Introduction

From a copyright perspective, there are four key questions posed by artificial intelligence (AI) technologies. They are:

1. Is the use of copyrighted material for machine learning (ML) legal?
2. Can Generative AI produce infringing material?
3. Can an AI machine be an author?
4. Is the AI algorithm and/or data-set copyrightable?

In this short memo, I discuss the first two issues and the role of collective management in this context. As to the latter two, I will only mention briefly that (a) there is a broad consensus, reflected in positions taken by various courts and governmental agencies, that a machine cannot be an "author" for purposes of copyright purposes, though a human author can use AI as a tool to create; (b) computer programs (written by human authors) are generally eligible for protection under copyright law; and (c) automatically generated large datasets are typically not copyrightable due to lack of originality but they may be protected under other legal doctrines.

In my analysis, I will use both domestic and international law, as appropriate. My focal point in discussing issues 1 and 2, above, will be the potential role of licensing. I will not discuss in detail the European Union (EU)'s sui generis protection of databases.[i]

This memo has three parts. I will first provide an overview of machine learning. I will then review essential aspects of copyright law applicable to AI. Finally, I will discuss the key role of licensing.

Overview of Machine Learning

Several jurisdictions have created or are considering creating incentives for the AI industry to develop in their territory. Indeed, the current situation has been described as an industrial "arms race."[ii] Those countries have adopted or are considering adopting legislation to address the legality, under copyright law, of using copyrighted material for ML purposes.

The term "machine learning" has become almost synonymous with AI because ML is the largest subset of the AI research field, with the fastest and most impressive advancements in recent years, especially those related to the creation of generative AI (GenAI) applications. ML is a field of study that provides algorithms capable of programing themselves by processing a data corpus. The ML process often has several steps. First, the machine is provided with (or is said to "ingest") a large amount of "raw" data, for example anything (within set parameters) it can find online. At that stage, the system makes a copy of the material to enable the system to analyze it, and additional copies can be made to accelerate data processing. That copy may also allow the reprocessing at a later point of the data corpus, for example to remove certain elements it contains.

At this point, the AI machine tokenizes the data in order to initiate its learning stage. For example, a Large Language Model (LLM-- such as GPT-4, which is foundational for ChatGPT) tokenizes

words or parts of words (strings of letters) while an image-based system tokenizes pixels of a digital image. A token typically takes the form of an integer. Once the machine has learned sufficiently to be able to answer prompts, which means that it can make predictions based on what it has learned from the data corpus, the data is usually "cleaned up" by humans, either by providing model answers to specific prompts, or by picking the best answer from several that the machine will provide to a prompt. The machine can be given instructions not to produce certain outputs, or not to answer certain prompts. For example, it can be instructed not to answer medical or legal questions or told not to produce outputs that contain strings of words (of a certain length or longer) to avoid copying existing texts or images. It can even be instructed to prohibit creating works in the "style" of living artists, as a lawyer from OpenAI announced the company would do in the near future.[iii]

Relevant Aspects of Copyright Law

Copyright law varies by jurisdiction, but almost every country recognizes a right of reproduction, a right of communication to the public (in the US, this is part of the right of public performance), and rights of translation and adaptation (in the US, this is part of the "right to prepare derivative works"). Those rights are contained in the most important international copyright treaty, the Berne Convention for the Protection of Literary and Artistic Works (hereinafter "Berne Convention").[iv] Most substantive provisions of the Berne Convention were incorporated in the World Trade Organization's (WTO's) Agreement on Trade-Related Aspects of Intellectual property Rights (hereinafter "TRIPS Agreement").[v] 115 countries (including the United States) and the EU are also party (as of June 2023) to the 1996 WIPO Copyright Treaty (WCT), which provides a so-called "making available" right.[vi]

Members of the WTO are not at liberty to adopt exceptions or limitations to copyright rights that do not comply with the *three-step test* that is used to measure the compatibility of an exception with international law. The test allows a WTO Member to challenge an exception to copyright rights adopted by another Member. While the test cannot be fully discussed in this short memo, one of its three "steps" considers the impact of the exception on the "normal exploitation" of the work.[vii] The TRIPS Agreement is "enforceable" via the WTO state-to-state dispute-settlement mechanism, which continues to function at the level of panels, and several Members have agreed to put in place a parallel appellate mechanism.

Importantly, licensing is a recognized form of "normal exploitation," as a WTO panel already recognized in a previous dispute.[viii] Indeed, licensing is present in the ML space as major companies have licensed data corpora to train AI machines. While some other users are relying on their view of fair use or other exceptions, there is not U.S. case law that specifically exempts all ML uses and international law has many nuances.

It is important to point out that, despite differences in national legal regimes, there will be, almost certainly, a part of the AI process that will go beyond the bounds of what is allowed under copyright law, including reproductions for machine learning purposes that subsist for long periods of time, or the production of outputs that are substantially similar (or indeed identical) to copyrighted material. The test for infringement in the ML context should not differ markedly from the test generally applicable to any other copyright infringement. That said, in the same way that a machine cannot be

an author or right holder, an AI machine cannot be said to be an infringer. However, the person who trains, sells, or uses the machine can be.

Though transient copies are generally allowed, the data corpus created during the ML process is often not transient and instead stored and repeatedly used for the reasons already mentioned above. This copying of a "substantial part of a copyright work is an act restricted by copyright, and so for legal reasons *it requires permission*." [ix] That permission can be granted by the right holder directly, by a collective management organization (CMO) on behalf of the right holder, or by law (exception). There is often an apparent misunderstanding in this context that something that is publicly available is "copyright-free." The fact that copyrighted material has been made available online does not mean it is not protected by copyright or is otherwise free to use. For example, many websites and platforms have terms of use that limit reuse and/or they implement technologies to prevent certain forms of copying and reuse. Moreover, there are, unfortunately, easily available illegal repositories of e-books online, and there are troubling reports that some of these were used for ML purposes.[x]

A number of jurisdictions have adopted specific legislation concerning the copyright aspects of the machine learning process. They include the European Union, Japan, Singapore, and Switzerland (see Annex 1). In the United States, there is uncertainty about the scope of the "fair use" exception in this context, a matter that should become more clear as federal courts issue decisions in several lawsuits pending as of this writing (June 2023). Meanwhile, commentators have been debating the impact of the decision by the Court of Appeals for the Second Circuit (New York) in Authors Guild v Google—the so-called "Google Books" case—that found that it was a fair use for Google to make copies of entire books for the purposes of allowing online search and the making available of "snippets."[xi] The case was cited with apparent approval in the recent Supreme Court opinion in Goldsmith v Warhol Foundation (May 2023).[xii] There are, however, notable differences between the fact pattern in *Google Books* and AI overall. Moreover, *Goldsmith* signals a possible tightening of the first fair use criterion, specifically when the *purpose* of the use is to create a work that can be used to "compete" with the plaintiff's work, as may happen with some forms of AI.

Finally, one should add that under article 12 of the WCT, the Contracting Parties must "provide adequate and effective legal remedies against any person knowingly […] remov[ing] or alter[ing] any electronic rights management information without authority." "Rights management information" is defined as " information which identifies the work, the author of the work, the owner of any right in the work, or information about the terms and conditions of use of the work, and any numbers or codes that represent such information, when any of these items of information is attached to a copy of a work or appears in connection with the communication of a work to the public." This was implemented in section 1201 of Title 17 of the U.S. Code, and in article 7 of the EU Directive on Copyright in the Information Society ("InfoSoc Directive").[xiii] This type of information is often removed during the ML process.

An argument is sometimes made that what AI does cannot *by its very nature* infringe copyright rights because unlike humans, machines only use works as "data." In Japan, this is referred to as "non enjoyment use."[xiv] Hence, an AI machine would be free to, say, translate any published book without infringing because it is only using the words in the book as data. This argument must fail for several reasons. First, the impact on the existing marketplace for works would be in serious jeopardy if AI

machines had a free pass. Second, while it is certainly true that AI machines neither "understand" nor "create" like humans, they do process semantic content and can produce commercially competitive outputs. Hence, from a human perspective, what AI machines do, especially when it comes to AI, is certainly relevant *for humans*—but then *for who else would it matter*? The exact way in which machines function should not be a decisive factor, in the same way that the exact way in which humans create and copy has not typically been a key factor.

The Key Role of Licensing.

Exceptions in domestic law cannot (under the above-mentioned three-step test) allow unlimited use and reuse of copyrighted content for ML. In other words, the demand for material for ML purposes will almost certainly exceed the bounds of exceptions—indeed it already does. The unavoidable gap can be filled by licensing. As the EU CDSM Directive notes, "[r]ightholders should remain able to license the uses of their works or other subject matter falling outside the scope of the mandatory exception provided for in this Directive for text and data mining for the purposes of scientific research."[xv]

Licensing is not just about granting permission to use a protected work; it is also about setting conditions for use and reuse. Though copyright is presented as an "exclusive right," including a right to say "no," right holders do not generate income by saying "no." They do so by saying, "yes, but," that is, allowing the use but imposing (negotiated) limits on the use of content they own. Absent appropriate licensing, a repeat of the fiasco of peer-to-peer (P2P) file-sharing, that is, massive infringement of copyright, is likely to occur.

Licensing can also address issues such as removal of rights management information and transparency requirements. Certain legislators have proposed measures designed to ensure transparency in the use (which works were used for ML) and reuse (including whether the output was produced by human or AI) of copyrighted works by AI.[xvi] The production of certain outputs (directly competitive with one or more works in the dataset) could be restricted, for example. In appropriate cases, attribution could be required, an application of the moral right protected under the Berne Convention.[xvii] In sum, licensing is poised to become a powerful and appropriate solution both to allow but also provide adequate boundaries for ML

Larger right holders may be able to negotiate direct licenses with major players in the ML space. Some of them already have. However, ML will happen throughout the world, and will be performed by both small and large companies and institutions. An effective and efficient solution would thus cover as much of the "repertory" of works that may be used for ML and be available in as many jurisdictions as possible. This points to a central role for CMOs active in the licensing of text and image works. Their international network and expertise would seem particularly appropriate in this context. Some of them also benefit under their national law from extended collective licensing arrangements. Licensing through CMOs would be a pillar "to safeguard intellectual property rights including copyright," and "to promote transparency, address disinformation … and how to responsibly utilize these technologies," objectives set by the G7 Summit recently held in Japan.[xviii]

Finally, as underscored recently by the *Harvard Business Review*, the current spate of lawsuits about the use of copyrighted material for ML purposes and specifically its use for GenAI purposes has raised doubts about the potential liability of companies interested in using those services:

> Businesses should evaluate their transaction terms to write protections into contracts. As a starting point, they should *demand terms of service* from generative AI platforms that *confirm proper licensure* of the training data that feed their AI. They should also demand broad indemnification for potential intellectual property infringement caused by a *failure of the AI companies to properly license data input* or self-reporting by the AI itself of its outputs to flag for potential infringement.[xix]

Licensing is an effective mechanism to ensure fair treatment of copyright owners, provide security to the businesses using AI services, and accelerate the development of the industry. It can also assist in providing transparency in protected works used to train AI machines. As there are innumerable right holders, all over the world, whose copyrighted material has been or can be used for ML purposes, a collective licensing solution would seem like the most logical step forward.

Annex 1

Examples of provisions adopted to-date (June 2023) to address the copyright aspects of ML

i. European Union: Articles 3 and 4 of the EU Directive on Copyright in the Digital Single Market (hereinafter the "CDSM" Directive")[xx] allow, first, reproductions for TDM made by "research organisations and cultural heritage institution" (art 3), with a parallel obligation to adopt appropriate "security and integrity" measures. The term "research organisation" is defined as a not-for-profit university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research" (art 2(1)).

The CDSM Directive then instructs EU member States to "provide for an exception or limitation" to the rights of reproduction, the right of communication to the public of works and right of making available for TDM in general (art 4(1)). However, this exception is subject to what is often referred to as an "opt out" because it applies "on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightsholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online" (art 4(3)).

ii. Japan: The Copyright Law of Japan was amended in 2009—that is, before the emergence of GenAI—to allow computerized "information analysis" defined as "to extract information, concerned with languages, sounds, images or other elements constituting such information, from many works or other much information, and to make a comparison, a classification or other statistical analysis of such information; the same shall apply hereinafter in this Article) by using a computer. The exception allows making a "recording on a memory, or [making an] adaptation (including a recording of a derivative work created by such adaptation), of a work, to the extent deemed necessary" (art 47*septiès*).[xxi]

iii. Singapore: In Singapore, the Copyright Law was amended in 2021 to make an exception to the rights of reproduction and communication to allow "computational data analysis" (CDA). CDA includes "using a computer program to identify, extract, and analyze information or data from the work or recording," and "using the work or recording as an example of a type of information or data to improve the functioning of a computer program in relation to that type of information or data."[xxii] There are limits to this exception. In particular, the user must have lawful access and the copy must be "made for the purpose of CDA or "preparing the work or recording for" CDA; and, importantly, the user must not "use the copy for any other purpose."[xxiii] There are also permitted forms of communication to the public, with specific restrictions. It is worth noting that any "contract term is void to the extent that it purports, directly or indirectly, to exclude or restrict" the uses permitted by this exception. [xxiv]

iv. Switzerland: In 2019, Switzerland amended its federal copyright statute to allow "[f]or the purposes of scientific research, [the reproduction pf] a work if the copying is due to the use of a technical process and if the works to be copied can be lawfully accessed."[xxv]

(June 2023)

References

[i] As provided for in Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.
[ii] For example, see Kevin Roose, "How ChatGPT Kicked Off an A.I. Arms Race," *New York Times*, 3 February 2023.
[iii] OpenAI To Bar Users From Copying Living Artists, Atty Says, Law360.
[iv] The original text of the Convention dates to 1886 but most countries are party to the latest revision of the Convention, the 1971 Paris Act. It protects, inter alia, the rights of reproduction, the right of communication to the public (arts 11, 11*bis* and 11*ter*); and the rights of adaptation and translation.
[v] Annex 1C of the Marrakech Agreement Establishing the World Trade Organization, 15 April 1994.
[vi] WIPO Copyright Treaty, 20 December 1996, art. 7, reads in part as follows: "authors of literary and artistic works shall enjoy the exclusive right of authorizing any communication to the public of their works, by wire or wireless means, including the making available to the public of their works in such a way that members of the public may access these works from a place and at a time individually chosen by them."
[vii] TRIPS Agreement, art 13; Berne Convention, art 9(2).
[viii] World Trade Organization, United States — Section 110(5) of US Copyright Act. Report of the panel, 15June 2000.
[ix] UK Government, Review of Intellectual Property and Growth, Supporting Document T: Text Mining and Data Analytics in Call for Evidence Responses, available at https://webarchive.nationalarchives.gov.uk/ukgwa/20140603093549/http://www.ipo.gov.uk/ipreview-doc-t.pdf.
[x] https://aicopyright.substack.com/p/the-books-used-to-train-llms.
[xi] Authors Guild v. Google 721 F.3d 132 (2d Cir. 2015).
[xii] Andy Warhol Foundation For The Visual Arts, Inc. V. Goldsmith Et Al, 598 U.S. ___ (May 18, 2023).
[xiii] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.
[xiv] Artha Dermawan, Text and data mining exceptions in thedevelopment of generative AI models:What the EU member states could learn from the Japanese "nonenjoyment" purposes? (2023) J. of World Intell. Prop. 1-25.
[xv] CDSM Directive, note xx above, Recital 18.
[xvi] For example, art. 13 of the proposed Regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, {SEC(2021) 167 final}, 21 April 2021.
[xvii] Berne Convention, note iv above, arts 6*bis* and 10*bis*(3).
[xviii] Quoted in Kazuaki Nagata, "G7 Digital Ministers Agree to Pursue Responsible AI as Chatgpt Booms," *Japan Times*, 30 April 2023.
[xix] Gil Appel, Juliana Neelbauer, and David A. Schweidel, Generative AI Has an Intellectual Property Problem, Harvard Business Review, April 07, 2023, available at https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem [emphasis added].
[xx] 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.
[xxi] Copyright Law of Japan, translated by Yukifusa Oyama, Copyright Research and Information center (CRIC).
[xxii] Republic of Singapore, Copyright Act 2021, ss. 243 and 244.
[xxiii] Ibid, s 244(2).
[xxiv] Ibid. s. 187(1).
[xxv] Switzerland, Federal Act on Copyright and Related Rights, as amended.