September 23, 2026

Submission to Innovation, Science and Economic Development Canada as part of its public consultation on AI transparency

Advancing Transparency of Artificial Intelligence Systems

Submitted by: Centre for Media, Technology and Democracy

Prepared by: Manon Paquet, Associate Director (Policy)[1]

Introduction

The Centre for Media, Technology and Democracy (CMTD) welcomes the opportunity to contribute to the Government of Canada's consultation on advancing AI transparency. Our response draws on research led by CMTD over the last two years, and complements the recommendations to the federal government found in Taylor Owen's Final Report to the AI Strategy Task Force.[2] Grounded in that evidence, this submission focuses on the four areas of the consultation where we can speak from primary research: AI-generated content, AI interaction, information about AI systems, and AI incidents.

Before turning to those areas, we offer a few framing principles. The first concerns the purpose of transparency. The consultation frames it as a means to enhance trust and, in turn, encourage adoption of AI by Canadians. We share the adoption goal, but it is dependent on building a trustworthy AI ecosystem based on transparency that can be verified. A transparency regime designed mainly to reassure users and accelerate adoption can look quite different in practice from one designed to serve the public interest. A regime oriented toward adoption measures success through uptake, whereas one oriented toward the public interest asks whether disclosure actually reaches the people affected, so that they can make informed decisions, including the decision not to adopt a system. Adoption should be the product of a safe and trustworthy ecosystem.

A second principle concerns the relationship between transparency and accountability. We recognize that transparency is only one pillar of a broader federal AI strategy, and that this consultation is deliberately scoped to transparency alone. We nonetheless emphasize that transparency is a precondition for accountability, not a substitute for it. Transparency must be paired with the duties (e.g., safety standards, independent verification, and the obligation to stand behind AI outputs) that make the ecosystem safe and trustworthy.

A third is the clear public appetite for binding measures. Public opinion in Canada points toward mandatory rather than voluntary rules. The results of our survey published in February 2026 show broad concern among Canadians about risks related to AI chatbots. Sixty-nine

[1] This memo was written with the support of Ben Steel, Esli Chan, Sonya Solomon, Aengus Bridgman, and Taylor Owen.
[2] Owen, Taylor. 2026. Final Report to the AI Strategy Task Force. Montreal: Media Ecosystem Observatory, Centre for Media, Technology and Democracy, McGill University.



percent want stricter regulation, and 74–80% support specific safeguards including age verification, parental oversight, data limits, and content restrictions.[3]

A workable regime also depends on who is expected to surface harm, and should draw on two complementary sources. Vendors are often best placed to detect emerging harms in their own systems, and some already engage in valuable unprompted disclosure; the regime should place developers and deployers under clear obligations to surface the risks and incidents they observe, building on the digital safety plans proposed in the *Safe Social Media Act* as a model for other AI applications. But self-reporting alone is inconsistent and cannot be the whole system; independent researchers play an indispensable role in holding developers and deployers to account — testing systems and surfacing harms that the value chain does not report or cannot see. CMTD's own audits and election monitoring, detailed below, have brought to light safety failures and information-integrity harms that were not otherwise disclosed. A regime built for the public interest should therefore recognize, protect, and adequately resource this independent-oversight function alongside its obligations on vendors.

Finally, we would also like to point to CMTD's *Gen(Z)AI: Youth Voices, AI Futures* (2026),[4] which convened a seven-month Youth Assembly on Artificial Intelligence and produced seventeen recommendations. Several of these recommendations are directly relevant to this consultation, including standardized content labelling, greater transparency in algorithmic recommendation, and authentication of content origins. As Canada designs transparency rules that will shape the information environment young people grow up in, the report models an approach of governing *with* youth rather than only *for* them, which we encourage the government to build on. Mandated consultation mechanisms would institutionalize democratic input into the design and oversight of AI policy.[5]

AI-generated content

There are two areas that we would like to address related to transparency with respect to AI-generated content: transparency about the sources an AI system draws on, particularly with respect to the use of news media; and labelling of the AI-generated content itself.

[3] Lavigne, Mathieu, Helen A. Hayes, Esli Chan, and Chris Ross. 2026. *Canadians' Perspectives on Governing AI Chatbots: A Policy Brief*. Montreal: Media Ecosystem Observatory, Centre for Media, Technology and Democracy, McGill University. https://doi.org/10.66536/ai-chatbots-perspectives_2026.

[4] Hayes, Helen A., Fergus Linley-Mota, Madeleine Case, Julian Lam, Alexander Martin, and Nonso Morah. 2026. *Gen(Z)AI: Youth Voices, AI Futures — Policy Recommendations for AI and Online Harms Governance in Canada*. Montreal: Media Ecosystem Observatory, Centre for Media, Technology and Democracy, McGill University. https://mediatechdemocracy.com/en/publications/genzai-youth-voices-ai-futures_2026/. [5] Owen, *Final Report to the AI Strategy Task Force*.

Sourcing Canadian journalism

The increased reliance on AI tools for news has significant implications for the sustainability of Canadian journalism. Our audit[6] on AI chatbots and news media found that in a majority of cases the systems reproduced enough of the original reporting that a reader would have little reason to visit the source, and in some cases drew on paywalled material. Evidence from the EU on Google's AI Overviews suggests these summaries are associated with reduced referral traffic;[7]the same dynamic is likely affecting Canadian publishers. Many AI systems reproduce journalistic content while returning neither consistent credit nor referral traffic on which news businesses depend.

We want to be clear about what attribution can and cannot do. While it is a necessary transparency measure and a precondition for any fair dealing between AI developers and publishers, it will not solve the underlying economic problem on its own. Attribution should therefore be paired with mechanisms that address the transfer of value from publishers to AI developers so that transparency about sources translates into a sustainable relationship rather than a polite citation; we would refer, in particular to recommendation 2.3 of the *Final Report to the AI Strategy Task Force.*[8] Unless this concern is addressed, Canadian journalism will struggle to remain financially viable in the age of AI.

Labelling output

We support requirements to label AI-generated content and welcome the fact that some requirements have been included in the proposed *Safe Social Media Act* as it relates to synthetic content circulated on social media platforms. Invisible watermarking and visible, point-of-exposure labelling are not competing options but complementary tools that do different jobs, and both are needed.

Watermarking, such as embedding a machine-readable signal in synthetic content, provide the infrastructure. These watermarks travel with the content, are more durable across re-posting than manual labels, and let platforms and tools detect what is synthetic. However, on their own, they do not reach the person scrolling a feed, who cannot see a machine-readable watermark unless they have the skills, tools and willingness to dig into it further. Visible labelling at the point of exposure would achieve that objective.

Neither tool is complete on its own. Watermarks can be stripped or degraded, coverage is uneven across generators, and determined malicious actors are likely to evade them; visible

[6] Bridgman, Aengus, and Taylor Owen. 2026. A*I News Audit: How AI Models Use and Distribute Canadian Journalism*. Montreal: Media Ecosystem Observatory, Centre for Media, Technology and Democracy, McGill University. https://doi.org/10.66536/ai-news-audit_2026.

[7] Lucchi, Nicola, *The Impact of Google AI Summaries and Google AI Overviews on Publishers' Revenue and Media Freedom Implications for the Information Ecosystem and Democratic Resilience in the European Union* (June 18, 2026). European Parliament Policy Department Briefing, PE 787.211, April 2026, Available at SSRN: https://ssrn.com/abstract=6963918 or http://dx.doi.org/10.2139/ssrn.6963918.

[8] Owen, *Final Report to the AI Strategy Task Force*.

labels depend on platforms actually applying them. These limits are reasons to require both and to keep obligations across the chain. Implementing standards for anti-circumvention or synthetic media detection automated within AI systems adds a further layer of defence against manipulated content that evades labels. We acknowledge the challenge that labelling regimes may create an inference about what is not labelled. It is therefore important to pair label obligations with public communications about what a label does or does not certify so that labels are not treated as a direct proxy for a healthy information environment. This aligns with recommendations 2.1 in the *Final Report to the AI Strategy Task Force.*[9]

AI interaction

The consultation asks what makes disclosure of AI interaction meaningful. Knowing that one is interacting with an AI is a necessary baseline, but on its own it may fall short. Our engagement with young Canadians suggests many already assume they are talking to an AI system; what they are far less aware of is how the system may be shaping the information they receive.[10] Meaningful disclosure should therefore convey not only that a user is dealing with an AI but the implications of doing so as a way to increase AI literacy amongst users.

One implication deserves particular attention: the system's incentives. Knowing that a system is automated tells a user nothing about what it is optimized to do. A chatbot that would be tuned to maximize engagement or session length, or to be agreeable and validating, is not a neutral information source, and a user weighing its output should be able to know that. This is not entirely a novel principle, but arguably a novel application of transparency measures already in force in some jurisdictions on social media platforms. Specifically, it could be understood as the AI equivalent of Article 27 of the EU *Digital Services Act*, which requires platforms to disclose the "main parameters" of their recommender systems, their relative importance and how users can influence them. In Canada, some of this information may surface in digital safety plans that AI chatbot operators would have to submit under the *Safe Social Media Act*, if passed, in relation to risks related to harmful content and harmful behaviour.

A disclosure obligation carries a risk the government should pre-empt: that deployers will use the notice to shield themselves ("we told you it was only an AI, so it is on you"). That would go against the public-interest purpose of transparency. The positive alternative is to pair the notice obligation with an explicit accountability duty, where the deployer has to be answerable for what its AI system tells users. While the debate over the legal status of AI chatbots continues, cases testing this principle have already reached the courts. For example, in *Moffatt v. Air Canada* (2024), the British Columbia Civil Resolution Tribunal held Air Canada responsible for inaccurate information its website chatbot gave a passenger, rejecting the airline's argument that the chatbot was "a separate legal entity responsible for its own actions"

[9] Ibid.
[10] Hayes et al., *Gen(Z)AI*.



and finding the company liable for negligent misrepresentation.[11] TThis principle should be enshrined in “binding legal mechanisms that assign clear accountability and remedies,”[12]thus ensuring that it need not be re-litigated case by case. Deployers should be accountable for the material representations, advice, and commitments their AI systems convey to users and clients.

The design remedies for engagement-maximizing or sycophantic patterns (e.g., usage limits, opt-outs, and stronger protections for minors) are duty-of-care measures rather than transparency measures; we note them only to be clear that transparency about incentives complements, and does not replace, those protections. Here again, the inclusion of AI chatbots in the *Safe Social Media Act* framework will help advance a safer, trustworthy AI ecosystem.

Information about AI systems

One of the most relevant concerns is where safeguards sit within AI systems, and how well they work at each layer, including input/output classifiers, account suspensions, and model safety. This is largely undisclosed, non-standardized, and difficult to verify from vendor statements alone. Our online-harms audit tested leading AI chatbots with hundreds of multi-turn adversarial attacks across the harm categories in Canada's proposed online safety framework. We found that protection varied enormously between comparable products.[13] That variation reflects choices by developers about the layer in which to implement safeguards.

A user or regulator currently cannot tell where a given product's protections sit or how well they perform. A model or system card that simply asserts a product is "safe" is close to meaningless; documentation should instead disclose, in standardized form, which safeguards are present, at which layer, against which harm categories, and how they routinely perform.

Our audit deliberately evaluated the developer system reached through the API separately from the consumer system reached through the web interface, and found they can diverge sharply. A Canadian business integrating a foundation model through an API may therefore ship a far less protected product than the vendor's own consumer app, without realizing it, because the safeguards it assumed it was inheriting live in a layer it never touches. This is precisely the information a business needs to deploy AI responsibly, and it is currently opaque.

Transparency here means disclosing, layer by layer, what protections are inherited at the model or API level versus what a deployer must add itself. None of this forces vendors to reveal trade secrets: a deployer and a regulator can be told that a product applies an output classifier for a given harm category, at a given layer, and how it performs against a standard test, without

[11] *Moffatt v. Air Canada*, 2024 BCCRT 149. https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html
[12] Owen, *Final Report to the AI Strategy Task Force*.
[13] Steel, Ben, Taylor Owen, and Aengus Bridgman. 2026a. *Online Harms AI Audit: Technical Brief*. Montreal: Media Ecosystem Observatory, Centre for Media, Technology and Democracy, McGill University. https://doi.org/10.66536/online-harms-ai-audit-technical-brief_2026.

being handed the classifier itself. Where a safeguard sits and how well it works against a standardized test is not a trade secret; the exact engineering of it can remain one.

Disclosure of this kind, however, is not self-verifying. The facts above surfaced only because our team tested the deployed systems directly; vendors' safety disclosures are insufficient on their own. Standardized safety reporting must therefore be paired with a durable capacity to verify it. Third-party AI assessment can strengthen accountability only when frameworks guarantee evaluator independence, cover the full range of potential harms, and treat assessment as one part of a broader regulatory toolkit rather than a standalone fix.[14] Concretely, we recommend protected access for accredited researchers and regulators to deployed systems with a safe harbour for good-faith adversarial testing and, where appropriate, regulatory sandboxes that allow systems to be examined under controlled conditions.

This would complement recommendations we have already made in our policy brief that accompanied the audit with respect to the trade-offs of regulating where the safeguards should reside.[15] First, tie duties to outcomes, not paperwork—defining "adequate" as demonstrated performance under realistic, adversarial testing, not the existence of a safety plan. Second, empower the regulator to test deployed products directly and publish the results, since a duty no one can check is a duty in name only. Third, decide deliberately where duties sit, anchoring them in the consumer app users meet while weighing how far to extend them to the developer system behind it. Fourth, resource the regulator like the UK's Ofcom or Australia's eSafety Commissioner, with in-house capacity to run its own audits.

Ultimately, a strong monitoring and compliance capacity needs three important and complementary: self-reporting requirements for the industry, a regulator with strong audit capacity, and protections and support for independent researchers.

AI incidents

The consultation's examples of AI incidents draw on product-safety templates, such as consumer products, medical devices, and transportation. While several recent incidents have stressed the need to be transparent about operational malfunction, we would suggest that a definition of incident not be limited to those but also encompass cases related to information integrity where AI tools are used to disrupt Canadians' public square and undermine our democratic processes.

The definition should be harm- and severity-based, which would capture new forms of harm as they emerge. For what makes an incident "serious," we suggest criteria such as the scale and reach of the affected audience; the vulnerability of those targeted; the reversibility of the harm;

[14] Miranda Bogen, Matt O'Shaughnessy & Anna-Maria Gueorguieva, Getting Third-Party AI Assessment Right, Center for Democracy & Technology (Sept. 15, 2026), https://cdt.org/insights/getting-third-party-ai-assessment-right/. [15] Steel, Ben, Taylor Owen, and Aengus Bridgman. 2026b. *Online Harms AI Audit: Policy Memo*. Montreal: Media Ecosystem Observatory, Centre for Media, Technology and Democracy, McGill University. https://doi.org/10.66536/online-harms-ai-audit-policy-memo_2026.

whether public safety or democratic processes are implicated; and whether the content is part of a coordinated or inauthentic operation. Severity is a function of impact, not of the underlying technique. This is the approach the EU *AI Act* and the Organisation for Economic Co-operation and Development (OECD) already take: both define a "serious incident" by its impact (e.g., death, serious or irreversible harm, or breaches of fundamental rights affecting many people) rather than by the technology involved.[16] This approach would contribute to the government's goal towards global interoperability.

These are not hypothetical risks. During the 2026 Quebec election, our monitoring documented AI-generated deepfake videos of party leaders circulating in a live campaign, including degrading synthetic videos targeting a female party leader.[17] The Alberta separation referendum offers an even sharper illustration for a consultation about transparency. Our monitoring found AI-generated imagery surging in the referendum debate and, most directly relevant here, an audit of the consumer AI tools many people now use to inform themselves found them unreliable on the basic facts.[18] Similarly, an AI chatbot audit during the provincial election showed that chatbots provided voting advice based on incomplete information on the position of political parties.[19] Deliberate manipulation of the tools' information sources risks compounding this unreliability.

Further, our *Slopaganda* investigation identified a network of twenty inauthentic YouTube channels with nearly forty million views that used human actors together with AI-generated voiceovers and avatars to impersonate Albertan voices and normalize secession and annexation narratives,[20] while investigative reporting documented a US-linked network of AI-generated fake websites apparently designed to be cited by AI chatbots ahead of the same

[16] OECD (2024), "Defining AI incidents and related terms", *OECD Artificial Intelligence Papers*, No. 16, OECD Publishing, Paris, https://doi.org/10.1787/d1a8d965-en.

[17] Media Ecosystem Observatory. 2026a. "Weekly Update 1: The Trade War Enters the Quebec Election Campaign." Centre for Media, Technology and Democracy, McGill University, September 4, 2026. https://mediatechdemocracy.com/en/publications/qc26-weekly-update-1_2026/.

~~[18]~~ Ross, Chris, Ben Steel, David Hobson, Saewon Park, Esli Chan, Bradley Wood-MacLean, Mika Desblancs-Patel, and Aengus Bridgman. 2026. *Alberta's Referendum, Online: Influencers, AI Content, Foreign Interest, and the Attention Economy Ahead of the October 19 Vote*. Montreal: Media Ecosystem Observatory, Centre for Media, Technology and Democracy, McGill University. https://doi.org/10.66536/ab-referendum-online_2026. [19] Media Ecosystem Observatory. 2026b. "Weekly Update 3: Quebec's First Election in the Age of AI." Centre for Media, Technology and Democracy, McGill University. https://mediatechdemocracy.com/en/publications/qc26-weekly-update-3_2026/.

[20] Ross, Chris, Ben Steel, Zeynep Pehlivan, Mika Desblancs-Patel, and Aengus Bridgman. 2026. *Slopaganda: The Inauthentic YouTube Network Selling Secession to Albertans*. Montreal: Media Ecosystem Observatory, Centre for Media, Technology and Democracy, McGill University. https://doi.org/10.66536/slopaganda-alberta-secession_2026.

vote.[21] Together these show an information environment in which the AI systems people consult are at once unreliable by default and actively targeted and exploited.

Beyond elections, information-integrity harms are chronic in the climate domain and spike during environmental crises. CMTD has documented the resonance of climate-delay narratives in Canada and the information environment around wildfires.[22] Wildfire disinformation carries a direct public-safety dimension and generative AI acts as an accelerant, from fabricated imagery to chatbots that reproduce misleading narratives. Content depicting or purporting to depict an active emergency or official emergency communication should therefore be treated as an aggravating factor in any test of severity and impact on public safety.

A recurring feature of these cases is who surfaced them: independent researchers and journalists, not the platforms that hosted the content or the developers whose tools produced it. The framework proposed in the *Safe Social Media Act* and the *Protecting Privacy and Consumer Data Act*, AI chatbot operators will be required to file digital safety plans with the future Digital Safety and Data Protection Commission, setting out the risks their systems pose, the safeguards they apply, the incidents they observe, incident response protocols, and updating them as those systems evolve. But a regime that relies on self-reporting alone will still systematically miss coordinated, cross-platform information-integrity operations that emerge as a result of organic information distribution or deliberately by malicious actors. The response should also strengthen regulators and support and protect independent researchers. The proposed legislative frameworks would ensure some significant progress towards this objective that was also highlighted in recommendation 2.2 of the *Final Report to the AI Strategy Task Force*.[23]

Recommendations and conclusion

Drawing on the analysis above, and following from a single principle — that transparency should be built for the public interest and paired with the accountability that gives it force — we recommend that the government:

AI-generated content

- Require default-on source attribution for AI outputs, paired with a value-transfer mechanism for news.

[21] *Canada's National Observer*. 2026. "A US-Linked Network of Fake Websites Is Promoting Alberta Separatism to AI Chatbots." September 4, 2026. https://www.nationalobserver.com/2026/09/04/investigations/network-fake-websites-alberta-separatism-ai-chatbots.

[22] Solomun, Sonja, Isadora Borges Monroy, Julia Bugiel, Esli Chan, Nikhil Gowd, Helen A. Hayes, Nina Hernández Jayme, Sequoia Kim, Hannah Tollefson, and Chris Ross. 2025. *Climate Obstruction: On the State and Spread of Climate Disinformation in Canada*. Montreal: Media Ecosystem Observatory, Centre for Media, Technology and Democracy, McGill University. https://doi.org/10.66536/climate-obstruction_2025.

[23] Owen, *Final Report to the AI Strategy Task Force*.



- Require both machine-readable watermarks and visible, point-of-exposure labelling of AI-generated content.

AI interaction

- Make disclosure meaningful — including what a system is optimized to do, not only that it is an AI.
- Codify deployer accountability for what types of information AI systems tell users; a disclosure notice is not a defence.

Information about AI systems

- Require standardized, comparable disclosure of where safeguards sit and how well they perform.
- Empower and resource an independent regulator to test deployed products and publish the results, with duties tied to demonstrated performance, and any accreditation regime designed to guarantee assessor independence.

AI incidents

- Define a “serious AI incident” by harm and severity, technology-neutral, including information-integrity and public-safety events.
- Three-fold approach to monitoring: require reporting by developers and operators; ensure strong regulator capacity to audit and support and protect independent monitoring.

Taken together, these measures would make transparency do more than reassure. Voluntary, default-off, and unverifiable disclosure has already fallen short in the systems Canadians use every day, and the harms that matter most are surfaced today largely by independent researchers rather than by the firms responsible. Transparency built for the public interest, and paired at every step with the accountability that gives it force, would begin to change that. We would welcome the opportunity to discuss this evidence and these recommendations further.