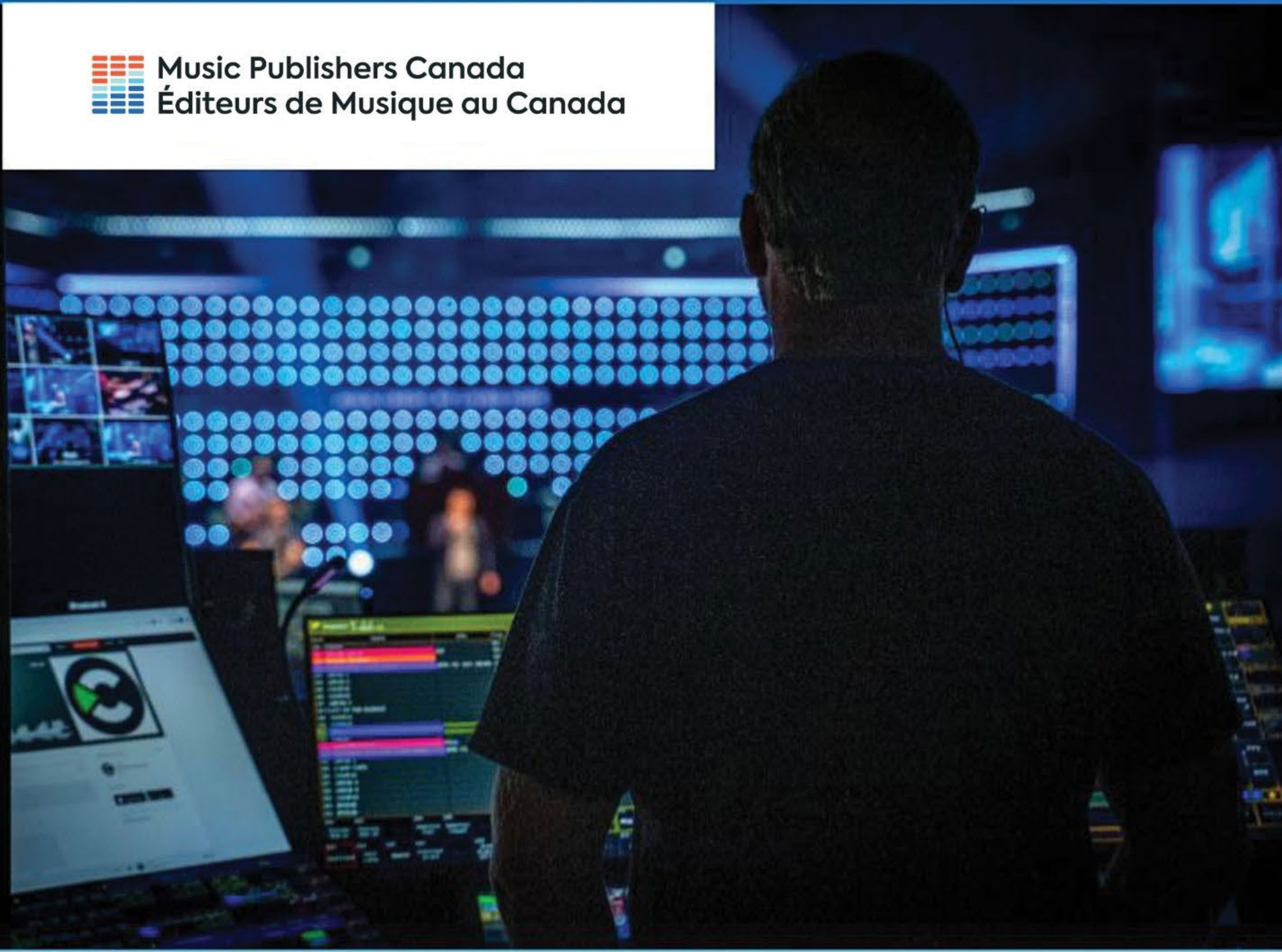
Music Publishers Canada
Éditeurs de Musique au Canada



# Opportunity to lead: Developing public policy on AI to protect and promote Canada's music publishing sector

## MPC submission to the Government of Canada's Consultation on Copyright in the Age of Generative Artificial Intelligence

DECEMBER 2023

Music Publishers Canada
Éditeurs de Musique au Canada

# Opportunity to lead: Developing public policy on AI to protect and promote Canada’s music publishing sector

## Introduction: Protecting copyright is key to AI’s enormous potential

Music Publishers Canada (“MPC”) is a membership-based organization that ensures the views of music publishers working in Canada are heard. It is our mission to create business opportunities for our members and to promote their interests and those of their songwriting partners through advocacy, communication, and education. Music publishers invest in thousands of Canadian songs and songwriters that are heard daily on the radio, on streaming services, in video games, and in film, television, and other screen-based productions around the world.

MPC recognizes that artificial intelligence has the potential to be enormously beneficial when it is implemented in a responsible and ethical manner, and MPC embraces that potential. In the music space, AI has the potential to support the valuable work of human creators, which in turn enriches Canadian culture and society. Our members are already exploring the benefits of this new technology.

However, the astonishing rate of both acquisition (or sometimes appropriation) of copyright-protected datasets and content on the input side, together with the development of generative

AI models on the output side, pose serious risks for Canada's creators and the companies that invest in them.

Copyright is the key protection that allows MPC's members to control and be paid for the use of their music. Copyright ensures that our members can share in the value if a third party wishes to use their music and ensure that the value is not appropriated solely by the user.

When an AI company uses music that has been scraped from the Internet without authorization, whether for training or other purposes, it prevents rights holders from controlling and realizing value for the use of their works. It also contributes to the destruction of a developing market for the licensing of copyright-protected content to AI developers before the market can flourish. Further, the development and commercialization of unlicensed AI model inputs and generative AI products can—and, in many ways, already are—creating serious market distortions, raising concerns about fair competition.

Human creation and expression, and their contributions to Canadian culture, must not be sacrificed at the altar of rapid technological progress. To strike the appropriate balance, Canada must approach generative AI in ways that respect creators and copyright and incentivize human expression. AI companies, like other commercial users, require permission from copyright owners to use copyright-protected content through negotiated licences.

The development of public policy surrounding AI is in its infancy. That gives the Government an important opportunity to lead the world in maintaining strong respect for copyright and the rights of creators.

# The use of AI in the music industry

MPC's members embrace technological change and invest in innovation. We know that there are already uses of AI in the music space. For example, AI is used to analyze and predict the audiences for an artist's music and to identify and target emerging artists. AI technology has also long played a role in the recording studio, including automation tools that can augment human-mixed recordings or even assist in the process of creating brand-new audio mixes.

But a new market is also developing for the licensing of music and other works to be used as training materials for generative AI models. Reported examples of the licensed use of copyright-protected content by AI companies include the following:

a) Meta's MusicGen tool was trained on 20,000 hours of licensed music from ShutterStock and Pond5;

b) Stability AI's new generative audio model, Stable Audio, was trained on a dataset provided under a licensing deal with a provider of stock music;

c) Universal Music Group (UMG) announced a collaboration with YouTube to experiment with generative AI tools predicated on human creativity and that account for the important interests of creators and copyright holders;

d) UMG has also announced collaborations with generative AI developers to explore how their technology can promote and enhance the creative process, such as with Endel, an AI tool that allows artists to generate content from their own sound recordings, and with Bandlab, the world's largest social music creation platform;

e) The generative AI imagery used during U2's new live show at the Spherem reportedly uses legally-ingested artwork by Es Devlin to spectacular effect.

As we discuss in more detail in the text and data mining section of this submission, large-scale licensing of copyright-protected works can be practicable and effective.

# AI developers use existing works as training material without authorization

AI developers obtain training material from multiple sources, including by scraping vast troves of content from the Internet and via pre-existing data sets. In many cases, the content is obtained without authorization from the rights holders and the AI developer does not disclose the source of the content.

This poses unique challenges in relation to musical works, which can be taken from the Internet in a variety of formats, including digital audio and audiovisual files containing musical works, song lyrics in text format, musical tablature and sheet music, MIDI files, and more.

Web-scraping occurs on a vast and indiscriminate scale and inevitably yields vast amounts of copyright-protected content. Scraping typically occurs without authorization of rights holders. In fact, it can be performed on webpages that, themselves, host and make available pirated content.

For example, it is reported that the training set used to train Google's T5, Meta's LLaMA, and other generative AI models, was trained using scraped content, including a subscription-based digital library, Scribd. OpenAI, a leader in generative AI, has also stated that its experience training AI models has involved "the use of large, publicly available datasets that include copyrighted works".

AI developers may also acquire training material from datasets collected and made available by third parties, some of which contain copyright-protected content obtained through large-scale web-scraping. A notable example is the Common Crawl dataset, which is a publicly available collection of large-scale web data used as the primary training corpus for most major LLMs.

OpenAI has reported that 60% of the training data for its GPT-3 model was drawn from Common Crawl.

From a copyright perspective, copying works to create a dataset, making the dataset available to be viewed and downloaded on the Internet, and exploiting the dataset to train, re-train, and fine-tune an AI model are each a separate and distinct exercise of a protected right.

# Licensing and transparency can mitigate liability

The Consultation asks, “What measures are taken to mitigate liability risks regarding AI-generated content infringing existing copyright-protected works?”

The best and most appropriate way to mitigate liability risk is for AI developers and dataset aggregators to obtain prior permission to exploit rights holders’ works, in accordance with Canadian copyright law and policy. The Government can incentivize that in two important ways.

First, a market is already developing for the licensing of music to AI companies. The growth and maturation of that market ought to be encouraged. Licensing large catalogues of music, including to new and disruptive technology companies, is what music publishers, copyright collective societies, and other rights holders do. MPC’s members, and the collective societies that represent them, have extensive experience negotiating bespoke licence agreements for the use of their repertoires by technology companies.  MPC implores the Government to permit the nascent market to develop and flourish, and not to eradicate it by introducing new or modified copyright exceptions for text and data mining or other AI activities.

Second, AI developers and data aggregators involved in any stage of training or testing AI models should also be required to disclose the dataset used to train or test the models and maintain complete and detailed records of that data. This requirement would ensure transparency and promote a functional licensing market, disincentivize unauthorized use of copyright-protected works, and restore the appropriate copyright balance by enabling rights holders to be compensated for the use of their works and to pursue enforcement options against infringers.

Certain market developments underscore the importance of ensuring that AI companies comply with existing copyright laws and norms. OpenAI, which as noted above has acknowledged training its models using vast amounts of protected content, is receiving massive investment from Microsoft, reportedly in the amount of $10 billion. In turn, Microsoft, Google, and other deployers of AI products have announced a commitment to indemnify users of certain of their AI products, if the users are sued for copyright infringement in connection with the use of those products. These developments, combined with the prospect of AI companies seeking to perform TDM activities in territories with the weakest copyright protections, suggest that, left unchecked, AI companies will continue to seek a competitive advantage built on a rampant neglect for the rights of creators and rights holders.

# Text and data mining

## 1. SUMMARY: TDM SHOULD BE LICENSED, NOT EXEMPT

The Copyright Act is intended to achieve a balance between promoting the public interest in the encouragement and dissemination of original works and obtaining a just reward for the creator or, more accurately, to prevent someone other than the creator from appropriating whatever benefits may be generated [Théberge v Galerie d'Art du Petit Champlain inc, 2002 SCC 34 at para 30].

This balance requires that AI developers obtain permission and pay for the use of copyright-protected works as AI training materials. TDM activities should not be given special status by introducing new copyright exceptions or modifying existing ones. In fact, a TDM exception would likely put Canada in breach of its international treaty obligations.

Copyright works add value to the AI training process. There is no legal or factual reason to allow AI developers to appropriate that value exclusively to themselves, especially by scraping online content on a vast and indiscriminate scale. To derive fair value for the use of their repertoires, music publishers routinely grant licences to technology companies. AI developers should be no different. The nascent market for licensing music to AI developers should be encouraged, including by requiring AI companies to disclose, and maintain records of, all their training data.

Finally, MPC urges the Government to reject any suggestion that TDM should engage a mere right of remuneration or be subject to an opt-out model.

## 2. NO NEW EXCEPTION FOR TDM

To maintain the proper copyright balance, the Government must reject calls for a categorical copyright exception for TDM. Rights holders must be able to control, and realize value for, the use of their works as AI training material, in accordance with Canadian copyright law and policy. Indeed, the ability to grant licences is central to the livelihood of creators and rights holders, and it is "a hallmark of copyright" [Euro-Excellence Inc v Kraft Canada Inc, 2007 SCC 37 at para 117].

Copyright content is a particularly valuable form of training material for an AI model. The quality of an AI model's output is proportional to the quality and quantity of its training materials. Copyright works are the products of human skill and judgment and the investment of time and resources. As such, many copyright works feature qualities that make them particularly valuable for use as AI training materials: nuance, richness, contemporary relevance, reduced "noise", integrity, reliability, and formatting consistency. All of this helps AI models turn out high-quality content that will attract and retain users. OpenAI has acknowledged that, if protected works are not used for training purposes, it would “lead to significant reductions in model quality” [ at footnote 33].

Rights holders must be entitled to control the use of their works as AI training material and share in the value created when they are used. That is best achieved through copyright protection and licensing.

Calls for a new copyright exception for TDM must be rejected. A new exception would eliminate the nascent licensing market before it can flourish and deprive rights holders of value for the use of their works as AI training material. An exception would appropriate the entirety of that value for the benefit of AI developers and data aggregators.

A categorical TDM exception would also violate the three-step test of the Berne Convention, violating Canada’s international treaty obligations. This test limits permissible exceptions to certain special cases that do not conflict with a normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the author” [Berne Convention for the Protection of Literary and Artistic Works (1979) at art 9(2)]. A categorical exception for TDM would not be limited to “special cases”. It would also eradicate the developing market for licensing works as AI training material, thus interfering with the normal exploitation of works and prejudicing the legitimate interests of rights holders.

Affording special status to TDM would also violate the principle of technological neutrality, which requires that copyright law operate consistently, and not favour or discriminate against any particular form of technology [Canadian Broadcasting Corp v SODRAC 2003 Inc, 2015 SCC 57, para 66].

The Consultation Paper notes that the existing exceptions for fair dealing and temporary reproductions for technological processes could potentially apply to TDM. While it is doubtful that these exceptions would apply to TDM, due in part to the application of the Berne Convention three-step test, the potential application of these exceptions is highly fact-dependent. It would not be prudent to attempt to address the potential application of these exceptions to TDM on a general or presumptive basis. That is a matter best addressed by the courts.

## 3. LICENSING MUSIC IS NOT AN INSURMOUNTABLE CHALLENGE

The music business is a licensing business. Any argument that licensing is impractical due to the quantity of data involved must be rejected.

Rights holders are experienced in licensing and administering large catalogues of works, including to technology companies. Canadian copyright collective societies like CMRRA and SOCAN process and license billions of lines of music data, or individual performances, each year. Any challenges that might arise are foreseeable and not insurmountable.

It also cannot be assumed that all AI models are trained on the same size of datasets. Introducing a TDM exception based on perceived challenges for AI models that train on massive datasets would destroy licensing markets for other AI training methods, such as the use of smaller and more carefully curated datasets.

In any event, any licensing challenges that may exist cannot justify the eradication of an exclusive right or the deprivation of a rights holder's ability to realize value for the use of its works.

## 4. RECORD-KEEPING AND DISCLOSURE IS CRITICAL

Transparency is critical to protect creators and rights holders and to strike an appropriate balance between fostering innovation in new technologies, on one hand, and protecting the legitimate interests and exclusive rights of rights holders, on the other. AI developers and deployers should be required to keep, and make readily available to rights holders, detailed and accurate records of their training data, the sources of that data, and the existence of any licences authorizing its use. Without those obligations, it will be extremely difficult, if not impossible, for rights holders to detect infringement and pursue enforcement options.

Further discussion on this topic can be found in our response to the section of the Consultation addressing infringement and liability.

## 5. THERE SHOULD BE NO COMPULSORY LICENCE OR OPT OUT SYSTEM FOR TDM

The Consultation Paper asks what "level of remuneration would be appropriate for the use of a given work in TDM activities." The appropriate level of remuneration should be determined in the developing market for the licensing of copyright-protected works for AI training and other uses, not by the Government.

To be clear, the Government should not entertain any suggestion that would eliminate a rights holder's exclusive right of reproduction in favour of either a compulsory licensing system or an "opt-out" system. Since a functional market for the licensing of musical works to technology platforms already exists, and is adapting rapidly to the needs of AI companies, there is no reason for the Government to impose either approach.

Compulsory licensing would be an extreme and prejudicial response to a non-existent problem. Among other things, it would (i) deprive rights holders of their right to contract freely in the market, preventing them from assessing and capturing fair value for the use of their works; (ii) prevent rights holders from choosing how their works are used and by whom, forcing them instead to allow the copying of their works for uses they cannot control or anticipate; and (iii) impose significant administrative burdens, including the creation of a needless and complicated infrastructure to administer and enforce the regime. Quashing an exclusive right of reproduction in favour of a right of remuneration would also raise serious concerns under Canada's international treaty obligations.

An opt-out system would also be antithetical to Canadian copyright law and policy and would drastically shift the copyright balance away from rights holders. In Canada, copyright is an opt-in system: prospective users of protected works must obtain advance permission from the copyright owner, who has an exclusive right to authorize—or refuse to authorize—the use. To depart from these fundamental principles would lead to an unwarranted erosion of copyright protection in Canada. It may also be contrary to the Berne Convention, which prohibits conditioning copyright protection on any formality requirement [Berne Convention, art 5(2)].

An opt-out system would place a disproportionate burden on creators and rights holders. It would require them to investigate and implement affirmative steps to prevent the infringement of their rights and to monitor compliance on a user-by-user basis. Rights holders who lack the legal or technological sophistication or resources to do so would be treated inequitably as unwitting licensors. In addition, because many rights holders do not control the websites on which their works appear, they would be unable to directly access the website's code to exercise an opt-out right. The same is true for online piracy sites that rights holders might not be aware of.

Finally, an opt-out approach would impose an all-or-nothing assumption on rights holders who may instead be willing to grant licences to use their works for specific purposes under certain conditions, including fair remuneration.

# Authorship and ownership of works generated by AI

Consistent with our submissions to the Government's "Consultation on a Modern Copyright Framework for Artificial Intelligence and the Internet of Things", MPC submits that Canada's existing copyright law framework is sufficiently robust and flexible to address issues raised by AI.

The Copyright Act is intended to incentivize human creativity and expression. For example, several provisions of the Copyright Act indicate that an author must be a human: sections 6, 7(1), and 9 link the term of copyright to "the life of the author", while section 14(1) imposes limitations on an author who is the first owner of the copyright based on "the death of the author". These provisions suggest that some degree of human involvement is necessary for a work to attract copyright protection.

Courts have confirmed on several occasions that authorship requires human involvement [see, for example, PS Knight Co Ltd v Canadian Standards Association, 2018 FCA 222 at para 147;

Setana Sport Limited v 2049630 Ontario Inc (Verde Minho Tapas & Lounge), 2007 FC 899 at para 4. They have also made clear that originality—the sine qua non for copyright protection—requires an author to exercise skill and judgment that is more than a purely mechanical exercise [see CCH Canadian Ltd v Law Society of Upper Canada, 2004 SCC 13 at paras 16, 25]. That too suggests strongly that a human author must be involved in the creation of a protected work.

The determination of copyright authorship and ownership rights related to AI-generated and AI-assisted works is highly fact-dependent. It should be made on a case-by-case basis. In the United States, the courts are now considering several cases that will assist with clarifying the boundaries of copyright law regarding generative AI content.

# Infringement and liability regarding AI

## 1. There are significant barriers to detecting infringement and enforcing rights

The Consultation Paper asks, "What are the barriers to determining whether an AI system accessed or copied a specific copyright-protected content when generating an infringing output?"

As noted in the Consultation Paper, to establish infringement by reproduction, a rights holder must establish that the defendant had access to the original copyrighted work, that the original work was the source of the copy, and that all or a substantial portion of the work was reproduced. A court may infer copying if the defendant had access to a plaintiff's work and there is substantial similarity between the works [Pyrrha Design Inc v Plum and Posey Inc, 2022 FCA 7, paras 38 & 48].

Without appropriate transparency, large-scale infringement might go undetected by rights holders. Rights holders only have access to the output of an AI system, from which it is nearly impossible to identify what copyrighted works were used in the dataset used to train the AI system. Even if a rights holder suspects infringement, it would be equally difficult, if not impossible, for a rights holder to establish that its work was used to train the AI model. This would create significant barriers to the rights holder's ability to obtain a remedy—and a right without a remedy is no right at all.

Thus, full transparency, including robust record-keeping and disclosure obligations, is necessary for rights holders to protect their intellectual property. Disclosure and record-keeping requirements would enable rights holders to know whether their works have been used. Standard copyright liability principles can then be applied to determine whether there has been an infringement, identify the infringer, and assess the resulting damages.

In addition, record-keeping and disclosure of the materials used in AI training and testing activities would accomplish three key objectives:

(i) promoting the development of a functional licensing market by incentivizing AI developers to seek authorization before using works and by disincentivizing unauthorized uses;

(ii) reinforcing the rights of creators and rights holders to control the use of their copyright-protected works and to obtain fair remuneration for such use; and

(iii) ensuring that the remedies in the Copyright Act are not nullified by the practical impossibility of detecting infringement and pursuing enforcement options.

Transparency would also serve consumer interests. While consumers, unlike rights holders, may not need to know exactly what data was fed into the AI system they are using, they should not have to guess whether the system was trained on legitimate, authorized copyrighted works rather than infringements or fakes.

For these reasons, developers and deployers of generative AI systems should be required to keep and make readily available detailed and accurate records of the data they have used for training, the source of that data, and the existence of any licences authorizing the use of that data. The records should be understandable to a layperson and detailed enough to identify (i) each specific work used in training, retraining, refining, or testing the AI model, or any similar use, (ii) any metadata associated with each work (e.g., title, author, owners), the immediate source of each work, (iii) the purposes for which each work has been used, and (iv) whether a licence has been obtained for each work.

Record-keeping and disclosure obligations should apply to every person involved at each stage of the training, retraining, refining, testing, and other development of the AI model, including dataset aggregators.

## 2. With record -keeping and disclosure obligations in place, the current copyright act would be sufficient to address AI -specific issues

The Consultation Paper asks, “Are there approaches in other jurisdictions that could inform a Canadian consideration of this issue?”

Requiring detailed logs of data used by an AI model is a necessary best practice. As an example of transparent disclosure obligations, the European Union’s draft Artificial Intelligence Act would require automatic logging of events while high-risk AI systems are operating. The logging capabilities must address, at a minimum, (i) the recording of the period of each use of the system; (ii) the reference database against which input data has been checked; (iii) the input data for which the search has led to a match; and (iv) the identification of the natural persons involved in the verification of the results" [European Commission, Proposal for a 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 ("EU Proposal"), at art. 12].

At the same time, the EU is also contemplating imposing similar record-keeping and transparency obligations on providers of so-called "foundational models", which are defined as AI system models that are trained on broad data at scale, are designed for generality of output, and can be adapted to a wide range of distinctive tasks. Also under discussion are provisions that would require providers of so-called "general-purpose AI systems" to maintain detailed technical documentation for at least 10 years. That would include "data requirements in terms of datasheets describing the training methodologies and techniques and the training data sets used, including information about the provenance of those data sets, their scope and main characteristics; how the data was obtained and selected; labelling procedures (e.g. for supervised learning), data cleaning methodologies (e.g. outliers detection)." [EU Proposal, at art. 50 and Annex IV].

Given that record-keeping by AI companies is important for purposes that extend beyond copyright protection, record-keeping obligations could be enacted outside the Copyright Act (for example, in the Artificial Intelligence and Data Act).

Importantly, AI developers should not be able to argue that they are shielded from liability for infringement because they do not retain copies of their training material once training is complete or that they did not compile the training data. Authorization from copyright owners is required before assembling and curating datasets that include copyright-protected works, much less before training AI models on those datasets. Whether or for how long the copies are retained is irrelevant.

With appropriate proper record-keeping and disclosure obligations in place, the current Copyright Act will be sufficient to address issues specific to AI. Liability could potentially arise for primary or secondary copyright infringement, moral rights infringement, removal of digital rights management information, and circumvention of technological protection measures.

# Conclusion

To strike the appropriate copyright balance, it is imperative that Canada approach generative AI in a manner that respects creators and copyright and incentivizes human expression. AI companies, like all technology companies, require permission from copyright owners before using copyright-protected content, whether to curate and assemble datasets or to train AI models on those datasets once assembled. That permission can and should be obtained through negotiated licences, not rendered moot by copyright exceptions, remuneration rights, or an opt-out system. Human creation and expression, and their contributions to Canadian culture, must not be sacrificed at the altar of rapid technological progress.

Canada should lead the international community in respecting creators. The development of public policy surrounding AI is in its infancy. That presents the Government with an important opportunity to lead the world in maintaining strong respect for copyright and the rights of creators.

Canada should not follow any international approaches to AI and copyright that would exempt or limit the scope of copyright protection in relation to TDM activities. MPC acknowledges and endorses commitments made by the G7, which broadly emphasize “multi-stakeholder” participation in the development of AI standards that prioritizes fairness, transparency, and adherence to existing law; commitment to “human-centric and trustworthy AI”; and continued discussion and analysis of how best to safeguard copyright and other IP rights [European Commission, “Hiroshima AI Guiding Principles and Codes of Conduct”; Government of Canada, “G7 Hiroshima Leaders’ Communiqué”].

## For more information

**Margaret McGuffin**
CEO
Music Publishers Canada
mmcguffin@musicpublishing.ca