Consultation on Copyright in the Age of Generative Artificial Intelligence

Submission by

Professor Michael Geist

Canada Research Chair in Internet and E-commerce Law

University of Ottawa, Faculty of Law

Centre for Law, Technology and Society

January 2024

A. Overview

I am a law professor at the University of Ottawa where I hold the Canada Research Chair in Internet and E-commerce Law and serve as a member of the Centre for Law, Technology and Society. I focus on the intersection between law and technology with an emphasis on digital policies. I have edited multiple texts on Canadian copyright law and appeared many times before House of Commons committees on copyright law and policy. I submit these comments in a personal capacity representing only my own views.

The consultation raises several questions related to generative AI and copyright. I have focused on three in this submission:

(1) Should Canada proceed with a text and data mining exception as recommended in the 2019 Copyright Act review?
(2) Should Canada introduce legislative reforms to address the use of copyright works in large language models (LLMs) that are central to the development of generative AI technologies?
(3) Should Canada introduce legislative reforms to address copyright-related questions arising from the outputs of generative AI systems?

My submission argues as follows:

1. **Introducing a text and data mining exception into Canadian copyright law is long overdue and should proceed as a copyright reform priority**. Similar provisions are widely used in other jurisdictions. Proceeding with the exception would ensure that Canada implements a copyright framework for AI that encourages innovation and investment while also providing appropriate protections for creators.

2. **It is premature to introduce legislative reforms on the use of copyright works within LLMs.** While there are both technical and copyright related issues related to LLMs, the copyright issue is currently before courts around the world in multiple cases that raise questions related to inclusion of copyrighted works within LLMs, whether such use constitutes infringement, and the potential application of limitations and exceptions. Given that these issues should become clearer as those cases progress, the government should maintain a watching brief to determine how the cases before the courts develop, whether market-based licensing alternatives emerge, and how the technology adjusts to reflect copyright-related concerns.

3. **It is similarly premature to introduce legislative reforms to address the outputs of generative AI systems.** While many have expressed concerns about the occasional similarities between generative AI outputs and copyrighted works that may have been included within LLMs, a deeper dive into the technology suggests that infringement is very rare. The courts will again be called upon to examine these claims and the government should await those outcomes before proceeding with any potential legislative reforms.

I also note that the next statutorily mandated Copyright Act review is due. Before proceeding with any reforms, it would be useful to conduct an assessment of the implementation of the recommendations from the last review conducted by the Standing Committee on Industry, Science and Technology and scope out a future review to update on outstanding issues and address emerging ones such as generative AI.

B. Text and Data Mining Exception

The inclusion of an explicit exception for text and data mining (sometimes referred to as informational analysis) within the Copyright Act's fair dealing provisions is long overdue. The adoption of a specific text and data mining exception is consistent with the 2019 Copyright Act review, which extensively examined the issue and recommended:

The evidence persuaded the Committee that facilitating the informational analysis of lawfully acquired copyrighted content could help Canada's promising future in artificial intelligence become reality. The Committee therefore recommends:

Recommendation 23
That the Government of Canada introduce legislation to amend the Copyright Act to facilitate the use of a work or other subject-matter for the purpose of informational analysis.

The federal government has invested millions to support research and commercialization of AI in the hopes of making it a world leader. However, the current state of Canadian copyright law undermines this investment by inhibiting innovation through the creation of legal uncertainty and high barriers faced by the very groups the investment aimed to attract.

AI research works by making machines smart. Whether this is focused on automated translation, big data analytics, or new search capabilities, it is dependent on data being fed into the system. Machines learn by scanning, reading, listening, or viewing human created works. The better the inputs, the better the outputs and the more inputs there are, the likelihood that results are biased or inaccurate decreases.

Canadian copyright law inhibits this because restrictive rules limit the data sets that can be used for machine learning purposes, resulting in fewer pictures to scan, videos to watch, or text to analyze. Without a clear rule to permit machine learning, the Canadian legal framework trails behind other countries that have reduced risks associated with using data sets in AI activities in a manner that fairly treats both innovators and creators. Under the Canadian system, researchers either must risk copyright infringement by using protected works to make their machines smarter (which has a chilling effect on innovation), or severely limit the data sets used, thereby producing less "smart" machines than would be possible under a more open copyright regime. This raises concerns of bias and discrimination.

Within the current framework, the fair dealing rules provide some protections and allow for some use of copyrighted work by AI companies without permission. Canadian courts have ruled that it is a right that should be interpreted in a broad and liberal manner and there are several purposes that would permit some text and data mining activities – notably exceptions for research, education, and private study.

The corporations and high-profile talent attracted by the investment in the Canadian AI system have been calling for such an exception for many years. In 2018, various technology groups noted that the current uncertainties in the *Copyright Act* limit the ability for Canadian companies to “access a basic necessary resource to train their algorithms”.[1] Indeed, as of the time of this writing, of the 121 companies on the Government of Canada’s approved AI vendor list, 87 are Canadian, with almost all other vendors coming from competing nations with TDM exemptions.[2] Canada is one of the top countries in the world for AI research talent, with rates of growth currently exceeding that of the United States, the UK, Germany, France, and Italy.[3] Unfortunately, we lag behind in AI commercialization.[4] A clearly articulated copyright framework that allows TDM for commercial use is an important step toward changing that tide.

Internationally, other countries have addressed this issue through text and data mining exceptions. In the above-mentioned discussion, Microsoft noted that there was a broad acceptance of text and data mining exceptions around the world and that Canada is posed to fall behind and be at a global disadvantage without one. They also noted that, unsurprisingly, the countries that have/are considering these exceptions are at the forefront of research involving data analytics and artificial intelligence.

Countries such as the United States and Israel have elected to open up their fair use provision to exempt text and data mining activities, with Israel’s Attorney General issuing guidance to inform the exception.[5] Others have created specific exceptions. Examples of such exceptions can be found in Japan, the European Union, the United Kingdom, and Singapore.

The EU enacted a mandatory exception for text and data mining for the purposes of scientific research but has permitted rightsholders to “contract out”.[6] The U.K.’s exception allows copies of works to be made without permission of the copyright owner for the purposes of automated analytical techniques to analyze text and data for patterns, trends, and other information. The law does not allow contracts to restrict data mining activities, but the exception is limited to non-commercial research. The approach adopted by Singapore specifically exempts text and data mining in their Copyright Act by exempting copies of works wherein the “copy is made for the purpose of computational data analysis; or preparing the work or recording for computational data analysis.”[7]

In order to not fall behind internationally and position ourselves as a world-leader, Canada needs to adopt a broad exception for text and data mining. Ultimately, machine learning does not harm the primary purposes of the original work – the goal is not to republish or compete with

[1] Michael Geist, “Want to Keep Canadian AI Thriving?: Create a Copyright Exception for Informational Analysis” (October 18, 2018) online (blog): *Michael Geist* <https://www.michaelgeist.ca/2018/10/elementaicopyright/>

[2] Number of approved suppliers per country/region: USA 19; EU 7; Japan 4; UK 3. Note that there was also one company based in India, which does not currently have TDM-specific exemption; Government of Canada, “List of interested Artificial Intelligence (AI) suppliers” (modified 28 November 2023), online: <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/list-interested-artificial-intelligence-ai-suppliers.html>.

[3] https://www2.deloitte.com/content/dam/Deloitte/ca/Documents/press-releases/ca-national-ai-report-2023-aoda-en.pdf at 9.

[4] https://www2.deloitte.com/content/dam/Deloitte/ca/Documents/press-releases/ca-national-ai-report-2023-aoda-en.pdf at 15.

[5] State of Israel, “Opinion: Uses of Copyrighted Materials for Machine Learning“ (18 December 2022), online (pdf): <https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf> at 15.

[6] Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on copyright in the Digital Single Market COM/2016/0593 final - 2016/0280 (COD), Articles 3-4.

[7] The Statutes of the Republic of Singapore, Copyright Act 2021, <https://sso.agc.gov.sg/Act/CA2021?ProvIds=pr243-,pr244->

copyrighted materials, but to ensure that researchers and AI companies can mine the text and data for informational analysis purposes – thus including commercial uses in the exception will not harm rights holders and will facilitate Canadian innovation.

Recognizing that there are concerns about the scope of a text and data mining exception, the government should consider including transparency requirements alongside the exception. There is a need for all stakeholders – copyright owners, users, and the broader Canadian public – to have easy access to disclosures about what content has been used in training AI systems. A mandatory transparency requirement would be akin to the attribution requirements in some fair dealing exceptions. By providing attribution in the form of transparent disclosures, the text and data mining exception would enable machine learning while also providing necessary safeguards for creators to better monitor and respond to the permitted use of their works within this exception.

C. Inclusion of Copyright Materials in LLMs

The inclusion of copyright materials in LLMs has emerged as a major source of concern for some rights holders, who argue that their rights are being infringed upon by virtue of the inclusion of their works without permission. I believe it is premature to introduce legislative reforms on the use of copyright works within LLMs. Indeed, notwithstanding the calls for immediate legislative reform, I believe that there are better approaches that balance the copyright concerns with the policy goals of developing beneficial generative AI systems that may support a wide range of activities including education, health care, and commerce. As the UK Minister for AI and intellectual property recently noted, there is no rush to regulate the AI field.[8]

First, there are a myriad of cases currently before the courts worldwide. These cases are likely to provide a first analysis of many of the copyright-related concerns with the inclusion of copyright materials within LLMs. For example, in *Andersen v Stability AI Ltd*, there are four different claims.[9] They are infringement of copyright, the removal of copyright management information under the DMCA, publicity claims where the defendant's knowingly used the plaintiff's names in their products (by allowing a user to request art in their specific style) and style by allowing a request in their artistic identities, unfair competition claims under the Lanham Act for the use of their art for commercial gain without permission or proper attribution. In *Authors Guild v OpenAI Inc*, the claims focus on the use of published works to train LLMs.[10] This is done by reproducing the works and that this act is central to the quality of the OpenAI product.

There are many other cases that will canvass these claims. Generative AI companies will likely point to uses that do not infringe copyright, the inclusion of materials not subject to copyright protection, and the temporary nature of the reproductions, largely for statistical and analytical purposes. The AI companies typically do not reproduce actual full text of the underlying materials found in the LLMs.

Given the legal uncertainties, it would be premature to intervene with legislative reform at this time. Rather, the government should maintain a watching brief on the litigation to see how these

[8] https://www.theregister.com/2023/11/17/uk_ai_regulation/
[9] *Andersen v Stability AI Ltd*, 3:23-cv-00201, (N.D. Cal.):
[10] *Authors Guild v OpenAI Inc*, 1:23-cv-08292, (S.D.N.Y.):

cases unfold and whether reforms may be required. Intervention with legislative reform runs the risk of altering or undermining both creator and user rights as the technology continues to evolve, market-based solutions emerge, and courts address the application of LLMs to current copyright laws. Rather than leaping into reforms that may have negative effects and entrench the power of a handful of AI and tech companies, it is preferable to better understand how the law has been applied to LLMs and generative AI tools and then identify potential gaps or reform solutions.

Second, while allowing the litigation process to unfold, the government could encourage several private sector developments to the benefit of all stakeholders. These include greater transparency of which materials included within LLMs, akin to an attribution requirement for some fair dealing purposes. It could also include work toward an AI version of the robot.txt standard for data scraping. While the robot.txt standard has worked well for decades within the context of Internet search, there are other considerations in generative AI. Unlike search, generative AI tools may not direct the end user to the original source material, suggesting the mutual benefit in search may not be replicated in AI. As the legal process unfolds, a new standard specific to generative AI and LLMs is needed to allow rights holders to opt out of the inclusion of their works within LLMs.

Third, the government moved quickly last year to develop Generative AI guidelines that address issues related to transparency, security, and fair practices. While there were some concerns expressed with how the guidelines were drafted, they provide a useful starting point for governing activities in the sector. Since the guidelines are only effective if implemented, the government should actively ensure that they are respected by AI companies and work to identify whether further provisions or amendments are needed.

D. Outputs of Generative AI Systems

Similar to the analysis on the inputs into generative AI systems, the outputs are also the subject of litigation. While there are some cases involving questions regarding the copyrightability of machine-generated works,[11] the more notable issue at the moment is whether works created by a generative AI system may infringe copyright if they appear to replicate an original work that may have been included within an LLM.

However, notwithstanding some fears expressed in the media, replication is likely rare given the vast amounts of training data that are used to train an AI system.[12] For example, a study on GitHub's Copilot found that reproduction of code took place only 0.1% of times.[13] A study on replication in the LAION Aesthetics dataset, which includes 12 million images, found that 1.88% of random outputs had a high similarity score with the training material, which was considered a

[11] For example, In *Thaler v Perlmutter,* the plaintiff has alleged that the Register of Copyrights and Director of the US Copyright Office denied the copyright registration of a generative AI which he has claimed will generate works that should qualify for copyright protection. The claim was denied due to lack of human authorship.

[12] Andres Guadamuz, "A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs" (26 February 2023) Available at
SSRN: https://ssrn.com/abstract=4371204 or http://dx.doi.org/10.2139/ssrn.4371204 at 12 at 30.

[13] Ziegler A, "GitHub Copilot research recitation" (30 June 2021) Github Blog, <https://github.blog/2021- 06-30-github-copilot-research-recitation>.

high incidence rate among current studies.[14] The study noted that the reason for the high similarity results, was due to the prevalence of popular images in the dataset.[15] The study also used a small sample set, which only included 0.6% of LAION's training data.[16]

In sum, it is fair to say that there is a small degree of memorization, which can lead to replication, of certain popular sources across AI models, but current studies reveal that the rate is quite low.[17]The process of training an AI necessarily involves breaking large quantities of data apart, clustering, putting things that are similar together and then passing them through a noise filter.[18] At the end of this process, there is little left of the original work in the AI model, with some exceptions.[19]

The technological realities of generative AI suggest that infringing outputs is likely very rare. The government should not intervene with legislative reform to address what may be a non-issue. Rather, it should await the outcomes of litigation that may examine these issues in greater detail. Intervening at this premature stage, could harm creator rights, the development of AI technologies, and Canada's competitiveness in a rapidly growing sector.

[14] Somepalli G and others, "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models" (12 December 2022) arXiv http://arxiv.org/abs/2212.03860.

[15] *Ibid*.

[16] *Ibid*.

[17] Guadamuz, *supra* at 26.

[18] *Ibid*.

[19] *Ibid*.