CENTRE FOR LAW AND DEMOCRACY

Canada

September
2021

Submission to the Consultation on the Government of Canada's Proposals to Address Harmful Content Online

Centre for Law and Democracy
info@law-democracy.org
+1 902 431-3688
www.law-democracy.org
fb.com/CentreForLawAndDemocracy
@Law_Democracy

Table of Contents

1. Introduction[1]

This Submission is pursuant to the ongoing consultation on the Government of Canada's proposed framework for regulating harmful online content. The Government released two documents as part of this consultation: a discussion paper, which outlines the Government's broad proposals,[2] and a technical paper,[3] which contains detailed legislative drafting instructions. This Submission collectively refers to both documents as "the proposal" but focuses on the technical paper to assess the proposal's specific features.

Addressing harmful content online is one of the most controversial topics in the world. While the widespread distribution and amplification of online content by social media companies have created unprecedented opportunities for people to connect and express themselves, the dark side of that freedom has become increasingly evident. Social media companies have been implicated in events such as the genocide of the Rohingya in Myanmar[4] and the 6 January 2021 Capitol riot in Washington D.C.[5] Accordingly, many countries, such as Singapore,[6] Nicaragua[7] and Ethiopia,[8] have introduced a wide array of laws and measures to regulate online content and social media companies.

Canada's proposal to regulate harmful content online is a mixed bag in terms of compliance with human rights. For instance, the proposed user-flagged content moderation system contains several positive features, notably the decoupling of social media companies' initial content moderation decisions from liability[9] and the independence of the various regulatory

---

[2] Canadian Heritage, Have your say: the Government's proposed approach to address harmful content online: discussion guide, 29 July 2021, https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content/discussion-guide.html.

[3] Canadian Heritage, Have your say: Government's proposed approach to address harmful content online: technical paper, 29 July 2021, https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content/technical-paper.html.

[4] UN Human Rights Council, Report of the independent international fact-finding mission on Myanmar, 18 September 2018, para. 74, https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/39/64.

[5] See, for example, Rory Cellan-Jones, "Tech Tent: Did social media inspire Congress riot?", BBC News, 8 January 2021, https://www.bbc.com/news/technology-55592752.

[6] Protection from Online Falsehoods and Manipulation Act 2019, No. 18 of 2019, section 7, https://sso.agc.gov.sg/Acts-Supp/18-2019

[7] Ley N. 1042 (Ley Especial de Ciberdelitos), 27 October 2020, Article 30, http://legislacion.asamblea.gob.ni/normaweb.nsf/($All)/803E7C7FBCF44D7706258611007C6D87

[8] Hate Speech and Disinformation Prevention and Suppression Proclamation *No. 1185/2020*, Articles 5, 7, https://www.article19.org/wp-content/uploads/2021/01/Hate-Speech-and-Disinformation-Prevention-and-Suppression-Proclamation.pdf.

[9] As explained below, the technical paper requires social media companies to make decisions about whether content is harmful content, but does not impose fines if those decisions are incorrect.

bodies that would be set up by the Act.[10] However, the proposal also contains several highly problematic features that should be overhauled, such as a vague scope of applicability that appears to include some private communications;[11] an unjustifiable 24-hour deadline for social media companies to take measures against harmful content;[12] an ill-defined obligation for social media companies to monitor and takedown harmful content proactively;[13] and an obligation for social media companies to proactively report content to law enforcement bodies, which requires them to make determinations about whether content that is hosted on their platforms is evidence of the commission of a crime.[14] Other aspects of the proposal are not inherently unacceptable but should be tweaked further, such as the definitions of hate speech[15] and terrorist content[16] and the approach to website blocking.[17]

This Submission assesses the proposal from the perspective of international human rights standards, although the Canadian constitutional framework and some laws and jurisprudence are briefly referenced. The Submission starts by laying out the key applicable international legal framework and the relevant human rights engaged by the proposal. Next, the Submission assesses the proposed definitions of harmful content, suggesting tweaks to the definitions for terrorist content and hate speech. The Submission then examines the scope of the proposal, arguing that it should be adjusted to exclude all private communications from its ambit. At this point, the Submission briefly reviews the five new regulatory bodies relevant to the proposal, finding that their independence from government through the Governor-in-Council (GIC) appointments process is key to the successful functioning of the proposal.

The Submission then discusses three additional substantive issues, starting with the user-flagged content moderation system, finding that it is largely in line with international standards, with the glaring exception of the 24-hour requirement to address content, which will result in a high rate of erroneous decisions at first instance and worsen an already excessive caseload for the Digital Recourse Council of Canada (Digital Recourse Council or Council), the new content moderation body. The Submission then examines the proposed obligations for online communication service providers (OCSPs) to proactively monitor content for removal and reporting to law enforcement, arguing that these obligations should be removed as they undermine privacy and over-incentivise content removal and reporting. The Submission concludes by recommending that further safeguards be built into the proposed system of website blocking.

---

[10] Technical paper, ss. 36-38 and 46-48.
[11] Technical paper, ss. 2-3.
[12] Technical paper, s. 11.
[13] Technical paper, s. 10.
[14] Technical paper, ss. 20, 22.
[15] Technical paper, s. 8.
[16] Technical paper, s. 8.
[17] Technical paper, ss. 120-123.

2. Applicable Legal Framework and Relevant Human Rights

2.1. Freedom of Expression

Article 19 of the International Covenant on Civil and Political Rights (ICCPR),[18] ratified by Canada in 1976, is the primary source of international human rights law's protection for freedom of expression. Article 19(2) of the ICCPR provides: "Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds…" It is plain that this right will be substantially engaged in any attempt to regulate online content; not just for the impact on would-be expressers of online content but also on the many users who have a right to receive that information.

The right to freedom of expression is not absolute under international human rights law. Restrictions of that right must pass the three-part test outlined in Article 19(3) of the ICCPR. Any restriction must by "provided by law", which means that it must be authorised by a validly passed law and be formulated with sufficient precision to enable individuals who are subject to it to regulate their conduct accordingly.[19] Second, the restriction must seek to protect at least one of the legitimate interests listed in Article 19(3): public order, public health, public morals, national security or the rights and reputations of others. Third, the restriction must be necessary to protect that interest which, among other things, includes an element of proportionality.[20]

The content of the protection in Article 19 of the ICCPR has been further developed and fleshed out in standards issued by a variety of international authorities, such as General Comment 34 of the UN Human Rights Committee, the treaty monitoring body for the ICCPR, and the Joint Declarations and thematic reports of the special international mandates on freedom of expression from the UN, Organization for Security and Co-operation in Europe (OSCE), African Union and the Organisation of American States (OAS).[21] This analysis draws on all of these documents.

While this Submission focuses on international human rights law, it is worth noting that freedom of expression is domestically protected in section 2(b) of Canada's constitutional Canadian Charter of Rights and Freedoms (Charter).[22] As a preliminary note, the above-stated international law test for protecting freedom of expression forms a floor for the Charter's domestic protection, as stated by Dickson C.J. in *Re: Public Service Employee Relations Act (Alta.)*:

---

[18] UN General Assembly Resolution 2200A (XXI), 16 December 1966, in force 23 March 1976.
[19] UN Human Rights Committee, General Comment No. 34, Article 19: Freedoms of opinion and expression, 12 September 2011, para. 25, https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf.
[20] *Ibid.*, para 34.
[21] For a full list of the Joint Declarations, see: https://www.osce.org/fom/66176.
[22] Part 1 of the Constitution Act, 1982, being Schedule B to the Canada Act 1982 (UK), 1982, c 11.

> The content of Canada's international human rights obligations is, in my view, an important indicia of the meaning of "the full benefit of the *Charter*'s protection." I believe that the *Charter* should generally be presumed to provide protection at least as great as that afforded by similar provisions in international human rights documents which Canada has ratified.[23]

The section 2(b) Charter analysis involves two steps. The first step is ascertaining whether there is a *prima facie* breach of freedom of expression.[24] The question then shifts to whether the *prima facie* breach can be justified under s. 1 of the Charter, which provides that the rights guaranteed by it may be subject to "reasonable limits prescribed by law as can be demonstrably justified in a free and democratic society." This implications of this were elaborated in some detail in the well-known case of *R. v. Oakes*.[25]

2.2. Privacy

The right to privacy is protected in Article 17 of the ICCPR, which provides: "No one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence, nor to unlawful attacks on his honour and reputation." Interferences with this right are only permitted where "authorized by domestic law that is accessible and precise and that conforms to the requirements of the Covenant", is in pursuit of "a legitimate aim" and "meet[s] the tests of necessity and proportionality."[26] Privacy and freedom of expression are interlinked,[27] since privacy "may empower individuals to circumvent barriers and access information and ideas without the intrusion of authorities" and "be the only way in which many can explore basic aspects of identity, such as one's gender, religion, ethnicity, national origin or sexuality."[28]

As with freedom of expression, the contents of the right to privacy have been further fleshed out in a variety of international statements, including the UN Human Rights Committee's General Comment 16,[29] and the thematic reports of the UN special mandates on the right to privacy and other privacy-relevant mandates such as the promotion and protection of human rights and fundamental freedoms while countering terrorism.

---

[23] [1987] 1 SCR 313, para. 59.
[24] See, for example, *Canadian Broadcasting Corp. v. Canada (Attorney General)*, 2011 SCC 2; and *Montréal (City) v. 2952-1366 Québec Inc.*, [2005] 3 S.C.R. 141.
[25] [1986] 1 S.C.R. 103.
[26] Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, 23 September 2014, para. 30, https://digitallibrary.un.org/record/781159/files/A_69_397-EN.pdf.
[27] See Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 17 April 2013, para. 79, https://undocs.org/A/HRC/23/40.
[28] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 22 May 2015, para. 12, https://undocs.org/A/HRC/29/32.
[29] 8 April 1988, https://tbinternet.ohchr.org/Treaties/CCPR/Shared%20Documents/1_Global/INT_CCPR_GEC_6624_E.doc.

Domestically, privacy is protected in s. 8 of the *Charter*, which provides: "Everyone shall have the right to be secure against unreasonable search and seizure." This right has been interpreted by the Supreme Court of Canada to include a right to privacy (an unreasonable search being a breach of privacy) that extends online. Indeed, the Supreme Court of Canada has recognised that aspects of informational privacy are especially important in the context of the Internet.[30] The s. 8 analysis comprises of two steps. First, there should be an assessment whether there has been a search or a seizure, requiring an assessment of whether there is a reasonable expectation of privacy in relation to the subject matter of the search or seizure.[31] If so, the second step assesses whether that search or seizure was reasonable, which comprises an analysis of whether the search or seizure was prescribed by law, whether the law was reasonable and whether the manner of the search or seizure was reasonable.[32]

2.3. The Rights of Others

It is important to acknowledge that freedom of expression and privacy are not the only rights at play in this proposal. Social media regulation can affect the free expression and privacy of expressers of content, but various other rights – notably those of the victims of harmful content – are also relevant. Terrorist content, hate speech, incitement to violence and the spreading of child pornography or non-consensual intimate images can have severe impacts on many important human rights.

For instance, hate speech can implicate the rights of others to be free from discrimination, protected in Article 26 of the ICCPR and the subject of an entire UN human rights treaty, the International Convention on the Ending of All Forms of Racial Discrimination (ICERD).[33] International human rights law considers the prevention of hate speech to be so important that Article 20 of the ICCPR requires States to prohibit hate speech, one of the ICCPR's few positive obligations to restrict speech. Similarly, terrorist content and incitement to violence can implicate others' rights to life and to security of the person, as protected in Articles 6 and 9 of the ICCPR. Content that sexually exploits children and the non-consensual sharing of intimate images can implicate others' rights to be free from cruel, inhuman and degrading treatment under Article 7 of the ICCPR and affect dignity, which is not a standalone right under the ICCPR but is at the very core of the concept of human rights[34] and features in the preamble of most international human rights treaties.[35]

---

[30] *R. v. Spencer*, 2014 SCC 43, para. 41, https://scc-csc.lexum.com/scc-csc/scc-csc/en/item/14233/index.do.
[31] See, for example, *R. v. Spencer*, *ibid.*, generally.
[32] *Ibid.*
[33] 21 December 1965, United Nations, Treaty Series, vol. 660, p. 195.
[34] UN Office of the High Commissioner for Human Rights, Universal Declaration of Human Rights – in six cross-cutting themes, 1996 – 2021, https://www.ohchr.org/en/udhr/pages/crosscuttingthemes.aspx.
[35] Including the ICCPR and Universal Declaration of Human Rights (UDHR).

Protecting the rights of others is a legitimate interest which may justify a restriction on freedom of expression and privacy under Articles 19(3) and 17 of the ICCPR respectively.[36] The key question that this Submission addresses is whether all of the proposal's restrictions on free expression and privacy are necessary and proportionate to the protection of legitimate interests, such as the rights of others or public safety.

3. The Definitions of Harmful Content

The proposal seeks to target five categories of online "harmful content", each of which largely tracks five categories of content that are already illegal under Canadian law: child sexual exploitation, terrorist content, incitement to violence, hate speech and the non-consensual sharing of intimate images.[37] The Act's definitions "borrow from the Criminal Code[38] but are "adapted to the regulatory context".[39] The only specific example that the technical paper provides of how this adaptation might look pertains to child sexual exploitation. The Act would cover material related to child sexual exploitation that may not constitute a criminal offence but still be harmful to children and victims when posted on an online communication service (OCS), such as screen shots of child porn videos that do not include the criminal activity but "refer to it obliquely".[40]

The five categories of harmful content must be sufficiently narrowly defined to pass the ICCPR's Article 19(3) test that limitations be "provided by law" and the Charter's s. 1 test that limits be "prescribed by law", both of which prohibit restrictions which are unduly vague. Most of the definitions are indeed sufficiently precise, with the exception of "terrorist content". The definition of "hate speech" is sufficiently precise, but given its notoriously subjective nature, its definition should specifically mention international standards on hate speech, such as the Rabat Plan of Action,[41] to guide OCSPs' content moderators and the Digital Recourse Council.

3.1. Terrorist Content

The technical paper does not provide a precise definition for "terrorist content", only explaining that such content is the kind which "actively encourages terrorism and which is likely to result in terrorism."[42] This definition is too flexible, leaving room to argue, for

[36] See note 27, para. 28.
[37] Technical paper, s. 8.
[38] R.S.C., 1985, c. C-46.
[39] Technical paper, s. 8.
[40] Technical paper, s. 8.
[41] UN General Assembly, Annual report of the United Nations High Commissioner for Human Rights: Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, pp. 6 – 15, 11 January 2013, https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf.
[42] Technical paper, s. 8.

instance, that certain political or religious opinions may not advocate for violence but encourage terrorism because they are uttered in fervent support of ideologies that have been associated with or co-opted by terrorists. Accordingly, the special international mandates on freedom of expression have explained that "Criminal responsibility for expression relating to terrorism should be limited to those who incite others to terrorism; vague concepts such as glorifying', 'justifying' or 'encouraging' terrorism should not be used."[43] While the proposal's prohibition of terrorist content pertains to regulatory rather than criminal responsibility, a more precise definition is still needed for the definition to pass the "provided by law" requirement of the ICCPR.

The technical paper does state that all the definitions of harmful content will be based on corresponding Criminal Code offences. However, it does not state precisely which terrorism-related offence in the Criminal Code the "terrorist content" restriction will be based on. Part II.i of the Criminal Code contains numerous terrorist-related offences, such as counselling[44] or facilitating terrorism,[45] although "terrorist content" is not defined. Terrorist content should be restricted to content which incites terrorist activities, with "terrorist activities" following the definition in s. 83.01(1)(b) of the Criminal Code. That definition requires an intention to cause serious bodily harm or death to a person by violence or to cause a serious risk of harm to the health and safety of the public, bringing it largely in line with the model definition of terrorism adopted by the UN High Level Panel on Threats, Challenges and Change.[46] To make sure that legitimate political, religious or ideological speech is not caught by the definition, the definition of "terrorist content" should also contain the following safeguard, adapted from s. 83.01(1.1) of the Criminal Code on terrorist activities:

> For greater certainty, the mere expression of a political, religious or ideological thought, belief or opinion does not constitute terrorist content.

3.2. Hate Speech

The technical paper defines hate speech in accordance with the definition in Bill C-36, which proposes amendments to the Canadian Human Rights Act.[47] The technical paper also states that the definition of hate speech must be in line with the jurisprudence of the

[43] Special international mandates on freedom of expression at the UN, OSCE, OAS and ACHPR, Joint Declaration on freedom of expression and responses to conflict situations, 4 May 2015, s. 3(b), http://www.law-democracy.org/live/wp-content/uploads/2015/05/JD-2015.final_.Eng_.pdf.
[44] Criminal Code, s. 83.221.
[45] Criminal Code, s. 83.19.
[46] Report of the High-level Panel on Threats, Challenges and Change: A more secure world: our shared responsibility, 2 December 2004, para. 164(d), https://undocs.org/A/59/565.
[47] R.S.C., 1985, c. H-6, https://laws-lois.justice.gc.ca/PDF/H-6.pdf; An Act to amend the Criminal Code and the Canadian Human Rights Act and to make related amendments to another Act (hate propaganda, hate crimes and hate speech), Second Session, 43rd Parliament, First Reading, 23 June 2021, s. 13, https://parl.ca/Content/Bills/432/Government/C-36/C-36_1/C-36_1.PDF.

Supreme Court of Canada, which has held that only the most extreme forms of speech would qualify for this title.[48] However, it would also be useful for the proposal, when made into law, to refer to some of the leading international standards on this issue. For instance, the Rabat Plan of Action,[49] the product of a series of UN-led expert consultations on hate speech, provides a useful six-part test for ascertaining when speech rises to the level of hate speech, focusing on the context, speaker, intent, content and form, likelihood of harm and imminence. The law should refer directly to the Rabat Plan of Action and other international instruments on freedom of expression and hate speech.

4. Who and What Will be Regulated

The main targets of the regulation would be what the proposal calls OCSs and the providers of those services (OCSPs). The proposal would define an OCS as "a service that is accessible to persons in Canada, the primary purpose of which is to enable users of the service to communicate with other users of the service, over the internet"[50] and should exclude services that only enable private communication.[51] This definition would catch all social media services with a public-facing element, such as Facebook, Youtube or Twitter. However, it is unclear whether this proposed definition would cover all of the services of dual-function OCSPs which provide both public-facing and private communications, such as the direct messaging systems of Instagram and Facebook Messenger. This is a reasonable interpretation since the proposed definition would only exclude services that exclusively enable private communications.

It is also unclear how the GIC, in consultation with the Digital Safety Commissioner of Canada (Digital Safety Commissioner, one of the newly created regulatory bodies), will define a "private communication" through regulation.[52] For instance, while services such as Whatsapp and Signal are generally understood to be for wholly private communications, they do allow for one-to-many communications through chat groups or message forwarding. The definition of "private communication" should thus be carefully tailored to ensure that the Act considers such services to be wholly private and thus exempted from regulation.

In any case, the proposal would empower the GIC, in consultation with the Digital Safety Commissioner, to use regulations to extend the Act's applicability to certain services that do not meet the definition of an OCS if the GIC "is satisfied that there is a significant risk that harmful content is being communicated on the category of services or that specifying the

[48] *Saskatchewan (Human Rights Commission) v. Whatcott*, 2013 SCC 11, paras. 57 and 116, https://scc-csc.lexum.com/scc-csc/scc-csc/en/item/12876/index.do.
[49] See note 41.
[50] Technical paper, s. 2.
[51] Technical paper, s. 2.
[52] Technical paper, s. 3(c).

category of services would further the objectives of this Act".[53] This would allow the GIC to expand coverage of the proposed Act to any other service, including private communication services, such as Whatsapp or Signal.

The proposal would oblige OCSPs to create a user-flagging content moderation system for harmful content[54] and to proactively monitor and make the five categories of harmful content inaccessible, including the use of automated systems (these obligations are detailed in greater depth in sections 6 and 7 of this Submission respectively).[55]

Subjecting private messaging services to these obligations would open a Pandora's box of privacy issues. For instance, the obligation on OCSPs to proactively monitor and report content to law enforcement is already problematic for freedom of expression when applied to public content, as explained below. However, if this obligation is applied to private messaging content, it would require OCSPs to monitor the private communications of all their users to identify harmful content and report some of that content to law enforcement. That would essentially be a form of mass surveillance that would be a flagrant violation of the privacy rights of millions of Canadian social media users and a gross breach of Canada's obligations to refrain from arbitrary interferences with privacy under Article 17 of the ICCPR. As stated by the UN Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism:

> The hard truth is that the use of mass surveillance technology effectively does away with the right to privacy of communications on the Internet altogether. By permitting bulk access to all digital communications traffic, this technology eradicates the possibility of any individualized proportionality analysis. It permits intrusion on private communications without independent (or any) prior authorization based on suspicion directed at a particular individual or organization.[56]

Including private messaging content would also create serious feasibility issues. Currently, the major social networks are already struggling with the immense burden of monitoring and moderating public content.[57] Including all of the private messaging on their platforms would quite clearly be beyond the capacity of OCSPs and would likely massively overload the user-flagging system and the caseload of the Digital Recourse Council. The Act should contain language that clearly excludes all forms of private messaging from its ambit in all circumstances.

---

[53] Technical paper, s. 3.
[54] Technical paper, s. 11.
[55] Technical paper, s. 10.
[56] Report of the UN Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, 23 September 2014, para. 12, http://s3.documentcloud.org/documents/1312939/un-report-on-human-rights-and-terrorism.pdf.
[57] See, for example, John Koetsier, "Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day", Forbes, 9 June 2020, https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=619f3dee54d0.

5. Five New Regulatory Bodies

The Act creates four new regulatory bodies to fulfil its aims: the Digital Safety Commissioner, the Digital Safety Commission, the Digital Recourse Council and the Advisory Board. A fifth relevant body, the Personal Information and Data Protection Tribunal, would be responsible for oversight of the penalties recommended by the Digital Safety Commissioner but is proposed in another bill, Bill C-11, which is currently before the House of Commons.[58]

The independence of the three bodies that have enforcement or adjudicatory functions – the Digital Safety Commissioner, Digital Recourse Council and the Personal Information and Data Protection Tribunal – is key to the appropriate functioning of the Act. For instance, the process for resolving appeals from content moderation systems will only meet international standards if the body deciding the appeals is free from political and commercial influence.[59] It is not so crucially important that the other two entities – the Advisory Board and the Digital Safety Commission – be independent of government, as the former merely serves a high-level advisory role[60] while the latter plays a supporting role for the other three bodies created by the Act,[61] although independence for both is still still highly advisable.

All three of the proposed adjudicatory and enforcement bodies appear to be sufficiently insulated from political or commercial influence. In terms of freedom from political influence, members of all three bodies are selected through Canada's GIC appointments process;[62] while the GIC is a political body, candidates must undergo a rigorous vetting process that ensures that they are ultimately chosen on the basis of merit and that adequate consideration is also given to diversity.[63] A specific subcategory of GIC appointees exists to further ensure independence from government, the GCQ category, for whom salary components cannot be determined by the GIC (unlike other federal government positions, which have a variable performance-based component that is determined by the GIC).[64] The members of the three adjudicatory and enforcement bodies should be GCQ positions, much

---

[58] An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make consequential and related amendments to other Acts, Second Session, 43rd Parliament, First Reading, 17 November 2020, https://parl.ca/Content/Bills/432/Government/C-11/C-11_1/C-11_1.PDF.

[59] Special international mandates on freedom of expression at the UN, OSCE, OAS and ACHPR, Joint Declaration on media independence and diversity in the digital age, 2 May 2018, s. 1(b)(v), https://www.law-democracy.org/live/wp-content/uploads/2018/12/mandates.decl_.2018.media-ind.pdf.

[60] Technical paper, s. 75.

[61] Technical paper, s. 60.

[62] Technical paper, ss. 36-37 and 46-47. C-11, Part 2, s. 6.

[63] Government of Canada, Governor-in-Council appointments, 1 Feb 2021, https://www.canada.ca/en/privy-council/programs/appointments/governor-council-appointments/general-information/appointments.html.

[64] Government of Canada, Terms and conditions applying to Governor in Council appointees, 1 April 2018, https://www.canada.ca/en/privy-council/programs/appointments/governor-council-appointments/compensation-terms-conditions-employment/terms-conditions.html.

like the members of Canada's other independent oversight bodies, such as the Canadian Human Rights Commission or the National Parole Board.[65]

In terms of freedom from commercial influence, members of the Digital Safety Commissioner and the Digital Recourse Council cannot be shareholders in an OCS or OCSP and members of all three adjudicatory or enforcement bodies must declare any conflicts of interest they have with regard to matters under their purview.[66]

6. The User-flagged Content Moderation System

6.1. The Proposed User-flagged Content Moderation System

The proposal would create a new, two-tier system of content flagging and moderation. The first tier is handled by OCSPs, which must create systems that enable their users to flag easily content which they believe falls within the scope of one of the five categories of harmful content.[67] Users who believe that these systems are inadequate may complain to the Digital Safety Commissioner, who may investigate and adjudicate that complaint.[68]

Once a user has flagged content, the OCSP must arrive at a decision within 24 hours on whether the content is harmful, although the GIC can prescribe a different timeline for some categories of content.[69] If the OCSP decides that the content is harmful, it must render it inaccessible to Canadian users; if it decides otherwise; the content may stay up.[70] In all cases, the OCSP must provide notice of its decision to the flagger and the author of the content.[71] The OCSP must also provide both parties with an option for an internal appeal to the OCSP (the Act's name for this process is "reconsideration").[72]

The second tier of the content moderation system is handled by the Digital Recourse Council. Either party to a content moderation dispute may file a complaint to the Digital Recourse Council if they are dissatisfied with the results of the internal appeal.[73] The Digital Recourse Council has the power to dismiss complaints that are "trivial, frivolous, vexatious, made in bad faith or on other grounds".[74] Once a complaint is filed, both parties must receive notice of the complaint and have the opportunity to make representations, which may include a hearing if the Digital Recourse Council considers that to be in the public

[65] *Ibid.*
[66] Technical paper, ss. 38 and 48; C-11, Part 2, s. 12.
[67] Technical paper, s. 12(a).
[68] Technical paper, ss. 12, 40-44.
[69] Technical paper, s. 11(a).
[70] Technical paper, s. 11(b).
[71] Technical paper, s. 12(b).
[72] Technical paper, s. 12(c).
[73] Technical paper, s. 49.
[74] Technical paper, s. 52.

interest.[75] These hearings can be private if the Digital Recourse Council and the Digital Safety Commissioner determine that "a public hearing would not be in the public interest, including where there is a privacy interest, national security interest, international relations interest, national defence interest, or confidential commercial interest."[76]

If the Digital Recourse Council finds that the content does not fall within one of the five categories of harmful content, it communicates its decision to all parties; the OCSP may then leave the content up or still decide to make the content inaccessible in accordance with its internal guidelines.[77] If the Digital Recourse Council finds that the content does fall within one of the five categories of harmful content, it communicates its decision to all parties and orders the OCSP to make the content inaccessible in Canada, if the OCSP has not already done so.[78] This order is to be shared with the Digital Safety Commissioner, who is to monitor the OCSP's compliance with the inaccessibility order.[79]

A failure to comply with a Digital Recourse Council's inaccessibility order is one of the bases for levying administrative fines of up to 3% of an OCSP's gross global revenue or up to ten million Canadian dollars, whichever is higher.[80] Failing to comply with an inaccessibility order could also constitute a criminal offence under the Act which can incur fines of 4-5% of an OCSP's gross global revenue or 20-25 million Canadian dollars,[81] although the proposal does not make it clear what threshold distinguishes the administrative penalty from the criminal penalty. To be clear, these fines are incurred if an OCSP defies an inaccessibility order issued by the Digital Recourse Council; the technical paper does not contemplate penalties where OCSPs issue initial content moderation decisions which are later overturned by the Digital Recourse Council.

6.2. Assessment of the User-Flagged Content Moderation System

The proposed user-flagged system of content moderation contains several useful protections for freedom of expression. Crucially, the system does not link the OCSPs' initial content moderation decision to liability, even if that decision is later reversed on appeal, thereby removing any incentive to be over-inclusive when removing content. Both the author of the flagged content and the flagging user have equal appeal rights,[82] which ensures procedural fairness for users and also likely reduces the likelihood of OCSP bias towards either content removals or takedowns for purposes of avoiding downstream engagement in the process. Notice to all concerned parties must be issued at every major

[75] Technical paper, s. 53.
[76] Technical paper, s. 59.
[77] Technical paper, s. 54.
[78] Technical paper, s. 55.
[79] Technical paper, s. 56.
[80] Technical paper, s. 108.
[81] Technical paper, s. 119.
[82] Technical paper, s. 49.

step of the decision-making process,[83] and appeals from content moderation decisions will be handled by an independent regulator, the Digital Recourse Council, with further recourse to judicial review by the Federal Court of Canada.[84]

However, one glaring problem in the user-flagging system is the 24-hour deadline to reach a decision in respect of all five types of harmful content. This ignores key differences between types of content and the relative urgency with which they need to be addressed. Overall, it is almost certain to result in poorer quality decisions across the board than if more time was allocated for this. While the salutary features of the system mentioned in the previous paragraph mean that decisions will not necessarily be poorer in a certain direction – for example, in the direction of over-removal – they will nonetheless be wrong more often.

The technical paper does not offer any justification for why the 24-hour deadline is necessary for any – let alone all five – of the categories of harmful content. It is true that in the case of child sexual exploitation, where the content is relatively easy to identify and normally easy at least to distinguish from political or other forms of public interest speech, and where its ongoing dissemination is especially harmful, a 24-hour deadline may be justifiable. In the case of non-consensual sharing of intimate images, a 24-hour deadline may also be justifiable given the extreme harm that can result from the dissemination of those images. It may be difficult to distinguish rapidly between consensual and non-consensual sharing of intimate images, since this assessment requires some analysis of context. However, intimate images, even if consensual, will rarely constitute public interest speech and are in any case already prohibited by the content standards of major social media companies.[85] The significant benefits of legally mandating the rapid removal of non-consensual intimate images may therefore outweigh the costs of doing so.

However, in the cases of hate speech, terrorist content and incitement to violence, the dividing line will often be hard to draw, making a one-day deadline far too short for a content moderator to arrive at a considered decision. As suggested above, hate speech and terrorist content can be difficult to distinguish from political or religious speech. Incitement to violence is sometimes clear-cut but may also involve complex assessments of nuance. One prominent example is the Facebook Oversight Board's decision to uphold Facebook's ban on US President Trump for alleged incitement to violence over the 6 January 2021 Capitol riot. It took five months for the oversight body to arrive at a decision, which was not unanimous, illustrating the highly subjective and complex nature of labelling speech as

---

[83] Technical paper, ss. 51, 52,

[84] Federal Courts Act, R.S.C., 1985, c. F-7, s. 18(1), https://laws-lois.justice.gc.ca/PDF/F-7.pdf.

[85] Facebook, Facebook Community Standards: Adult Nudity and Sexual Activity, 2021, https://transparency.fb.com/policies/community-standards/adult-nudity-sexual-activity/; Twitter, Sensitive media policy, November 2019, https://help.twitter.com/en/rules-and-policies/media-policy; and Tiktok, Community guidelines: Adult nudity and sexual activities, December 2020, https://www.tiktok.com/community-guidelines?lang=en#30.

incitement to violence.[86] Content moderators cannot consistently make high-quality decisions on these matters within 24 hours. Line content moderators may also wish to escalate especially difficult decisions to superiors with better training, which can also take more than 24 hours.

The result will likely be a high proportion of content moderation decisions that are decided incorrectly. It is not clear at this point whether these decisions would, in aggregate, tend towards the over-removal of legitimate content or over-maintenance of harmful content. Since the proposed Act does not levy penalties on OCSPs for making content moderation decisions that are later overturned by the Digital Recourse Council, there are no obvious incentives towards over-removal, although there may well be more subtle ones, such as a tendency to respond to the creaky wheel, i.e. the user who complains about content. Authors raising sensitive issues may also be reluctant to defend their content vigorously, which could again result in a bias within the system. In any case, poor decisions will still harm freedom of expression, since at least a percentage will involve the inappropriate removal of content at the first instance which could only be restored on appeal or perhaps never if no appeal is forthcoming. On the flip side, hurried decisions that lead to harmful content erroneously being left up also defeat the Act's primary purpose of removing harmful content online.

Another negative implication of the 24-hour limit is that more social media users are likely to lose trust in social media companies' content moderation processes. This may be the case whether the user is the author of political content that has been mistaken for terrorist content or a person of colour who has identified racist hate speech that has been left up by a content moderator forced to make a hasty decision. And rushed first-level decisions will almost certainly lead to far more appeals being lodged with the OCSP's internal appeal process, and potentially with the Digital Recourse Council, especially as users observe a reasonably high rate of initial decisions getting overturned. This, in turn, will increase what can reasonably be expected to be a fairly massive caseload placed on these bodies.

A reasonably obvious solution is to give OCSPs more time to address content that has been flagged as hate speech, incitement to violence or terrorist content, while child sexual exploitation material and non-consensually shared intimate images could still be addressed within 24 hours. A 72-hour initial deadline that can be extended through a simple procedure for a further week if necessary should provide OCSPs with enough time to make considered decisions about content.

6.3. Caseload of the Digital Recourse Council of Canada

[86] Facebook Oversight Board, Case decision 2021-001-FB-FBR, 5 May 2021, https://www.oversightboard.com/sr/decision/2021/001/pdf-english.

The Digital Recourse Council comprises three to five members[87] and is required to review all appeals from OCSPs' initial moderation decisions, with the only limitation being the Council's power to dismiss complaints that are "frivolous, vexatious, trivial, made in bad faith or on other grounds".[88] The GIC may introduce other grounds for dismissing complaints[89] but, otherwise, the Council is required to adjudicate all complaints that have merit. This may be contrasted with the Facebook Oversight Board, which only adjudicates the few cases that are "difficult, significant and globally relevant".[90]

It is not possible to predict with any certainty the volume of complaints that the Council will receive but statistics from other systems give some insight into this. For example, in the 2nd quarter of 2021, internally appeals were lodged against about 1,400,000 of Facebook's initial content moderation decisions just regarding hate speech worldwide.[91] Scaling down for Canada's population,[92] that roughly translates into about 77 internal appeals for every day of the year. If just 25% of those internal appeals were subject to complaints before the Council (a potentially conservative estimate, given that this costs nothing), that would be 19 complaints every day just in relation to hate speech, and arising from Facebook alone. Of course scaling of this sort is notoriously unreliable but it seems reasonable to assume that the number of complaints to the Council would be enormous.

This has important implications in terms of the operations and budgets of the Council and the Digital Safety Commission that supports it. We note that it is imperative that it be able to process complaints rapidly since, otherwise, wrong decisions – whether to allow harmful content to remain online or to block access to legitimate content – will remain in place, undermining the credibility of the system and harming freedom of expression. In terms of operations, the Council will need to have staff processing complaints, whether or not these are ultimately signed off on by the Council's members, where these staff are housed in the Council itself or in the supporting Digital Safety Commission. To process a large volume of complaints rapidly, that staffing complement would need to be significant. And to process complaints properly, the staff will need to be very professional. All of which suggests that the government should be prepared to commit significant resources to sustain the operations of the Council.

7. OCSPs' Proactive Obligations

[87] Technical paper, s. 46.
[88] Technical paper, s. 52.
[89] Technical paper, s. 52.
[90] Facebook Oversight Board, Appealing Content Decisions on Facebook or Instagram, https://oversightboard.com/appeals-process/.
[91] Facebook, Community Standards Enforcement Report – Hate Speech, August 2021, https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/.
[92] Worldometer, Canada Population, 21 September 2021, https://www.worldometers.info/world-population/canada-population/.

7.1. OCSPs to Take "All Reasonable Measures" to Identify and Make Harmful Content Inaccessible

Alongside the user-flagging content moderation system, the Act also obliges OCSPs to "take all reasonable measures" to identify harmful content on their platforms and to make that content inaccessible in Canada.[93] The technical paper does not provide specifics on what these measures would entail, although it explicitly contemplates the use of automated systems.[94] The Digital Safety Commissioner may prescribe rules in this area by regulation, which presumably covers both what would constitute "all reasonable measures" and how to make content inaccessible.[95] These measures cannot result in discrimination, as described in Canada's anti-discrimination legislation, the Canadian Human Rights Act, but no other constraints are set out here.

There are serious problems with placing legal obligations on OCSPs to monitor content on their systems. The volume of material most OCSPs host means that monitoring can only really be done with automated tools, since there is too much material for humans to monitor, at least for the first pass. Using such tools to identify content that involves child sexual exploitation is relatively less sensitive, since such content is generally easier to identify and more difficult to mistake for public interest content. But using automation to identify hate speech, terrorist content, incitement to violence and the non-consensual sharing of intimate images is highly problematical given that such tools remain relatively crude.[96] For instance, Facebook, Youtube and Twitter's automated systems have on multiple occasions taken down evidence of mass atrocities and war crimes by misidentifying it as terrorist content or incitement to violence.[97] This is why international standards make it clear that it is not legitimate to place a positive obligation on OCSPs to monitor for illegal content.[98] In this context, the technical paper's specific reference to automated systems is especially troubling, including insofar as it would allow the Digital Safety Commissioner to direct OCSPs specifically to rely on automation for content identification and perhaps even removal.

---

[93] Technical paper, s. 10.

[94] *Ibid*.

[95] *Ibid*.

[96] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 6 April 2018, para. 29, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf?OpenElement.

[97] Human Rights Watch, "Video Unavailable", Social Media Platforms Remove Evidence of War Crimes, 10 September 2020, https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes.

[98] Special international mandates on freedom of expression at the UN, OSCE, OAS and ACHPR, Joint Declaration on Freedom of Expression and the Internet, 1 June 2011, s. 2(b), http://www.law-democracy.org/wp-content/uploads/2010/07/11.06.Joint-Declaration.Internet.pdf; and Council of Europe, CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries, 7 March 2018, s. 1.3.5, https://rm.coe.int/1680790e14.

The allocation of sweeping power to the Digital Safety Commissioner to regulate how OCSPs must proactively identify and remove content is also a violation of Article 19 of the ICCPR, which requires restrictions on freedom of expression to be both provided by law and narrowly tailored. As stated by the UN Human Rights Committee, "A law may not confer unfettered discretion for the restriction of freedom of expression on those charged with its execution."[99] The only constraints that are placed on the exercise of these powers are that they must be "reasonable" and cannot fall foul of non-discrimination protections,[100] which signally fail to meet the standard set out by the Human Rights Committee.

Instead of allocating broad authorisation to the use of automated content identification and removal systems, the proposal should establish guardrails around the use of such systems. Such guardrails might, for example, require initially flagged content to be reviewed by a human being, providing an opportunity to fix automated errors. Automated systems which entail the use of content filtering – the pre-emptive blocking of content triggered by certain metrics, such as keywords – are not legitimate. Such systems, when not controlled by the end-user, have been condemned by the international special mandates as being "a form of prior censorship and not justifiable as a restriction on freedom of expression."[101]

The technical paper is not as clear as it should be regarding the right of users to appeal against measures taken by OCSPs against their content as described above. This seems to be provided for,[102] but any legislation should clarify this in favour of equal rights to appeal for all affected users, whether they are flaggers or authors. It is also important that provision be made for users whose content is affected by OCSP monitoring and proactive content moderation measures to be notified as soon as any decision is made, thereby enabling them to appeal.

7.2. OCSPs to Have New Reporting Obligations to Law Enforcement and Intelligence Services

OCSPs must also report regularly to the Digital Safety Commissioner on a broad range of data about their services in Canada, including the volume and type of harmful content on them, the volume and type of content they have moderated, the resources and personnel dedicated to content moderation and how they "monetize harmful content".[103] To fulfil these and other obligations, OCSPs must have adequate record management systems and practices in place.[104] These features should be kept within the proposal as they would result

[99] See note 19, para. 35.
[100] Technical paper, s. 10.
[101] Special international mandates on freedom of expression at the UN, OSCE, OAS and ACHPR, Joint Declaration on Freedom of Expression and the Internet, 1 June 2011, s. 3(b), http://www.law-democracy.org/wp-content/uploads/2010/07/11.06.Joint-Declaration.Internet.pdf.
[102] For example in s. 12(c).
[103] Technical paper, s. 14.
[104] Technical paper, s. 15.

in increased transparency about harmful content on OCSPs and measures taken to address it, a positive development.

The second type of reporting obligation is to law enforcement, and the technical paper indicates that the Government of Canada is contemplating two options here.[105] The first option would obligate OCSPs to notify the Royal Canadian Mounted Police (RCMP) if they have reasonable grounds to believe that content that falls within the five categories of harmful content, and is therefore likely criminal in nature, poses an imminent risk of serious harm to any person or property.[106] The second option would obligate OCSPs to report to the relevant law enforcement agency – which could be the RCMP, the Canadian Security and Intelligence Services (CSIS) or others – in respect of certain crimes (to be prescribed by the GIC through regulation) based on content that falls within the five categories of harmful content.[107] Thus, the notification option is only for the RCMP and for content that poses an imminent risk of harm, while the reporting option is for any relevant law enforcement agency and covers prescribed crimes covered by the five categories of harmful content. Both options are to be subject to regulated standards on timing, the type of information to be provided, the thresholds of severity that trigger notifications or reports, and the formats of notifications or reports.[108]

A key problem here is that OCSPs are required to monitor content in the first place, discussed above, absent which it is not clear how these notification or reporting requirements could be discharged.

A second issue is that these obligations essentially deputise OCSPs to make subjective determinations on law enforcement issues which they are not qualified to do and which are best left up to law enforcement bodies. These include, respectively, whether content poses an imminent risk or harm or represents criminal behaviour. In the exceptional case of the crime of spreading child sexual exploitation content, the benefits of reporting may outweigh the risks, especially given that, as noted earlier, identification is less controversial in this case. Otherwise, however, this sort of reporting obligation is not appropriate. We note that notification or reporting by OCSPs is not value neutral; rather, it will likely trigger a police investigation. As such, unreliable reporting, especially where it is over-inclusive, exposes users to unjustified interactions with law enforcement bodies, which is not legitimate.

The technical paper fails to make it clear what type of information would need to be included in notifications or reports, which is left up to future regulations.[109] We note that this should be limited to the content of the social media post and not include user

[105] Technical paper, s. 20.
[106] Technical paper, s. 20(a).
[107] Technical paper, s. 20(b).
[108] Technical paper, s. 20.
[109] Technical paper, s. 20.

information, such as name, email address, phone number or IP address. To require the provision of that sort of information to law enforcement officials without judicial authorisation would be a serious breach of the right to privacy under international human rights law. As the UN Human Rights Committee has stated: "[S]ubscriber information may be issued with a warrant only ."[110] The technical paper does specifically provide that OCSPs should preserve and retain basic subscriber information that is pertinent to their reporting obligations,[111] which may suggest that this information would only be releasable pursuant to a warrant, but this should be made crystal clear in the wording of the Act.

8. Website Blocking

If an OCSP persistently defies orders to block content relating to child sexual exploitation or terrorist content, and if all other enforcement mechanisms have been exhausted, the Digital Safety Commissioner may apply to the Federal Court of Canada to request that telecommunications service providers wholly or partially block access to the offending OCSP in Canada.[112]

Website blocking is an extreme measure that can only be justified in highly exceptional circumstances, as the special international mandates on freedom of expression have stated:

> Mandatory blocking of entire websites, IP addresses, ports, network protocols or types of uses (such as social networking) is an extreme measure – analogous to banning a newspaper or broadcaster – which can only be justified in accordance with international standards, for example where necessary to protect children against sexual abuse.[113]

That said, the proposed modalities for website blocking do contain safeguards. Blocking can only be exceptionally employed against sites that consistently defy orders issued by the Digital Recourse Council with respect to terrorist content or child sexual exploitation,[114] the latter highlighted above by the special mandates as one instance where website blocking could be justified. All other enforcement mechanisms must have been exhausted and the blocking orders can only be issued by the Federal Court of Canada upon an application by the Digital Safety Commissioner.[115] Furthermore, the technical paper states that the Act

---

[110] UN Human Rights Committee, Concluding Observations on the Fourth Periodic Report of the Republic of Korea, 3 December 2015, para. 43, http://docstore.ohchr.org/SelfServices/FilesHandler.ashx?enc=6QkG1d%2FPPRiCAqhKb7yhshdNp32UdW56DA%2FSBtN4MHy9iuSMtUiNSvrbV9%2BJuD7JMLvy0Ju%2FXKLNHlCvzsdHK1rJtIsosm9tfQBiOl2kvBgjNYQMFXBklPP6Cl8vcuw0.

[111] Technical paper, s. 23.

[112] Technical paper, s. 120.

[113] See note 101, s. 3(a).

[114] Technical paper, s. 120.

[115] *Ibid*.

should direct the Digital Safety Commissioner to ensure that the blocking orders it requests are proportionate, taking into account the risk of excessive blocking.[116]

Further tweaks are nonetheless needed to strengthen safeguards for freedom of expression. While the language about blocking orders needing to be proportionate is welcome, more specific safeguards are needed given the extreme nature of website blocking. Either the Act or its regulations should address the technical challenges of blocking individual pages of a website, rather than a whole OCSP.[117] Where it is technically impossible to tailor blocking orders, considerations of proportionality may require orders to err on the side of leaving websites up instead of blocking them. The law should also clearly require the Digital Safety Commissioner to maintain a public and up-to-date list of blocked sites.

Finally, for the website blocking regime to be legitimate, it is crucial that the problems with the vague definition of terrorist content, highlighted in s. 3.1 of this Submission, are addressed. If the Act fails to provide a clear definition of terrorist content or if that definition is not strictly restricted to content which incites terrorist activities, then it would enable the blocking of websites that host content which is controversial but not prohibitable under international law.

[116] Technical paper, s. 121.

[117] Internet Society, Internet Society Perspectives on Internet Content Blocking: An Overview, 24 March 2017, https://www.internetsociety.org/resources/doc/2017/internet-content-blocking/.

Recommendations

- The definition of "terrorist content" should be strictly limited to "content that incites terrorist activities" and the definition of terrorist activities should be linked to or mirror the definition in s. 83.01(1)(b) of the Criminal Code, including the safeguard in s. 83.01(1.1) of the Code.
- The definition of "hate speech" should incorporate by reference international human rights standards on hate speech and freedom of expression.
- The scope of the legislation should clearly exclude all private communications in all circumstances.
- Members of the Digital Safety Commissioner, Digital Recourse Council and the Personal Information and Data Protection Tribunal should be appointed as GCQ-level positions to promote their independence from government.
- The 24-hour requirement to address harmful content should only apply to content about child sexual exploitation and the non-consensual sharing of intimate images. For hate speech, incitement to violence and terrorist content, the deadline should be 72-hours and OCSPs should be able to extend that by an additional week in challenging cases.
- Significant resources should be allocated to the Digital Recourse Council so that it can handle its caseload in a timely manner.
- The legal obligations for OCSPs to proactively monitor and remove content should be removed, perhaps other than for child sexual exploitation content.
- The legal obligations for OCSPs to proactively report or notify law enforcement bodies about the content found on their platforms should be removed.
- The rules on blocking of websites should contain additional safeguards for freedom of expression and transparency, such as directing the Digital Safety Commissioner to maintain a public list of blocked sites and rules that require careful tailoring, within technical constraints, of any blocking measures to ensure that blocking is proportionate and, in particular, that innocent content is not blocked.