Written Comments of

Copyright Clearance Center, Inc.
(submitted February 13, 2019)

In the content of the
WIPO Conversation on
Intellectual Property (IP) and Artificial Intelligence (AI)
Second Session
(Draft Issues Paper on Intellectual Property and Artificial Intelligence, Jan. 14, 2020)

INTRODUCTION AND BACKGROUND

Copyright Clearance Center, Inc. ("CCC") is providing these comments to the World Intellectual Property Organization (WIPO) concerning the impact of artificial intelligence ("AI") technologies on intellectual property law and policy, particularly with respect to copyright matters, in response to WIPO's questions as published on December 13, 2019.[1]

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, whether 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 and unpatented inventions and technologies

[1] https://www.wipo.int/about-ip/en/artificial_intelligence/call_for_comments/index.html

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.[2] 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 millions of 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 software applications, and – importantly – there is no broad, commonly accepted definition. In common parlance, these applications cover a wide spectrum of human activities, ranging from playing chess and speech recognition to autonomous driving and digital assistants (e.g., Amazon's Alexa, Apple's Siri). But no known application of AI exists that meets the ambitious goal (set by the pioneers of the field in the 1950s) of creating machines that can "think like humans." [3] 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 called machine learning (ML) and, in particular, a subfield of machine learning called "deep learning," which is a class of algorithms in which data is processed through multiple layers, from raw input to greater and greater levels of abstraction – each layer providing a more granular representation of reality (and thereby enabling the machine to perform, and to teach itself to perform, more and more sophisticated tasks).[4] 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 so on. 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,

---

[2] 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

[3] http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p005.htm

[4] However, we are aware that not all machine learning algorithms are trained in this manner.

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).[5] Current 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.[6] 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 also already using machine learning, and AI generally, towards gaining greater operational efficiencies.

As noted above, many AI practices involve the ingestion of copyrighted content, including the content found in journals, newspapers, books and databases, the rights for which comprise 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[7] 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 widely considered one of the most valuable assets for businesses and other organizations, deployed to operate successfully and efficiently.[8]

---

[5] 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 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.

[6] https://aws.amazon.com/rekognition/

[7] 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.

[8] 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.

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 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.

Against this background, CCC offers the following responses to questions raised by WIPO that are relevant to our views.

COPYRIGHT AND RELATED RIGHTS:

Issue 6: Authorship and Ownership

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 most AI algorithms, just as natural persons now author software. We see no reason for exceptions or *sui generis* protections to be written and implemented simply because software – a long-established technology – does something new. The old rules (answering the questions "Who is an author?" and "What is protected subject matter?") have over time been adjusted as society and technology have evolved. In our view there is insufficient reason to assume that they cannot continue to do so.

CCC is of the view that for copyright to attach, a human 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 algorithmic contributors which merely organize or provide data or code elements at or under the direction of another (human) 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 relevant jurisdictional copyright acquis.[9] 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,[10] such as the commonly used "free" licenses promoted by Creative Commons (e.g., CC-BY[11]), and that, in certain statutory regimes, it is governed by nonwaivable author rights.[12]

Issue 7: Infringement and Exceptions

With respect to the ingestion of copyrighted content for AI, machine learning and mining technologies, to the extent that copies are made, the rightsholder in that copyrighted content is entitled to prohibit such use as an infringement unless such use is licensed or such infringement is excused, most prominently in the U.S. by the case-specific fact-dependent doctrine of fair use (17 U.S.C. Section 107), or in the U.K., Japan and now the E.U. under fairly limited copyright exceptions.

Licensing for copying (whether transactional or collective) is, of course, a very old 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 "normal exploitation" of those materials (as the rapidly increasing quantity of newly-created content exceeds the ability of human beings to read and understand it sufficiently). As such, those materials are more or less at the point where bulk or systematic copying for mining and AI extraction can never qualify as "fair use" under U.S. law because such use will always (1) exceed the scope of exceptions and limitations permitted by the "three-step test" of the Berne Convention, and (2) cause market harm. 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" in the U.S. simply because it is non-commercial. For another example, the commonly-misunderstood Japanese text mining exception expressly recognizes that it does not apply where it conflicts with the normal exploitation.[13]

---

[9] See, e.g., in the U.S., 17 U.S.C. § 201(a) and (c).
[10] https://www.law.cornell.edu/wex/adhesion_contract_(contract_of_adhesion)
[11] https://creativecommons.org/licenses/by/4.0/
[12] For, e.g., German law treatment, see https://openjur.de/u/126846.html , OLG Hamm, Urteil vom 07.08.2007 - 4 U 14/07, cited before as: Az. 4 O 329/01; for, e.g., French law, see generally http://www.jurizine.net/2005/09/02/10-les-droits-moraux-de-lauteu.
[13] See https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30_hokaisei/pdf/r1406693_05.pdf 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

Generally, our view is that existing copyright law does not favor wholesale and/or 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.[14] It is worth noting that, to a substantial extent, the acts performed by AI systems are similar to those performed by humans, although typically on a different (vastly greater) scale. Thus, if it is, in the normal case, an infringement for organizations to make unauthorized copies of entire works for humans to learn from, i.e., 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, ephemeral posts on Reddit or Twitter are, as a matter of law, 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 the U.S., a fair use analysis of these particular circumstances would suggest that 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.[15]

As a point of comparison, 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), as 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 Convention problem associated with such an "opt-out"). 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.[16]

---

[14] Or explicitly exempted by statute.

[15] https://medium.com/@nturkewitz_56674/sustainable-text-data-mining-part-ii-us-and-fair-and-unfair-uses-770e4aad705

[16] See footnote 9 above.

We note that rulemaking in this area is especially difficult, as the "facts on the ground" change rapidly in terms of markets, usage, and rights required. It is for this reason we believe that licensing, with its constant interaction between the rightsholders and users, is the most fit-for-purpose tool in establishing norms for mining and extraction.

CONCLUSION

In sum, CCC is of the view that public policy should, for some time to come, align with the assumption 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 applicable copyright law and will often be dealt with as a matter of contract, work for hire or similar status under the relevant national regime. 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 clearly the most effective 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