DEPARTMENT OF COMMERCE
Patent and Trademark Office

Docket No. PTO–C–2019–0038

Request for Comments on
Intellectual Property Protection for Artificial Intelligence Innovation
(84 Fed. Reg. 58141, October 30, 2019)

WRITTEN COMMENTS OF
COPYRIGHT CLEARANCE CENTER, INC.

January 10, 2020

INTRODUCTION AND BACKGROUND

Copyright Clearance Center ("CCC") is providing these comments to the United States Patent and Trademark Office ("USPTO") concerning the impact of artificial intelligence ("AI") technologies on intellectual property law and policy, particularly with respect to copyright matters, in response to the USPTO's questions published in the Federal Register on October 30, 2019.

CCC has served for more than 40 years as a licensing hub primarily for text-based copyrighted materials, enabling the issuance of licenses on behalf of tens of thousands of rightsholders to users of all kinds, including academic, business, government and non-profit organizations. We offer our services in a large variety of ways: transactional (across many different uses) or repertory (sometimes called blanket) licenses, centralized at our own website or decentralized at the websites of participating rightsholders, domestic-only or multinational, ourselves or in partnership with either rightsholders themselves or with peer organizations in other countries. We enjoy longstanding and close relationships with publishers, authors, and their representative associations and other groups, both in the United States and elsewhere. And, almost uniquely among text-based collective licensing organizations, CCC offers its services on an entirely voluntary basis – we represent the rights and works solely of those rightsholders who, directly or through their agents, sign agreements with us and we issue licenses solely to users who choose to buy them.

In support of those many licensing services, CCC developed – first for its own use and ultimately for the use of its customers – a variety of patented inventions and technologies that make copyright licensing and management more convenient and efficient. As a result, CCC is deeply

involved with issues of technological innovation concerning rights and content.[1] As two examples most closely connected to the issues raised by the Request for Comments, (i) CCC offers a license on behalf of almost 60 publishers along with access rights to aggregated and normalized scientific articles for text mining (which includes the right to extract information and data so that employees may engage in directed/assisted machine learning for internal AI development and the internal use of the licensee), and (ii) CCC also provides metadata, tagging and enrichment services to publishers and users designed to, among other things, enable our clients to better use machine learning and AI technologies.

The term "Artificial Intelligence" or "AI" covers a broad range of technologies, and there is no broad, commonly accepted definition. For our purposes, we propose the following working definition: "AI systems facilitate the automation of tasks, normally performed by humans, by incorporating information from the data that they process in order to adjust the outcome of the task." In the past decade major advances have been made in a subfield of AI that is called machine learning and, in particular, a subfield of machine learning called deep learning, which is a class of algorithms in which data is processed through layers from raw input to greater and greater levels of abstraction, at each layer providing a better representation of reality (and thereby enabling the machine to perform, and to teach itself to perform, more and more sophisticated tasks).[2] Some of these successful applications of machine learning and deep learning relate to tasks that rely on the processing of copyrighted materials, such as photographs, audio recordings, videos, books, journal articles, and other digital assets. In data terms, these types of content are generally regarded as "non-structured" and therefore more difficult to analyze (as compared to "structured" data, usually in the form of tables and graphs of numbers, which are more readily analyzed by traditional software), and the higher quality the unstructured data are, the higher quality the ultimate uses of them (for example, through deep learning) are likely to be.

All applications in this field start with a purpose or objective which can be commercial, non-commercial or scholarly. A party then builds a model algorithm. Where "training" is involved, the algorithm is trained by using appropriate data sets, and then the party applies the trained algorithm to new data and content to generate certain outputs (a "data first" approach).[3] Current

---

[1] Recent awards for technical innovation that CCC has been awarded include a Readers' Choice Award from KMWorld for its content discovery tool. https://www.kmworld.com/Articles/Editorial/Features/2019-Readers-Choice-Award---Best-E-Discovery-Copyright-Clearance-Center-134855.aspx. See also https://www.kmworld.com/Articles/Editorial/Features/KMWorld-Trend-Setting-Products-of-2019-132295.aspx. In 2019, EContent's readership selected CCC for inclusion in the magazine's annual list of "Companies That Matter Most in the Digital Content Industry." http://www.econtentmag.com/Previous_EContent100_Winners

[2] However, not all machine learning algorithms are trained in this manner.

[3] See Drexl, J., et al., *Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective* (Max Planck Institute for Innovation and Competition Research Paper No. 19-13, October 22, 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3465577; Scollo

consumer-oriented applications of this type of machine learning model include tailored search capabilities in search engines such as Google or Bing, voice recognition software such as that embedded in Alexa by Amazon, and image recognition software such as Amazon Rekognition. Researchers are also using machine learning to quickly identify key research data in order to identify common patterns and processes among diverse data, with goals such as improving healthcare outcomes and developing new health practices and pharma products. Finally, businesses are using machine learning, and AI generally, for internal operational efficiencies as well.

As noted above, many AI practices involve the ingestion of copyrighted content, including content from journals, newspapers, books and databases that are part of CCC's repertories available for licensing. The result of significant ideas and research, thoughtful analysis of facts and theories, and conscientious and (hopefully) clear writing skills, this kind of copyrighted content has driven scientific, political, economic and business decision-making for hundreds of years. And it is the qualities of this type of content that make it most desirable for training and as datasets in various forms of AI applications, just as it has been used for the training of *humans* since (at a minimum) the advent of writing and has formed the "datasets" (usually called research materials) for those humans. This point about quality is widely recognized: for example, a September 2019 WIPO conversation on AI and intellectual property[4] reported that "[a] common misunderstanding is about the quantity of data needed for machine learning when in reality the **quality** of data is really the key" (emphasis added). In fact, quality data inputs, including inputs of copyrighted content, are now one of the most valuable tools for businesses and other organizations to operate successfully and efficiently.[5]

Recognizing the value of high quality data in the form of copyrighted content and the need for business models to ensure its future creation, CCC is of the view that, when such content is used for commercial AI projects – whether they are for training, testing and de-biasing, application in real life, verification, transparency or other purposes – licensing the use of the content is generally the appropriate model. This is especially true where licensing is practical, the content is professional, and the user is specifically choosing that content for its unique value. Licensing is also appropriate for non-commercial projects to the extent that the reuse would otherwise

---

Lavizzari, C., *Artificial Intelligence: How Machines Learn and What It Means for Authors, Publishers and Media Businesses* (presentation at Fordham International IP Conference, April 2019), http://fordhamipinstitute.com/wp-content/uploads/2019/04/Lavizzari-Carlo-Scollo-AI-IP-Presentation-2.pdf.

[4] Several dozen high-level experts were convened by WIPO in September 2019 for a set of structured discussions of AI and intellectual property, and the WIPO Secretariat then summarized the outcomes of the discussions. See in particular https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_ge_19/wipo_ip_ai_ge_19_inf_4.pdf at ¶ 88.

[5] Marmanis, B., The Concept and Importance of Knowledge Supply Chains https://www.copyright.com/blog/the-concept-and-importance-of-knowledge-supply-chains/, August 21, 2019; Reed, J., Using AI-powered Text Mining to Re-use Research Insights Published in Scientific Literature, https://www.copyright.com/blog/ai-powered-text-mining-research-insights-scientific-literature/, November 12, 2019.

prejudice the reasonable commercial interests of a rightsholder who ordinarily sells or licenses her works into that non-commercial market (such as textbooks to educational institutions). In our own business, both rightsholders and users seem to agree that a marketplace for licenses to use such high-quality content is reasonable and suitable: CCC's current license for text and data mining – itself an application of AI – including access to convenient forms of the content, today covers more than 8,000 journals and 11 million articles from 59 scientific-journal publishers.

RESPONSES TO QUESTIONS

Against this background, CCC offers the following responses to the specific questions contained in the USPTO's Request for Comments that are relevant to our views:

- **Question 1:** Should a work produced by an AI algorithm or process, without the involvement of a natural person contributing expression to the resulting work, qualify as a work of authorship protectable under U.S. copyright law? Why or why not?

- **Question 2:** Assuming involvement by a natural person is or should be required, what kind of involvement would or should be sufficient so that the work qualifies for copyright protection? For example, should it be sufficient if a person (i) designed the AI algorithm or process that created the work; (ii) contributed to the design of the algorithm or process; (iii) chose data used by the algorithm for training or otherwise; (iv) caused the AI algorithm or process to be used to yield the work; or (v) engaged in some specific combination of the foregoing activities? Are there other contributions a person could make in a potentially copyrightable AI-generated work in order to be considered an "author"?

Response to Questions 1 and 2: With respect to the question of authorship and the involvement of natural persons in AI technologies, CCC notes that there are already a number of types of works – such as databases, software, and collective works of all kinds – where natural persons interact with systems and technologies to create works protectable by copyright, including word processing, computer aided design, and music and motion picture editing software. While deep learning algorithms offer new possibilities in the creation of new copyrighted material, for the foreseeable future natural persons will still be heavily involved in designing the models and algorithms, identifying useful training data and standards, determining how the technologies will be used in commerce and research, guiding or overriding choices made by the algorithms, and selecting which outputs are useful or desirable. Likewise, for the foreseeable future natural persons will author AI algorithms, just as natural persons now author software. We see no reason for new rules simply because what is now an old technology – software – does something new; just as the old rules ("who is an author?" has evolved for 230 years and the work for hire doctrine for at least 100) have adjusted as society and technology have evolved, there is no reason to believe that they cannot continue to do so.

There are a number of existing scenarios for software, audiovisual works, compilations and other collective works where co-authorship and collective contributions are identified among a number of different types of creators. However, the general view of most courts is that, to be considered

an author or co-author, the contributor must have general oversight of the work, have a general understanding of how different elements will be combined to create the totality of that work, and provide actual intellectual or creative input. CCC suggests in response to the USPTO's **Question 2** that a contributor must be involved in the design, refinement, and ultimate outcomes of an AI work in order to be deemed an author or co-author, and that contributors who merely organize or provide data or code elements at or under the direction of another contributor should not be viewed as co-authors. The involvement and recognition of contributors other than co-authors would generally be resolved as a matter of contract, or under rules and regulations of the employing entities involved, and similarly the rights of contributors to collective works or joint works will be governed by contract or by the Copyright Act itself.[6] Further, we note that, in a typical licensing model author recognition (if any) is specified by contract, whether negotiated or in the form of a contract of adhesion, such as the commonly used "free" licenses promoted by Creative Commons (e.g., CC-BY at https://creativecommons.org/licenses/by/4.0/).

- **Question 3:** To the extent an AI algorithm or process learns its function(s) by ingesting large volumes of copyrighted material, does the existing statutory language (*e.g.*, the fair use doctrine) and related case law adequately address the legality of making such use? Should authors be recognized for this type of use of their works? If so, how?

Response to Question 3: With respect to the ingestion of copyrighted content for AI technologies, to the extent that copies are made, the rightsholder in that copyrighted content is entitled to prohibit such use as an infringement (Section 106 of the Copyright Act) unless such use is licensed (Section 106) or such infringement is excused, most prominently in the U.S. by the doctrine of fair use (Section 107). Licensing for copying (whether transactional or collective) is, of course, a very old practice in American law and practice, and even in connection with AI licensing these uses has become increasingly common; for example, CCC itself has enabled collective licensing of published content for text and data mining for the purposes of research by scholarly and commercial researchers, and is well-aware of individual rightsholders who do the same on their own. Moreover, especially for certain types of research, use of materials for AI and mining purposes is itself becoming a "primary use" of those materials (as the rapidly increasing quantity of newly-created content exceeds the ability of human beings to read and understand it sufficiently) and, as such, is more or less at the point where it cannot qualify as fair use because such a primary use will necessarily exceed the scope of exceptions and limitations permitted by the "three-step test" of the Berne Convention to which the United States is a party.[7] The issue of Berne adherence, and specifically whether the application of fair use or another exception or limitation in a particular situation violates the three-step test, applies regardless of whether the use is commercial or non-commercial. For example, as text mining of materials created for non-commercial markets such as education and research becomes a primary use, the use cannot be excused as "fair" simply because it is non-commercial.

[6] 17 U.S.C. § 201(a) and (c).

[7] See discussion and citations to various versions of the Berne Convention, and to the manner in which it has been incorporated into U.S. law, in Circular 38A of the United States Copyright Office, https://www.copyright.gov/circs/circ38a.pdf.

Courts have recognized that the licensing potential of "novel" uses of copyrighted works will at times defeat a fair use defense. This can be true even under an allegation that the use was somehow "transformative," where the allegedly "transformative" use was merely a form of copying that allows the user to make a use for free that is well within the reasonable scope of the markets that the copyright holder is currently exploiting or can reasonably be expected to exploit in the short term. For example, in the Second Circuit decision in Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169 (2d Cir. 2018),[8] the court found that the creation of an index as a result of the ingestion of video clips had a competitive commercial impact on the ability of the rightsholder to license just such a use in an environment where users were beginning to demand it and rightsholders were beginning to license it, and that conclusion tilted the analysis of the four fair use factors towards a finding of infringement. Generally, our view is that the correct understanding of U.S. fair use law is that it does not favor wholesale and systematic ingestion of copyrighted content for clearly commercial purposes, when, regardless of ultimate use, (i) such "ingestion" is merely a form of copying – an act reserved to the rightsholder for the entire history of copyright law, (ii) the content has been made available to the public specifically for purchase, subscription or licensing, and (iii) the content copied has been specifically chosen for such purpose because of its value for such purpose; such activities amount to direct copyright infringement unless licensed. It is worth noting that, to a substantial extent, the acts performed by AI systems are similar to those performed by humans, even if on a different scale. Thus, if it is an infringement for organizations to make unauthorized copies of entire works for humans to learn from, to study, and to read, it is *a fortiori* an infringement to do so at scale for similar machine use.

Of course, some data and content that creators and other rightsholders make available online, often on social media platforms, may be impractical for licensing due to its high volume, small size and often "orphan" status (if the copyright holder does not make herself readily known or findable). For example, posts on Reddit or Twitter are protected by copyright (as writings fixed in a tangible medium of expression), but there is no intention by their authors to derive economic value from their posting. In a fair use analysis of these particular circumstances, copying of these materials would not lead to market harm, as there is no market nor is one likely to develop; the outcome of such a fair use analysis will very likely be different from the analysis performed on curated content from a professional creator or distributor of copyrighted materials.

We note that U.S. copyright law as it exists provides a framework for analyzing the legality of a particular use, even if it is suboptimal. As practitioners know, evaluating fair use is complicated and, in the often many years it takes for a final court ruling, the "facts on the ground" usually change in terms of markets, usage, rights, and licenses, rendering final decisions less useful in guiding forward looking behavior. While specific legislation creating new rules regarding the applicability of the Copyright Act to AI and text mining could, in theory, provide more certainty, the legislative process itself is slow moving and its outcomes are uncertain.

[8] See summary of the case provided by the Copyright Office at https://www.copyright.gov/fair-use/summaries/fox-news-network-tveyes-02272018.pdf.

- **Question 4:** Are current laws for assigning liability for copyright infringement adequate to address a situation in which an AI process creates a work that infringes a copyrighted work?

Response to Question 4: The ability to hold a natural or legal person responsible when AI infringes is an issue that will eventually need to be addressed. In the event that it is not "obvious" what person is responsible for the activity of the AI, then, in our view, liability should attach to the beneficiaries – that is, the persons or entities who benefit from the infringement, either directly or indirectly. Likewise, as is already true under U.S. law, the creation of an AI designed for infringement, even in the absence of a direct financial motive, should give rise to criminal and civil liability.

Any copyright infringement analysis requires a determination of whether "actual copying" (or another act identified by Section 106 of the Copyright Act as reserved to the copyright holder) is involved. In evaluating whether any such reserved act has occurred in the creation of an output, courts should consider not only whether the AI ingested a particular work at the direction of a natural or legal person, but also whether the AI was enabled by a person to ingest the work on its own. Meanwhile, access and substantial similarity will continue to serve as part of the proof of infringement where direct evidence of copying is lacking.

With respect to secondary liability that could arise when the operator of a commercially-oriented AI project is not directing that the AI identify and ingest significant and substantive copyrighted content, but where such ingestion occurs incidentally, CCC believes that existing standards for computer trespass[9] and copyright infringement should be adequate to protect copyrighted works and databases from unauthorized access. Generally, copyright infringement is a strict liability tort.

- **Question 13:** Are there any relevant policies or practices from intellectual property agencies or legal systems in other countries that may help inform USPTO's policies and practices regarding intellectual property rights (other than those related to patent rights)?

As a point of reference, the European Commission and Parliament have recognized a need under EU law for more clear regulation of use of copyrighted content in some instances of extraction and mining, as articulated in the 2019 Directive on Copyright in the Digital Single Market (Directive 2019/790) and expected to be transposed into the national law of Member States by mid-2021. Under this Directive, non-commercial scientific research using licensed content for text and data mining is a permitted exception under Article 3, and other lawfully accessible online content is also available for short-term mining/extraction under Article 4, provided that the rightsholder has not reserved its rights. These provisions provide for a "copyright exception"-based approach to the use of content ingested for AI purposes, and contemplate a viable market for licensing content for commercial AI use (subject to the possible Berne

[9] Under the federal Computer Fraud & Abuse Act, 18 U.S.C. § 1030.

Convention problem associated with such an "opt-out").[10] Japan, which like the EU does not include fair use in its copyright law per se, also allows some text mining and other forms of data analytics, so long as (1) the materials are lawfully acquired, (2) the use does not unreasonably prejudice the interests of rightsholders, (3) the use is "minor" in terms of the amount of each work used in the TDM effort, and (4) license terms are respected, including those forbidding such use in the absence of specific permission.[11]

CONCLUSION

In sum, CCC is of the view that natural persons will continue to play the predominant role in engineering and directing AI projects and thus will continue to be authors and contributors to copyrighted works produced through an AI mechanism to the extent that such works are fixed and otherwise non-functional and therefore copyrightable. AI works that involve multiple contributors should be analyzed as collective or joint works under U.S. copyright law and will often be dealt with as a matter of contract or work for hire status. CCC is also of the view that many "data-driven" AI projects will involve the ingestion of the copyrighted works of third-party rightsholders as such works are often specialized and otherwise useful for such projects. Licensing is the obvious market solution for the ingestion of professionally-produced and -curated copyrighted works for commercial purposes (otherwise, such use amounts to infringement) or where the use supplants existing markets.

Contact:

Roy S. Kaufman
Managing Director, Business Development and Government Relations
Copyright Clearance Center, Inc.
222 Rosewood Drive
Danvers, Massachusetts 01923
(978) 646-2463
rkaufman@copyright.com

[10] As copyright exceptions, they should be read narrowly. Thus, because these provisions apply to the reproduction right and not the making available or communication to the public right (an explicit part of European copyright law), there is no copyright exception with respect to the display or distribution rights relating to the output of the text and data mining effort by the user. This means that the direct output is intended to be limited to the internal use of the miner, further suggesting the need for licensing for any such use.

[11] See current Article 47-5 of the Japanese Copyright Act, unofficially translated here: http://www.japaneselawtranslation.go.jp/law/detail/?ft=1&re=02&dn=1&co=01&ia=03&ja=04&x=0&y=0&ky=copyright&page=24, and paragraph 2(4) of what we understand to be the Cabinet Order issued by the Japanese government to implement the statutory provision, available only in Japanese here: https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30_hokaisei/pdf/r1406693_05.pdf. For Japanese text of the law and related materials: https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30_hokaisei/pdf/r1406693_06.pdf; and https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30_hokaisei/pdf/r1406693_08.pdf.