SERVING AUTHORS WORLDWIDE
AU SERVICE DES AUTEURS DANS LE MONDE
AL SERVICIO DE LOS AUTORES EN EL MUNDO

SG23-0049
Source language: English / Written on: 20/10/2023
Document prepared by CISAC Secretariat (CISAC)

CISAC Comment on the US Copyright Office's Notice of Inquiry (NOI) on Copyright and Artificial Intelligence

The International Confederation of Societies of Authors and Composers (CISAC) welcomes the opportunity to engage with the U.S. Copyright Office ("USCO", the "Office") in its request for comment on copyright law and policy issues raised by artificial intelligence ("AI") systems.

CISAC is the leading worldwide organisation of authors' societies. We represent more than 5 million creators from all geographic areas and all artistic repertoires (including music, audiovisual, drama, literature, and visual arts) through our 225 members. The position of CISAC is not just a reflection of its members, but of its long history centred on defending the livelihood of creators and supporting creativity for future generations.

This comment focuses on legal and policy issues pertaining to the use of AI, specifically generative AI tools and their applications by end-users, with particular regard to the following:

1) the use of copyright-protected works as training data (input) for developing AI models; and
2) the recognition of copyright in the context of AI-generated or AI-assisted outputs.

Through this submission, we hope to provide a clearer picture of emerging legal issues as viewed through the lens of foundational principles of copyright law, as well as identify several opportunities for improving the current state-of-play.

Table of Contents

I. Introduction

Society benefits from the presence of creative works, which is made possible through the application of a balanced system of copyright law. To achieve this balance, copyright laws must enable creators to make a livelihood from their works, while allowing the public to access and consume such works within the creative marketplace.

In current discussions surrounding the development of AI[1], it is evident that large volumes of creative works are required in order to develop and improve the effectiveness and quality of an AI application designed to generate content ("generative AI"). Volumes of data containing information related to creative works, including copyright protected works, is typically gathered using so-called "data scraping" and "web scraping" tools and methods. These methods often include the use of automated webcrawling processes to gather information, sometimes applied indiscriminately and without seeking permission to use the content gathered for training purposes.

Thus far, the process of obtaining and using such a volume of creative works has largely gone unchecked: this is due to the lack of transparency from AI companies concerning the datasets and processes used to train their generative AI models, the lack of explicit guidance on the application of existing copyright law to generative AI training, and AI companies' overreliance on the blanket assumption that such a use is "fair" or otherwise subject to an exception. **As such, the current process actively disrupts the essential balance of copyright law by creating a strong bias in favour of AI technologies. This bias threatens the livelihoods of members of the creative community by, for example, possibly replacing the need for works that are the result of human creativity.**

Furthermore, outputs generated through the use of AI tools should be carefully scrutinised before considered protectable under copyright law. Though AI has important facilitative potential in creative processes, it should only be viewed as a tool in the hands of a human creator. Likewise, the products of human creativity should be recognised distinctly and apart from elements generated using AI tools. Without appropriate licensing, outputs generated through the use of AI tools could also potentially infringe on the creative works used to train such tools. **Though there is some considerable grey area to consider in this task, we highlight below several conceptual guideposts and practical measures which may facilitate more legally accurate decision-making with regards to determining the protectability of works incorporating elements of generative AI.**

II. The Use of Copyrighted Works to Train AI Models

Foundational models used for generative AI applications and tools require the ingestion of large datasets in order to train the model effectively. This necessity for amassing large amounts of data creates incentives for AI developers to gather as much data as possible, and such practices become problematic when performed on a wide-scale and on an indiscriminate basis. Artists and creators using the online space to promote and disseminate their works often remain unaware that their works have been digested by a foundational model. In section II. A., we argue that exempting AI developers from performing basic due diligence on the taking and use of copyrighted works for training purposes is detrimental to the welfare of creators, who are prevented from authorizing such uses, and asserting their rights to remuneration.

Regulatory solutions may further be considered in order to balance the interests at stake. As discussed below, there seem to be several means available for addressing the issues: first, by making voluntary licensing options available for creators whose works are used for training purposes (II., B.); and by creating clear

[1] "AI" and "AI technologies" are used here as encompassing terms, including reference to foundational models (trained through various machine learning techniques), which are then adapted for downstream "AI applications" (i.e., general purpose AI systems or AI systems designed for specific purposes).

reporting obligations on AI developers to make their training data sets available in a way that puts rightsholders on notice that their works are being used for training purposes (II. C.).

A. Legal Significance of the Use of Copyrighted Works as Training Data

1. Substantial Use of Copyrighted Source Material

As a preliminary issue, much current legal debate concerns the use of copyrighted works as training data, and whether or not such a use qualifies as a violation and/or infringement of an existing copyright.

Most significantly, as already observed within the visual arts, audiovisual and music communities, generative AI models are capable of taking and reproducing elements of a copyright protected work which are the "heart"[2] or the "essence" of the source material. While the "style" or "feel" of a creators' repertoire of works may not be copyrightable itself, unlike a human who draws creative inspiration from such source material, AI replicates the style and feel of creators' works by using copies of their works. The automated process implemented by the generative AI model may feature some element of random choice, but despite this randomness (and in order to stay accurate), AI developers are rather incentivised to design their systems to be able to reproduce as closely as possible source materials which are requested via prompt by the user.

This replication of elements of copyrighted works by generative AI models would therefore clearly fall above the threshold of being a mere "token" use of existing works, or a use as purely raw data. Any argument sustaining the impression that uses of copyrighted content for training data purposes are only minimal and insubstantial overgeneralises the reality of the situation, and understates the danger these practices pose to the livelihood of creators relying on the same user demand for creating and selling their own works.

It is further worth noting that the recognisability and identifiability of copyrighted elements of source works, when reproduced in the outputs of generative AI, should not be the dispositive measure of whether the use of the work for training merits a license for use and/or remuneration. Nor should it be the case that AI developers should be tasked with extracting only elements of a work which would not have, independently of the work, been considered a protectable element under copyright**. Instead, we suggest that, as a default, permission from the author for the use of a work for the purposes of use as training data should be requested in all instances.** This default rule would effectively recognise the inherent commercial value in the use of a work as training data, while providing a means for offsetting the loss in revenues caused by the presence of competing works trained using copyrighted source material.

2. Inapplicability of Exceptions or Limitations to Use of Copyrighted Works for Training AI Models

In order for copyright's essential balance between private rights and public interests to be struck effectively, copyright law enables certain exceptions, allowing the public to make specific uses of copyrighted works under special circumstances, whereas the application of "external" safeguards such as the recognition of the balancing of fundamental rights, help to establish the outermost boundaries of the recognition of such exceptions. Necessarily, copyright exceptions are recognised and applied in circumstances that do not cause undue prejudice towards a creator's ability to benefit from the uses of his or her work.

AI developers so far have relied on exceptions to copyright law to freely gather and use data derived from the internet to train their models, in particular the doctrine of fair use. Following the arguments made above, with emphasis on the element of unreasonable prejudice to the legitimate interests of creators at stake, **we**

[2] Under the fair use doctrine, the Supreme Court has recognised the importance of considering the characteristics of the content taken in determining the substantiality of the use of copied content [third factor]. (See, e.g., Acuff-Rose at 587: "[The Court of Appeals held that] '[w]hile it may not be inappropriate to find that no more was taken than necessary, the copying was qualitatively substantial. . .. We conclude that taking the heart of the original and making it the heart of a new work was to purloin a substantial portion of the essence of the original.' . . .The Court of Appeals is of course correct that this factor calls for thought not only about the quantity of the materials used, but about their quality and importance, too.") Of course, the third factor of the fair use test should be assessed in the context of the other four factors. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994).

maintain that AI developers should not benefit from the application of an exception to copyright with respect to the use of copyrighted works as training data (i.e., a "fair use"), because extending such exceptions in this instance violates the "three-step test" as recognized by international treaties, to which the U.S. is bound.

The introduction of additional exceptions, including those providing for text and data mining (TDM) that permit AI systems to exploit copyright works and performances without authorisation or remuneration, must therefore also be avoided.

B. Licensing of Works Used for Training Purposes

It is essential for creators to retain the option to license their works for the purposes of training AI models.

The use of copyright-protected works by AI should be considered a use subject to copyright authorisation, namely because it involves at least the rights of reproduction, preparation of derivative works, and distribution.[3] In recognising this, fair licensing practices should be encouraged between creators and AI innovators interested in using copyrighted works as training data, and such licenses may be most effective when applied under standardised rules. While text and data mining licensing for AI purposes is still a nascent market, we expect that the rightsholders would achieve market-value compensation for the use of their works in this context.

Additionally, as developed below, no exception to obtaining a license and/or obtaining permission for use should apply. This means that current interpretations of existing exceptions may need to be revisited to ensure that the development of fair licensing practices will not be hindered.

C. Transparency and Disclosure Requirements of Training Data

AI developers have claimed that, during the course of data gathering, it is not feasible to identify all works subject to copyright law, nor is it possible to identify the creators from whom approval should be sought. **We, however, maintain that indiscriminate data gathering practices should not undermine the recognition of rights in a creative work.** In support of these rights, several practical steps can be readily applied by AI developers in order to create a system that encourages fair treatment of creators. One example currently under discussions in the EU relates to its draft regulation on AI ("EU AI Act").[4] The proposed provision mandates that AI developers disclose the relevant information about the copyrighted works used for training purposes.[5]

When constructing practices which prioritise responsible reporting of datasets used by generative AI, developers – including those who focus on end-user tools and creators of third-party models – should further acknowledge and disclose when third-party inputs have been used in training, as well as outputs and data

[3] Statement of Society of Audiovisual Authors, "Artificial intelligence must serve society and enhance human creativity" 4 October 2023. https://www.saa-authors.eu/file/1287/download.

[4] See "Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))" (hereinafter "Proposed Amendments to EU AI Act Draft Text") https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

[5] Ibid., Art. 28 (b)(4), Obligations of the provider of a foundation model:
"Providers of foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video ("generative AI") and providers who specialise a foundation model into a generative AI system, shall in addition . . .

a) comply with the transparency obligations outlined in Article 52 (1);
b) train, and where applicable, design and develop the foundation model in such a way as to ensure adequate safeguards against the generation of content in breach of Union law in line with the generally-acknowledged state of the art, and without prejudice to fundamental rights, including the freedom of expression,
c) without prejudice to Union or national or Union legislation on copyright, document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law.

provided by users which are used to feed back into the data pool. The AI developers are best positioned to access such information and develop the means to monitor such activities.

Overall, it is important to recognise that transparency obligations in the disclosure of data sets are a means to an end – if improperly applied, these obligations would only provide a cursory overview of the data being used for training of AI models, thus failing to achieve its true objective: to enable rightsholders to better understand when their works are being used, and where a license should be applied.

To successfully achieve this end, it is essential for national governments to facilitate cooperation between copyright rightsholders and AI developers to determine whether rights have been violated. In such cases, fair, voluntary licenses can be made available by rightsholders to remedy the situation, and AI developers should make all effort to comply with applicable reporting/transparency obligations.

As for the future of developing better and more accurate technological standards to aid in the proper identification of works and authors, there is ample precedent for the development of such technologies. For years, CISAC, rightsholder groups, and many other organisations managing rights on behalf of creators internationally, have channelled their resources into developing and improving technical measures and infrastructures for enabling creators to receive their fair share from their creative efforts. This has included the creation of reliable systems such as industry-standard ISWC and ISRC identifiers, among other similar tools, which ultimately streamline the task of determining copyright authorship in works.

We believe that further collaborative efforts between AI developers and rightsholders are essential to improve the infrastructure surrounding works identification, and that such collaborations should be encouraged and facilitated at the policy level.

III. Legal Status of Outputs of Generative AI (Protectability)

A. User-Assisted/User-Prompted Outputs of Generative AI

As repeated in numerous recent documents released by the US Copyright Office,[6] in cases where a work contains both AI-generated material and human authorship, the Office may conclude that the components of the work (whether AI generated or not) over which the human author exhibits sufficient creative control are entitled to copyright. In practice, this has led to the rejection of a number of applications of registration of a work due to, e.g., the presence of generative AI elements where a sufficient level of human creative control could not be exhibited.

In other words, when a human has selected or arranged AI-generated material in a sufficiently creative way, such works may merit copyright protection. In those cases, and as currently maintained by the USCO, copyright should protect the sufficiently human-authored aspects of the work, while disclaiming copyright protection over elements created or added primarily or exclusively through the use of AI. **We believe that clear rules are required in this domain in order to ensure that AI tools are available to creators, and that the creators using such tools – in order to further their own human creative vision and practice – are not unduly limited because of their use of AI tools to aid their human efforts.**

B. "Fully autonomous" Creations by AI

As put by Dan Burk, the prospect of fully autonomous AI in creating works is an error in framing: "AIs seem autonomous in isolation just as a paintbrush or pencil might seem magically and mistakenly autonomous if

[6] E.g., U.S. Copyright Office. "Response to May 1, 2023 Letter re: AI Material" copyright.gov. https://www.copyright.gov/laws/hearings/Response-to-May-1-2023-letter-re-AI-Material.pdf; U.S. Copyright Office. "Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence" 16190 Fed. R., Vol. 88, No. 51, Rules and Regulations. March 16, 2023. copyright.gov. https://copyright.gov/ai/ai_policy_guidance.pdf.

we were to ignore the human hand that holds it."[7] **We also maintain that, absent a sufficient level of human creativity, outputs fully created by AI should not be entitled to protection under copyright law.**

Given the current state of technology, it is still the case that AI cannot independently produce fully original works; the recognition of copyright in works created fully autonomously by AI without any human intervention still only exists as a legal fiction. Nevertheless, in the future, it may be necessary to acknowledge the presence of such works on the creative marketplace in some way. **We suggest that, as the technology advances, the legal protections surrounding works created fully autonomously through AI should be further clarified.**

C. Resolving Practical Issues Concerning the Registration of Works Containing Elements of AI

A number of recent cases dealt with by the USCO have involved circumstances where an initial registration was granted for works incorporating generative AI technologies, but was later revoked by the Office due to a lack of appropriate initial disclosure of the use of AI.

This experience suggests that a lack of clarity regarding the protection over works featuring elements created with the assistance of AI will signal a new wave of application disputes to be handled by the USCO and other national copyright offices. **As such, it is in the best interest of all parties to simplify and streamline the recognition of works to the greatest extent possible, ensuring that creators can easily understand the threshold of protection when benefitting from the use of AI tools in their creative practice.**

Likewise, the Office's early idea of disclaiming portions of a whole work featuring generative AI elements may not be the most feasible solution in the short- or long-term, due to the added administrative burden created by dividing every single work into protected and non-protected pieces. Nevertheless, **we agree that, at the very least, new protocols should be established by the Office and other national offices which require proper disclosure of information on the nature of the creators' use of AI in their creative workflow.**

D. Addressing Market Displacement Caused by Generative AI Outputs

As the overall use and acceptance of AI technologies in our daily lives continues to grow, it is anticipated that the volume of outputs by generative AI will also dramatically increase. The disruptive effect is likely to be felt both in the number of new outputs that will compete on the creative marketplace against the creations of human authors, and in the changes to the production costs (and expected returns) of creative works overall.[8]

It should therefore become a priority, at the earliest stage possible, to create protocols which can assist consumers in discerning the products of human creativity within a marketplace saturated by AI outputs. This can potentially be accomplished through the development of labelling or tagging systems. In the EU, the proposed EU AI Act[9] acknowledges this issue by requiring that "[g]enerative foundation models should ensure transparency about the fact the content is generated by an AI system, not by humans".[10] AI developers, who are best positioned to identify works produced with the use of AI tools, should bear the obligation to appropriately label or tag AI outputs.

Finally, accurate registration information on works created using AI (to any degree) will also gain importance as the number of AI-assisted works grows in the creative marketplace.

[7] Burk, D., "Thirty-Six Views of Copyright Authorship, by Jackson Pollock" 58 Hous. L. Rev. 263, 266 (2020).
[8] "AI threatens to lower the costs of creation itself, so that creative works—however easy to access—are cheap to produce in the first place." Burk, D., "Cheap Creativity and What It Will Do" 57 Ga. L. Rev. 1669, 1680 (2023).
[9] "Proposed Amendments to EU AI Act Draft Text", n.3.
[10] Ibid., Amendment 101 Recital 60 g (new).

IV. Conclusion

Internationally, there is a pressing need to understand and anticipate the many impacts AI innovations will have in our day-to-day lives. Governments can ensure stability in the short- and long-term through establishing clear obligations for AI developers which can, in turn, create a culture centred on responsible innovation. However revolutionary, it is imperative that AI technologies be developed and utilised in a controlled manner, especially to avoid damaging the livelihoods of creators.

Copyright law, if respected to the level indicated in this submission, may be the most useful tool in the toolbox for regulators to ensure that the future of creativity is secured. Copyright has always existed as a body of law designed at the service of human creators. As such, copyright laws can serve as an important shield for creators against the initial consequences of AI's influence on the creative marketplace.

Likewise, copyright exceptions do not exist to undermine the overall functioning of the copyright system and the interests of rightsholders, but to restore the equilibrium between the interests of rightsholders and the interests of the public. Taking this principle into account, even in the context of open norm systems such as fair use, rights may not be stripped from rightsholders simply for the sake of innovation. AI can provide many opportunities for improving society, but such benefits should not be won at the price of human creativity.