Government of Canada Consultation on the Proposed Approach to Address Harmful Content Online

Submission by

Professor Michael Geist

Canada Research Chair in Internet and E-commerce Law

University of Ottawa, Faculty of Law

Centre for Law, Technology and Society

September 2021

A. Overview

I am a law professor at the University of Ottawa where I hold the Canada Research Chair in Internet and E-commerce Law and serve as a member of the Centre for Law, Technology and Society. I focus on the intersection between law and technology with an emphasis on digital policies. I submit these comments in a personal capacity representing only my own views.

My submission raises serious concerns with the government's proposed approach. I raise many specific concerns, but there are eight general comments that need to be raised.

1. The proposed approach does not strike an appropriate balance between addressing online harms and safeguarding freedom of expression. Indeed, after a single perfunctory statement on the benefits of Online Communications Services (OCSs) which says little about the benefits of freedom of expression, the document does not include a single mention of the Charter of Rights and Freedoms or net neutrality. There is surely a need to address online harms, but doing so must be Charter compliant and consistent with Canadian values of freedom of expression. I believe the proposed approach fails to adequately account for the freedom of expression side of the ledger.

2. Rather than adopting a "made in Canada" approach consistent with Canadian values, the plan relies heavily on policy developments elsewhere. Yet the reality is that those models from countries such as France, Germany, and Australia have met with strong opposition and raised serious concerns of unintended consequences. Indeed, France's approach has been ruled unconstitutional, Germany's model has resulted in over-broad removal of lawful content and a lack of due process, and Australia's framework is entirely unproven. An evidence-based approach would better account for these experiences rather than seek to mirror them.

3. The proposed approach mistakenly treats a series of harms - spreading hateful content, propaganda, violence, sexual exploitation of children, and non-consensual distribution of intimate images – as equivalent and requiring the same legislative and regulatory response. While there is a commonality between these harms as so-called "illegal speech", there are also significant differences. For example, it makes no sense to treat online hate as the equivalent of child pornography. By prescribing the same approach for all these forms of content, the efficacy of the policy is called into question.

4. There are lingering concerns about scope-creep with this proposal. Government officials have previously referenced the need to address "harmful" or "hurtful" comments, raising the prospect of expanding the model far beyond the current five forms of illegal speech cited in the proposal. Moreover, the government has indicated that these rules apply only to OCSs, identifying Facebook, Youtube, TikTok, Instagram, and Twitter as examples. It notes that there will be an exception for private communications and telecommunications such as wireless companies, Skype and WhatsApp (along with products and services such as TripAdvisor that are not OCSs). Yet during a briefing with stakeholders, officials were asked why the law shouldn't be extended to private communications on platforms as well, noting that these harms may occur on private messaging. Given that the government previously provided assurances of the exclusion of user generated content in Bill C-10

only to backtrack and make it subject to CRTC regulation, there is a need for renewed assurances about the scope of the rules.

5. The proposed approach envisions a massive new bureaucratic super-structure to oversee online harms and Internet based services. Due process concerns dictate that there be a suitable administrative structure to address these issues. However, some of the proposed models are ill-conceived that will not scale well nor afford the much-needed due process. For example, adjudicating over potentially tens of thousands of content cases is unworkable and would require massive resources with real questions about the appropriate oversight. Similarly, the powers associated with investigations are enormously problematic with serious implications for freedom of the press and freedom of expression.

6. The proposed approach threatens Canada's important role as a model for the rest of the world. Some of the proposals risk being deployed by autocratic countries to suppress freedom of expression with Canada cited as an example for why such measures are reasonable. The government should be asking a simple question with respect to many of its proposals: would Canadians be comfortable with the same measures being implemented countries such as China, Saudi Arabia, or Iran. If the answer is no (as I argue it should be), the government should think twice before risking its reputation as a leader in freedom of expression.

7. The proposed approach also threatens to harm the very groups it purports to protect. Without full due process and with clear incentives to remove content, there are real fears that the rules will be used to target BIPOC communities and vulnerable groups. Those groups could be silenced by a process that is weaponized by purveyors of hate with their voices removed due to poorly conceived rules that do not feature adequate due process.

8. During the last election campaign, the government promised to move forward within 100 days of its mandate. Given that commitment – as well as the structure of the consultation that reads more like a legislative outline rather than a genuine attempt to solicit feedback – there are considerable doubts about this consultative process. Consultations should not be a box-ticking exercise in which the actual responses are not fully factored into policy decisions. The challenge of reading, processing, analyzing and ultimately incorporating consultation responses within a three month period appears entirely unrealistic. The government should provide assurances that there will be no legislation without taking the consultation responses fully into account.

B. Specific Concerns

1. 24 Hour Takedowns

The proposed approach includes a requirement for OCSs to implement measures to identify harmful content and to respond to any content flagged by any user within 24 hours. The OCSs would be required to either identify the content as harmful and remove it or respond by concluding that it is not harmful. The OCSs can seek assistance from the new Digital Safety Commissioner on content moderation issues. The proposed legislation would then incorporate a

wide range of reporting requirements, some of which would be subject to confidentiality restrictions, so the companies would be precluded from notifying affected individuals.

By mandating such rapid takedowns of content, there is a clear risk of over-removal of content since it is difficult to give the content a proper assessment to understand its context. Furthermore, since many companies will use automatic systems to meet their legal obligations, experience elsewhere suggests that there will be significant over-removal of otherwise lawful content.[1]

a. Germany

The proposed approach appears largely modeled on the German *NetzDG* law. The German approach sparked international criticism stating with fears it would seriously harm free speech when it was adopted. It imposes a 24-hour time limit to remove obviously illegal content and allowed for up to 7 days to make a decision in circumstances where it is not clearly illegal – where an argument could be made that it was legal. This provides more nuance than the Canadian model.[2]

Since enactment, the German experience has demonstrably been shown to lead to over-removal of content as Internet services respond to high penalties by erring on the side of content removal.[3] For example, In 2018, Facebook took down a picture of a traffic sign with a bikini top on it from both Facebook and Instagram. The picture was created by an anonymous artist whose work consists of humorous and politically pointed alterations to public signs – they have won awards for their work.[4]

That same year, Twitter blocked the account of the satirical magazine *Titanic,* after they published a tweet parodying the far-right populist Alternative for Germany (AfD) party's Islamophobia. The tweet pretended to be coming from a leading AfD politician complaining about German police using Arabic numerals, which are of course standard throughout the west. Again, critics pointed to this as showing that NetzDG is over-blocking because something this clearly satirical was taken down.[5]

A large part of the problem with a 24 hour takedown requirement is that it does not allow for a fulsome analysis of edge cases. For example, in July 2018 YouTube and Twitter presented their first transparency reports for the first half of 2018 with the law. In YouTube's case, it received hundreds of thousands of takedown requests, but found only 27% justified such action. Further,

---

[1] Daphne Keller, "Empirical Evidence of Over-Removal By Internet Companies Under Intermediary Liability Laws: An Updated List" (February 8, 2021) online: *Stanford Center for Internet and Society* <http://cyberlaw.stanford.edu/blog/2021/02/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws>

[2] Darryl Carmichael & Emily Laidlaw, "The Federal Government's Proposal to Address Online Harms: Explanation and Critique" (September 13, 2021) online: *ABlawg* <https://ablawg.ca/2021/09/13/the-federal-governments-proposal-to-address-online-harms-explanation-and-critique/>

[3] Svea Windwehr & Jillian C York, "Turkey's New Internet Law Is the Worst Version of Germany's NetzDG Yet" (July 30, 2020) online: *EFF* <https://www.eff.org/deeplinks/2020/07/turkeys-new-internet-law-worst-version-germanys-netzdg-yet>

[4] Jefferson Chase, "Facebook slammed for censoring German street artist" (January 15, 2018) online: *DW* <https://www.dw.com/en/facebook-slammed-for-censoring-german-street-artist/a-4215521>.

[5] *Ibid*.

hundreds of cases required more than a week to reach a determination and dozens required external counsel to provide assistance.[6] Meanwhile, Twitter found that only 11% of takedown requests were justified and also reported that hundreds required more than 24 hours to reach a determination.[7] The lesson is clear: trading expediency for due process and careful examination of takedown claims invariably leads to over-removal of lawful content.

b. France

France has also endeavoured to establish rapid takedowns. In May 2020, it adopted a controversial online hate speech bill, known as the *Avia Bill* that required social media platforms and search engines to remove flagged hateful content within 24 hours and flagged terrorist propaganda and child sexual abuse material within one hour. Failure led to the threat of high fines.[8]

While the law may have influenced the proposed Canadian approach, it is important to note that it was struck down by the French Constitutional Court as unconstitutional. The court ruled that the 24-hour time window was "particularly brief"[9] and that this time limit to take down "manifestly illicit" online posts "could only encourage operators of online platforms to remove content that's flagged to them, whether or not it's manifestly illicit". Further, it concluded that the law constituted "an infringement of the right to free expression and communication that isn't necessary, appropriate and proportionate".[10] It also struck down the one-hour limit on taking down content deemed child pornography or terrorist content. Finally, in a statement, the court said that "freedom of expression and communication is all the more precious since its exercise is a condition of democracy and one of the guarantees of respect for other rights and freedoms."[11]

c. United States

In 2017, the United States passed the *Allow States and Victims to Fight Online Sex Trafficking Act* (FOSTA) intending to penalize sites that hosted speech related to child sexual abuse and trafficking. This is somewhat different than the online harms legislation in Germany and France because it has a far narrower scope. However, the law had the similar impact, leading to large and small Internet platforms censoring broad swaths of speech that contained adult content. This

---

[6] Thomas Wischmeyer, "'What is illegal offline is also illegal online': the German Network Enforcement Act 2017" in Bilyana Petkova & Tuomas Ojanen, eds, *Fundamental Rights Protection Online* (Cheltenham: Edward Elgar Publishing Limited, 2020) 28 at 54.
[7] *Ibid.*
[8] Laura Kayali, "France gives final green light to law cracking down on hate speech online" (May 13, 2020) online: *Politico* <https://www.politico.eu/article/france-gives-final-green-light-to-law-cracking-down-on-hate-speech-online/>
[9] Mathieu Rosemain, "France's top court rejects core of law targeting online hate speech" (June 18, 2020) online: *Reuters* <https://www.reuters.com/article/us-france-tech-regulation/frances-top-court-rejects-core-of-law-targeting-online-hate-speech-idUSKBN23P32O>.
[10] Sam Schechner, "French Court Strikes Down Core of New Hate-Speech Law" (June 18, 2020) online: *Wall Street Journal* <https://www-proquest-com.proxy.bib.uottawa.ca/docview/2414465405/ADD9B3730D7B4A0FPQ/2?accountid=14701>
[11] Asia News Monitor, "France: French constitutional court blocks large portion of online hate speech law" (June 22, 2020) online: *Asia News Monitor* <https://www-proquest-com.proxy.bib.uottawa.ca/docview/2414779631/65AA1F6DFD404946PQ/3?accountid=14701>

had devastating consequences for marginalized communities and those that served them, especially organizations that provide support and services to victims of trafficking and child abuse, sex workers, and groups/individuals promoting sexual freedom.[12]

Indeed, the law had particularly devastating consequences on already vulnerable sex workers. There had been a broad movement for sex workers to move online to better protect themselves from both the dangers of the job and from police harassment. The online setting provided online forums, client-screening capabilities, "bad date" lists, and other intra-community safety tips. Countless amounts of these sources were either taken down or had to charge significantly more in the wake of FOSTA. Ironically, the law has made the position of sex workers *more* precarious since it forces sex workers back "on the streets" or back to a pimp - and has led to significant financial instability for them.[13]

The proposed approach risks raising many of the same concerns and problems experienced elsewhere. To be clear, there is a need to establish a system for the removal of illegal content and OCSs should be expected to comply with those takedown rules. However, it is critical to ensure that takedown requirements adequately account for due process and contain essential freedom of expression safeguards. The government's proposed approach as articulated in the consultation does not meet that standard.

2. Proactive Monitoring

The proposed approach envisions pro-active monitoring and reporting requirements that could have significant negative implications. For example, it calls for pro-active content monitoring of the five harms, granting the Digital Safety Commissioner the power to assess whether artificial intelligence tools used to identify illegal content are sufficient. Moreover, the OCSs would face mandatory reporting requirements of users to law enforcement, leading to the prospect of an AI identifying what it thinks is content caught by the law and generating a report to the police. This represents a huge increase in private enforcement and the possibility of Canadians garnering police records over posts that a machine thought was captured by the law. Given the risks outlined below associated with AI and bias, the risk of machine generated police reports is particularly pronounced for BIPOC communities.

The issue of proactive monitoring has been the subject of opinions from three UN Special Rapporteurs in the context of an Indian law focused on online content regulation (Special Rapporteurs for the promotion and protection of the right to freedom of opinion and expression; on the rights to freedom of peaceful assembly and of association; and on the right to privacy).[14] The Special Rapporteurs expressed concern about the obligations of companies to monitor and rapidly remove user-generated content, which they feared will likely undermine the right to

[12] Corynne McSherry & Katitza Rodriguez, "O (No!) Canada: Fast-Moving Proposal Creates Filtering, Blocking and Reporting Rules – and Speech Police to Enforce Them" (August 10, 2021) online: *EFF*
[13] Danielle Blunt & Ariel Wolf, "Erased: The Impact of FOSTA-SESTA" online (PDF): *Hacking//Hustling* <https://hackinghustling.org/wp-content/uploads/2020/01/HackingHustling-Erased.pdf>
[14] Katitza Rodriquez & Kurt Opsahl, "India's Draconian Rules for Internet Platforms Threaten User Privacy and Undermine Encryption" (July 20, 2021) online: *EFF* <https://www.eff.org/deeplinks/2021/07/indias-draconian-rules-internet-platforms-threaten-user-privacy-and-undermine>

freedom of expression. They noted that intermediaries could over-comply with takedown requests to limit their liability and develop digital recognition-based content removal systems or automated tools to restrict content. They added that the programs are unlikely to accurately evaluate cultural contexts and identify illegitimate content. Moreover, they worried that the short deadlines, coupled with the potential criminal penalties, could lead service providers to remove legitimate expression as a precaution to avoid sanctions. The concerns mirror those arising from the Canadian government's proposed approach.

As the demand for content moderation has increased, especially through proactive monitoring provisions, companies have moved toward automated versions of monitoring flagged content – both to ensure the wellbeing of human moderators and to be able to do it quicker – but this poses a major risk to the freedom of expression online. Automated systems are not capable of consistently identifying content correctly. Human communication is complex and context-dependent. AI misses this. Reports have shown that automated process take down large amounts of legal speech, and if there is no appeals process then the speech stays down.[15]

Such automated process have been shown to disproportionately remove some content over others, penalizing Black, Indigenous and LGBTQ+ people.[16] Several studies have shown that AI models for processing hate speech were more likely to flag content from black Americans than white Americans. One study from researchers at the University of Washington[17] found that AI was 1.5 times more likely to flag tweets by black Americans over white Americans, and 2.2 times more likely to flag tweets written in African American English.[18] Another study[19] found similar evidence of substantial racial bias against black speech in five widely used academic data sets for studying hate speech – it totalled around 155,800 twitter posts.[20]

The risks associated with the proposed proactive monitoring approach cannot be overstated. The proposals risks over-removal of content and increased reliance on AI-based monitoring systems that raise significant concerns of bias. These policies are most likely to harm the very people that the policy purports to help.

---

[15] Svea Windwehr & Jillian C York, "Facebook's Most Recent Transparency Report Demonstrates the Pitfalls of Automated Content Moderation" (October 8, 2020) online: *EFF* <https://www.eff.org/deeplinks/2020/10/facebooks-most-recent-transparency-report-demonstrates-pitfalls-automated-content>

[16] Digital Rights Watch, "Explainer: The Online Safety Bill" (February 11, 2021) online: *Digital Rights Watch* <https://digitalrightswatch.org.au/2021/02/11/explainer-the-online-safety-bill/#:~:text=The%20Online%20Safety%20Bill%20was%20introduced%20in%20December,scheme%2C%20to%20remove%20material%20that%20seriously%20harms%20adults%2C>

[17] Maarten Sap et al, "The Risk of Racial Bias in Hate Speech Detection", (2019) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 1668 <https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf>

[18] Shirin Ghaffary, "The algorithms that detect hate speech online are biased against black people" (August 15, 2019) online: *Vox* <https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter>

[19] Thomas Davidson, Debasmita Bhattacharya & Ingmar Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets" (2019), *Proceedings of the Third Workshop on Abusive Language* 25 <https://arxiv.org/pdf/1905.12516.pdf>

[20] Ghaffaray, *supra* note 18.

3. Website Blocking

If the OCS does not comply with the order to remove certain content, the proposed approach introduces the possibility of website blocking with orders that all Canadian Internet service providers block access to the online communications service. The implications of these provisions are enormous, raising the likelihood of creating a country-wide blocking infrastructure within all ISPs with the costs passed on to consumers in the form of higher Internet and wireless bills.

The government's approach may be modelled on the Australian *Online Safety Act*, which grants the eSafety Commissioner the power to issue a non-negotiable request that ISPs block domains, URLs, or IP addresses hosting 'seriously harmful content'. The Commissioner does not need to observe any requirements of procedural fairness for these requests. The notices cannot be longer than 3 months, but there is no limit to how many times they can be renewed.[21] The Australian Act passed both houses of the legislature on June 23, 2021 and received Royal Assent on July 23, 2021.[22] However, the eSafety Commissioner announced that the bill will not take effect until January 23, 2022.[23] At this stage, the effects and effectiveness of the Australian law remains unknown.

However, there are numerous concerns with website blocking, particularly a state-sanctioned approach as envisioned by the government's proposal. The danger of over-blocking legitimate websites raises serious freedom of expression concerns, particularly since experience suggests that over-blocking is a likely outcome of blocking systems. The Council of Europe Commissioner for Human Rights issued a report in 2014 on the rule of law on the Internet in the wider digital world, noting,

blocking is inherently likely to produce unintentional false positives (blocking sites with no prohibited material) and false negatives (when sites with prohibited material slip through the filter). From the point of view of freedom of expression, the most problematic is widespread over-blocking: the blocking of access to sites that are not in any way illegal, even by the standards supposedly applied.[24]

The costs associated with site blocking can run into the millions of dollars with significant investments in blocking technologies and services, employee time to implement blocking orders, and associated service issues. Website blocking orders applied broadly to the myriad ISPs in Canada would have an uneven impact: larger ISPs may find it easier to integrate blocking technologies and processes into existing systems (some already block child sexual abuse material), whereas hundreds of smaller ISPs would face significant new costs that would affect their marketplace competitiveness. In fact, larger ISPs might ultimately benefit from higher fees passed along to subscribers and reduced competition. By harming the competitiveness of many

[21] Digital Rights Watch, *supra* note 16.
[22] <https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/Bills_Search_Results/Result?bId=r6680>
[23] eSafety Commissioner, "Online Safety Act 2021: Fact sheet" (July 2021) online (PDF): *eSafety Commissioner* <https://www.esafety.gov.au/sites/default/files/2021-07/Online%20Safety%20Act%20-%20Fact%20sheet.pdf>
[24] Council of Europe, Commissioner for Human Rights, The rule of law on the Internet and in the wider digital world, (2014) at 13, online: <rm.coe.int/16806da51c>

smaller providers, website blocking may jeopardize efforts to extend affordable Internet access to all Canadians.

4. Enforcement

The proposed approach identifies several measures to ensure enforcement. These include providing the public with the ability to file complaints with the Digital Safety Commissioner. The new commissioner would be empowered to hold hearings on any issue, including non-compliance or anything that the Commissioner believes is in the public interest. The Digital Safety Commissioner would have broad powers to order the OCSs "to do any act or thing, or refrain from doing anything necessary to ensure compliance with any obligations imposed on the OCSP by or under the Act within the time specified in the order." Moreover, there would also be able to conduct inspections of companies at any time:

"The Act should provide that the Digital Safety Commissioner may conduct inspections of OCSPs at any time, on either a routine or ad hoc basis, further to complaints, evidence of non-compliance, or at the Digital Safety Commissioner's own discretion, for the OCSP's compliance with the Act, regulations, decisions and orders related to a regulated OCS."

In fact, the inspection power extends to anyone, not just OCSs, if there are reasonable grounds that there may be information related to software, algorithms, or anything else relevant to an investigation.

Should a company declines to take down content, the public can also file complaints with the new Digital Recourse Council of Canada. This regulatory body would have the power to rule that content be taken down. Hearings can be conducted in secret under certain circumstances. Layered on top of these two bodies is a Digital Safety Commission, which provides support to the Commissioner and the complaints tribunal.

These proposals raise several concerns. The broad inspection power is similar to that found in Australia, where the *Online Safety Act* provides the eSafety Commissioner with the power to obtain information about the identity of an end-user of a 'social media service', a 'relevant electronic service' or 'designated internet service'. It also provides the Commissioner with investigative powers, which includes a requirement that one provides "any documents in possession of the person that may contain relevant information".[25]

The risk of overbroad or overzealous enforcement is very real. For example, in Australia Digital Rights Watch has raised concerns that investigative powers could extend to encrypted services – 'relevant electronic service' includes email, instant messaging, SMS, and chat. They note that the Commissioner has already spoken out against encryption because it makes investigations into online child sexual abuse more difficult. If extended to this realm, it could place Canadians' privacy and security at risk. The breadth of the inspection power suggests that it could also extend to journalists (raising issues involving protecting sources and freedom of the press), Internet service providers (raising privacy concerns and telecom regulatory issues), and any other

[25] Digital Rights Watch, *supra* note 16.

business or person with any link to the investigation. A far more circumscribed power with real oversight is needed.

There are also concerns about the potential caseload and the ability for the Digital Recourse Council of Canada to provide fulsome review with appropriate due process. If claims run into the thousands, the system will simply not scale in a manner commensurate with demands. While that points to the challenges of moderating online content, a different system that better accounts for the likely demands is required.