 4 and 5, available at: https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1.pdf.
20. EUROPEAN PARLIAMENT, Resolution on intellectual property rights for the development of artificial intelligence technologies (2020/2015(INI), 20.10.20, available at: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0277_EN.html.
21. Kublik V. EU/US Copyright Law and Implications on ML Training Data. Valohai, 2024. available at: https://valohai.com/blog/copyright-laws-and-machine-learning/.
22. Kop M. Machine Learning & EU Data Sharing Practices (March 3, 2020). Stanford - Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust and IPR Developments, Stanford University, Issue No. 1/2020, Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3409712
23. https://www.japaneselawtranslation.go.jp/ja/laws/view/4207 (unofficial English translation of the Copyright Act of Japan (Act No. 48 of 1970)).
24. Grimmelmann J. Copyright for Literate Robots. Cornell Law Faculty Publications, 2016
25. Joint comment to WIPO on copyright and Artificial Intelligence. Available at: https://infojustice.org/archives/42009
26. Lemley MA, Casey B. Fair Learning (January 30, 2020). Available at SSRN: https://ssrn.com/abstract=3528447 or https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3528447