ECIJA

Calle Serrano, 69. 28006 Madrid
Tel: +34 917 816 160
www.ecija.com

Training AI with copyrighted material: USA vs Europe. Match point.

October, 2023

Madrid, Spain

Over the recent months, a surge of lawsuits has targeted major AI services such as ChatGPT, DALL-E, Midjourney, and Stable Diffusion, with a significant majority originating from the USA[1].

These lawsuits primarily focus on a pivotal aspect of AI functionality: the alleged use of copyrighted materials during the training phase without permission from the copyright owners.

When training AI systems, the use of public domain content, creative commons, and other open-access resources is just the tip of the iceberg. A significant portion of the training data often comes from copyrighted materials, leading to legal concerns about using such content without the consent of copyright owners.

The question here is do they need permission from the copyright owners to train the AI systems? What does it mean exactly to train an AI system? And how and why do they need copyright protected content?

First, we need to understand what is training in this context: training an AI system, particularly in the context of machine learning and deep learning, refers to the process of feeding the system a vast amount of data and content to help it learn and make predictions or decisions without being explicitly programmed for that task. During this process, the system adjusts its internal parameters to optimize its performance based on the feedback it receives from the training data. Essentially, the AI learns from past examples and refines its understanding to better predict future outcomes[2]. Hence, AI systems require high-quality and diverse datasets to be effectively trained. In numerous situations, AI training often relies on copyrighted materials[3] because they offer high-quality, valuable content that reflects current technological advancements and prevailing social and cultural trends. Consequently, for optimal performance, AI systems may need to utilize, replicate, adapt, and even disseminate such content to deliver the most effective results to users and benefit the broader public. In copyright terminology, AI systems may need to reproduce, communicate and even transform such copyrighted protected content.

AI system developers assert that incorporating copyrighted material can greatly enhance an AI model's accuracy and scope. For example, when training a model to understand human

[1] List of some current cases: https://copyrightalliance.org/current-ai-copyright-cases-part-1/

[22] https://blog.adobe.com/en/publish/2020/02/27/copyrights-in-the-era-of-ai

[33]"The Process of AI Training." Clickworker Customer Blog: https://www.clickworker.com/customer-blog/process-of-ai-training/

language, copyrighted materials such as literary works[4], articles, and other writings are invaluable. Relying solely on public domain or any other non-copyrighted content for AI training might not be adequate given the fast-paced advancements in today's world. This method could yield obsolete and constrained outcomes, particularly when limited to content from authors who died more than 70-80 years ago[5].

The pressing question remains: Is it permissible for them to use copyrighted content without the consent of the copyright holders? The answer is complex. While training AI models on copyrighted data might be considered fair use in some contexts, generating content from these models could pose legal challenges[6].

Central to this debate is whether the use of copyrighted materials for AI training falls under the "fair use" doctrine in the USA. Alternatively, in the European Union, the discussion revolves around any potential limitations on the exclusive rights granted to copyright holders. It's highly probable that these cases will escalate to the highest judicial courts: the U.S. Supreme Court ("**SCOTUS**") and the Court of Justice of the European Union ("**CJEU**").

The outcomes of these cases will undoubtedly influence AI operations, their monetization strategies, the range of materials they can utilize, and consequently, the comprehensiveness and efficacy of these AI systems.

Interestingly, there's a possibility that the SCOTUS and the CJEU might arrive at contrasting verdicts on the matter.

The fair use doctrine of the U.S. copyright statute

Now, let's explore deeper into the situation in the USA, where most of these cases are currently under scrutiny.

The US Copyright Act of 1976 provides in Section 107[7] that notwithstanding the exclusive rights of the owner of the copyright, *the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research,* ***is not an infringement*** *of copyright.*

[4] https://www.reuters.com/technology/more-writers-sue-openai-copyright-infringement-over-ai-training-2023-09-11/

[5] The duration of copyright protection can vary based on the country and its specific regulations. While many countries have adopted a standard of "life of the author plus 70 years," the Berne Convention sets a minimum protection period of 50 years after the author's death for the protection of copyrighted materials.

[6] According to META it should be considered fair use: https://www.reuters.com/legal/litigation/meta-tells-court-ai-software-does-not-violate-author-copyrights-2023-09-19/

[7] **Section 107 17 U.S. Code § 107 - Limitations on exclusive rights: Fair use**: *Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—*
(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.
The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.

Under Section 107, while certain examples like criticism, news reporting, or teaching are highlighted, fair use is not confined to specific and limited unauthorized uses, as is the case in the European Union. Instead, fair use encompasses a broader range of uses. The legality of these uses is determined by the Courts, which evaluate and balance four specific factors to ascertain whether a particular use qualifies as fair use.

These factors are:

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.*

Drawing from SCOTUS precedents on fair use, one could make several arguments in favor of the proposition that training AI systems using copyrighted works may constitute fair use:

1. **Purpose and Character of the Use**: In Campbell v. Acuff-Rose Music, Inc. (1994), the Supreme Court emphasized the transformative nature of a work as a key element in determining fair use. Training AI is arguably a transformative use of copyrighted material. When an AI model like ChatGPT is trained on various texts, it doesn't simply reproduce those texts. Instead, it learns patterns and structures from the data, enabling it to generate entirely new content. The resulting model, though informed by the training data, is a fundamentally new and different creation.

    a. **Commercial vs. Nonprofit**: While OpenAI might have commercial aspects (they charge 20USD per month to premium subscribers), there is also a strong argument that advancements in AI have significant educational, research, and societal benefits. This dual-purpose might help weigh in favor of fair use. Even in Campbell, the commercial nature of 2 Live Crew's parody did not preclude a finding of fair use.

2. **Nature of the Copyrighted Work**: Given the vast and varied nature of the internet, AI models are likely trained on a mixture of factual data, creative works, and everything in between. Relying on Sony Corp. of America v. Universal City Studios, Inc. (1984), it could be argued that not all copyrighted works should be treated equally, and training on a broad swath of diverse content reduces the emphasis on any single copyrighted work's nature. AI systems are using, mixing, and analyzing at the same time protected works, in whole or in part, together with other non-protected content.

3. **Amount and Substantiality of the Portion Used**: Although AI training might involve large datasets, the Supreme Court has indicated that the amount used should be evaluated considering the purpose of the copying. In Campbell, even though the entirety of the song "*Oh, Pretty Woman*" was used, the transformative nature of the use justified it. For AI training, using substantial amounts of data can be necessary to achieve the educational and research goals of developing effective models.

4. **Effect on the Market**: A strong argument is that training AI models doesn't replace or substitute the need for original works. Consuming an AI-generated response or image isn't a substitute for reading a copyrighted book or viewing copyrighted art. In Sony Corp., the Court found that time-shifting did not harm the potential market for the copyrighted TV shows. Similarly, one could argue that AI training doesn't harm the market for original works, as the works themselves aren't being distributed or replicated

in a recognizable form. While AI systems have the capability to distill novels or essays into summaries and emphasize key arguments or ideas, they cannot truly replace or replicate the depth and nuance of the original works. AI systems shouldn't necessarily be viewed as competitors or replacements for the sale of traditional books. Definitely not following Campbell.

Lastly, in the landmark case of Google LLC v. Oracle America Inc. in 2021, the U.S. Supreme Court ruled in favor of Google, determining that its replication of Java API qualified as fair use.

This decision was crucial in the realm of software development and intellectual property. The Court's ruling was heavily influenced by two main factors: the broader public interest and the transformative nature of Google's use of the Java API. These considerations are not just limited to software development but can also be applied to the rapidly evolving field of artificial intelligence (AI). Specifically, when examining AI training processes, one can identify similarities in terms of serving the public interest and the transformative utilization of data. Just as the Court recognized the importance of innovation and adaptability in the tech industry, there's a growing consensus that AI training, when done responsibly, can be seen in a similar light, emphasizing progress and the greater good.

Limitations of the exclusive rights: the European vision.

On the other hand, and in the other side of the Atlantic Ocean, European Union (EU) countries do not follow a "fair use" system like in the USA. Instead, the EU adopts a system of specific exceptions and limitations to copyright. While the U.S. "fair use" doctrine provides a broad and flexible framework that allows courts to determine, on a case-by-case basis, whether a particular use of a copyrighted work is "fair" or not, the EU's approach is more rigid.

In the EU, the exceptions and limitations to copyright are enumerated in Directives and national legislations. These exceptions are specific and cover detailed areas like private copying, quotation, parody, and use for educational purposes, among others. Each exception has its own set of criteria and conditions that must be met.

Using the Spanish Copyright Act as a reference, the Act provides a list of exceptions and limitations to these rights, allowing certain uses of copyrighted works without the need for permission from the copyright owners.

However, upon examining these exceptions, it becomes evident that training AI systems with protected works of authorship does not fit within any of them, making such an act potentially illegal in Spain. Here are some of the most relevant limitations applicable to this analysis:

1. **Temporary Reproductions & Private Copy (Article 31)**: This article allows for temporary reproductions that are part of a technological process, primarily aimed at facilitating network transmissions or lawful uses. It also permits individuals to reproduce disclosed works for private use without commercial intent. However, training an AI system is not a mere temporary reproduction; it involves the permanent ingestion and processing of data. Moreover, the use of copyrighted works to train AI systems often has commercial implications, as these systems are typically developed for profit-making applications.

2. **Security & Official Procedures (Article 31 bis)**: This exception allows for the reproduction, distribution, or public communication of works without the author's consent for public safety or official administrative, judicial, or parliamentary processes. Training AI systems does not fall under public safety or any official procedure, making this exception inapplicable.

3. **Citations, Reviews, and Educational Illustrations (Article 32)**: While this article permits the use of fragments from other works for teaching, analysis, or critique, it does not give carte blanche permission to use entire works or large datasets, which is often required for AI training.

4. **Current Affairs Works (Article 33)**: This exception is specific to works on current events shared via social media. It does not extend to the vast array of copyrighted works that might be used to train an AI system.

5. **Database Usage Rights (Article 34)**: Even though legitimate users of a protected database can access its content without the author's permission, this does not grant them the right to use the database to train an AI system.

6. **Orphan Works (Article 37 bis)**: Orphan works are defined as creations where the rights holders remain unidentified or cannot be located even after thorough and diligent searches. Spanish legislation permits certain public institutions, such as educational centers, museums, and libraries, to reproduce and provide access to these works for non-commercial purposes, especially when aligned with their public interest missions like conservation, restoration, and facilitating cultural and educational access. However, this allowance does not inherently authorize the utilization of these works for AI training purposes.

7. **Parody (Article 39)**: Training an AI system is not a parody of the original work, so this exception is not applicable.

8. **Access to Culture (Article 40)**: This article addresses the posthumous hiding of works and does not relate to AI training.

9. **Three-steps Rule (Article 40 bis)**: This provision underscores the importance of ensuring that exceptions do not adversely affect the author's rights or the customary usage of the cited works. The utilization of AI systems for training purposes poses potential risks to the author's interests. This is particularly true if the content generated by the AI directly competes with, or reduces the significance of, the original work. While this approach may seem optimal in certain contexts, it's crucial to understand that this article is designed to serve as interpretative tools for the enumerated list of exceptions, rather than as standalone provision. Interestingly, a recent decision by the Spanish Supreme Court has taken a unique approach to this provision, as it will be described below. The court determined, applying this article, and not in conjunction with any other provision, that a particular use was permissible as an exception, drawing parallels to the U.S.'s fair use doctrine.

10. **Data Mining**: Spain has adopted the provisions of 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 and amending Directives 96/9/EC and 2001/29/EC*, by providing an exception for text and data mining. Specifically, it allows for reproductions of works and other performances that are legitimately accessible for the purpose of text and data mining.

    However, there are several reasons why this exception does not fully apply to AI training:

    - The exception is designed for the purpose of extracting patterns and knowledge from large amounts of data, not for training AI models that might be used in commercial applications.

- While the reproductions and extractions can be stored for the time necessary for text and data mining purposes, AI training often requires permanent storage and continuous processing of data.

- The exception emphasizes the respect for personal data protection norms and digital rights, which might be compromised in AI training processes.

- The Act allows rights holders to expressly reserve the use of works for mechanical reading or other suitable means, which could exclude AI training.

- The exception is more geared towards research organizations and cultural heritage institutions for scientific research purposes. Commercial AI training does not fit within this context.

- Even if a work is legitimately accessible, the exception does not permit reproductions and extractions for AI training that goes beyond text and data mining as described in the Act.

In conclusion, the Spanish Copyright Act's exceptions and limitations are not designed to accommodate the unique challenges and implications of AI training. Or, in other words, AI training does not fall into any of the numbered cases and types of uses that limit the exclusive rights of the copyright owners.

However, it's worth noting that while the Spanish and EU system is more prescriptive, some member states, as mentioned, have implemented broader exceptions that resemble "fair use" to some extent. However, these are not as open-ended as the U.S. system.

For instance, these might the case of the Spanish Supreme Court in the Google case of 2012 (STS 3942/2012) which is often cited as an instance where the Spanish Supreme Court has seemed to adopt a stance resembling the "fair use" approach.

In this case, the Spanish Supreme Court had to determine whether Google's practice of displaying snippets of the web's content in its search results constituted a copyright infringement.

The court ruled in favor of Google, stating that such use was a transformative one and did not compete with the original use of the copyrighted news articles. The court emphasized the importance of search engines in modern society and recognized the value they bring in terms of disseminating information.

Several factors made this decision resemble a "fair use" approach:

a) **Transformative Use**: The court considered whether the use was transformative, a key factor in U.S. fair use analysis. Google's use was deemed transformative because it provided a new function - helping users find information - rather than replicating the original purpose of the website.

b) **Nature of the Use**: The court acknowledged the importance of search engines in the digital age and the societal benefit they provide by facilitating access to information and content.

c) **Effect on the Market**: The court assessed the impact of Google's use on the potential market for or value of the copyrighted work. It concluded that Google's snippets did not harm the potential market for the original content of the website.

d) **Amount and Substantiality**: Even though Google used only small snippets, the court considered the qualitative value of the snippets in relation to the whole content.

While the ruling did not overtly embrace the "fair use" doctrine, although it was repeatedly cited in its Judgement, the criteria evaluated, and the justification given by the Spanish Supreme Court bore a striking resemblance to the principles underpinning the U.S. fair use framework. This interpretation has prompted many legal scholars and practitioners to speculate that the Spanish judiciary may, under specific circumstances, be inclined to adopt a more adaptable stance towards copyright exceptions, drawing parallels with the "fair use" doctrine prevalent in the U.S. Notably, the Supreme Court's groundbreaking move to singularly apply Article 40 bis (Three-step rule) without pairing it with any of the enumerated exceptions in the Copyright Act was seen as a significant departure from traditional interpretations.

The Court underscored the importance of striking a balance within the Copyright Act. It cautioned against allowing the scope of copyright to extend excessively or in ways that might seem unreasonable. Such overreach, while causing only minimal or no harm to the rights holder, could disproportionately obstruct services that offer substantial advantages to the broader community. It is essential to ensure that the Copyright Act serves its primary purpose of protecting creators without stifling innovation or impeding societal progress.

The principle of protecting intellectual property is deeply rooted in the American legal framework. Specifically, Article 1, Section 8, Clause 8[8] of the U.S. Constitution articulates this commitment by granting Congress the power to protect the rights of inventors and authors. This clause ensures that inventors receive protection for their inventions through patent law, while authors are safeguarded for their writings under copyright law. The underlying rationale for these protections is to "*promote the progress of science and useful arts*." By offering these protections, the Constitution aims to incentivize innovation and creativity, recognizing their pivotal role in advancing society and contributing to the nation's cultural and technological growth.

Preliminary conclusions

In summary, while there are similarities in the objectives of promoting creativity, innovation, and access to knowledge, the mechanisms by which the U.S. and EU achieve these objectives in copyright law differ significantly.

Using copyrighted materials to train AI systems without explicit consent it might be potentially considered "fair use" in the United States. However, this practice might contravene copyright laws in the European Union. As the realm of AI rapidly advances and finds broader applications, there's an emerging necessity for legal structures to evolve, offering more precise directives on such matters.

This divergence in legal perspectives could pose operational challenges for U.S.-based AI systems operating within the European Union. Ultimately, they might be compelled to seek formal authorizations and potentially remunerate copyright holders. This could set a precedent, prompting them to adopt similar practices with American copyright holders.

An analogy can be drawn with data protection norms. For practical reasons, many U.S. tech giants have adopted the European Union's General Data Protection Regulation (GDPR) as a

[8] *The Congress shall have power:*
../...
To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries;

global standard for their services. However, it's plausible that, before universally adopting such standards, AI companies might contest these regulations. They could await a definitive ruling from the Court of Justice of the European Union, especially if the U.S. Supreme Court has previously ruled that their use of copyrighted materials for AI training falls under "fair use."

A further challenge, warranting its own article and research, is determining how AI providers measure the extent of their usage of specific copyrighted content and how they establish the corresponding compensation to content owners. While in the music industry it's somewhat straightforward to measure how frequently a song is played on radio stations, TV, or streaming platforms, and subsequently calculate royalties, is it as simple to determine the extent of copyrighted material usage when training an AI model?

Carlos Rivadulla[9]
Lawyer. Manager of the TMT team at ECIJA.
crivadulla@ecija.com

[9] The content, arguments, and structure are the original creations of the author. ChatGPT has aided in making grammatical improvements and suggesting alternative vocabulary.