September 13, 2021

The Federal Government's Proposal to Address Online Harms: Explanation and Critique

By: Darryl Carmichael and Emily Laidlaw

Commented On: The Federal Government's proposed approach to address harmful content online

In late July, the Federal Government introduced its proposal for online harms legislation for feedback. It comprises a discussion paper outlining the government's approach to regulating social media platforms and a technical paper that provides more detail on the substance of the proposed law. The proposal is part of a suite of law reform efforts by the Canadian government concerning what can broadly be categorized as platform regulation and content regulation issues. They include Bill C-10 to reform broadcasting laws, which stalled when it hit the Senate floor (for now at least) and proposed legislation to combat hate speech and hate crimes. The timing of the online harms and hate speech proposals has been a point of contention so close to the election call. Regardless of the election result in September, it is worthwhile analyzing this proposal because the Canadian government will need to prioritize law reform in this area. Online harms legislation is sweeping the globe, and Canada is well overdue to address these issues. For better or worse (as remains to be seen), new laws have been proposed or passed in Europe, the United Kingdom (UK), Australia, India, and Turkey, to name a few.

All blog posts include a caveat that the analysis is not fulsome, but it seems crucial to emphasize that here. The scope of online harms is broad and can include anything and everything posted online, and the regulatory environment is global, even if what is discussed is domestic law. Indeed, the broad scope of this proposal is a point of criticism, with scholars such as Cynthia Khoo arguing that this should be broken down into subject-matter specific legislation. All this to say that what is offered here are the highlights of some of the key issues of debate and concern.

The Department of Heritage is open for feedback on the proposal until September 25th, depending on the election result. Therefore, this post is organized to provide such feedback. The analysis focuses on some of the major points of reform: scope and definitions, the proposed regulator, proactive monitoring, 24-hour takedown requirements, website blocking, mandatory reporting, and transparency obligations. Each point is explained and critiqued, and recommendations are made to address the criticisms. Because of the nature of this post and the breadth of the proposal, many of the recommendations are relatively general and have the same theme: implementation of this proposal needs to be slowed down and significant consultation undertaken.

By way of introduction, the proposal aims to regulate social media platforms concerning how they moderate certain types of harmful content: terrorist content, content that incites violence, hate speech, non-consensual sharing of intimate images, and child sexual exploitation content. It proposes the creation of a new regulator, the Digital Safety Commission, which would provide

recourse concerning specific items of content, oversee and investigate platforms concerning their moderation systems, and enable major administrative penalties to be levied against non-complying platforms. The proposal would also impose significant new obligations on platforms, such as to action content within 24 hours of it being flagged and to proactively monitor content on their services.

To set the scene of the complexity of content moderation, let's use a famous example in content moderation circles. In 2016, Facebook famously grappled with whether to remove a Pulitzer Prize-winning photograph of the "Napalm Girl". We all know the photo. It is the haunting image of a naked and sobbing Vietnamese girl running from a napalm attack. It was shared on Facebook as an example of photographs that had changed the history of warfare. Facebook initially removed the photo, and after severe criticism, reversed its decision and reinstated it. However, the decision of what to do is not as easy as it might seem. It is an iconic and newsworthy photo of historical significance, but it is also a brutal and intimate image of a child at one of the most horrific moments in their lives. Years later, the girl depicted in the photo, now 44 years old, commented, "[t]he more the picture got famous, the deeper the cost to my private life." But someone had to make that decision – and for platforms, it is a content moderator, automated system, or both depending on how their system is designed.

Let's play out how this image might be treated in this proposal. One of the categories of harm in the proposal is intimate images as defined in the *Criminal Code*, RSC 1985, c C-46, "but adapted to a regulatory context" (technical paper, para 8), which presumably means a broader rather than a narrower scope. The photo might be an intimate image pursuant to s 162.1 of the *Criminal Code* because it shows nudity and was taken and shared without consent. It does not depict a sexual activity. However, depiction of sexual activity is not required in the definition an of intimate image. Rather, s 162.1(2) requires the depiction of nudity *or* sexual activity, although the section is bundled under the heading of "Sexual Offences" with crimes such as sexual exploitation and voyeurism. The photo must also be shared in circumstances giving rise to a reasonable expectation of privacy, and that is a matter of debate in these circumstances. There is a defence that the conduct serves the public good. One argument is that taking and sharing this image – by the original photographer and by the user on Facebook – serves the public good. However, a public good defence is not a newsworthy defence, although they overlap. The question is whether there is some public benefit. Certainly, there is a strong argument here that taking and sharing the image is a public good, but it is not obvious given that this is a child and the intimate circumstances.

Now let's take a step back. This is not a decision by the police about whether to investigate or the Crown about whether to prosecute an individual nor is this a decision by a court whether to convict, all of which would be highly unlikely. This is a social media platform deciding whether to remove content from circulation in light of its obligations under this new online harms proposal if made law. Platforms would have 24 hours from the moment the content is flagged to make the assessment. This should be understood in the context that hundreds of million photos are posted to Facebook daily. Conservative advice to the platform would be to take the image down. But that requires the image to be de-contextualized and advice provided purely based on risk avoidance, and that is the problem. The risk to the platform is a potentially enormous administrative penalty, even if remote, and while the image is defensible, it is not obviously so. Blunt legislation forces

blunt responses by those regulated by it, and we all lose because the richness and complexity of how we converse will be neutered.

Scope & Definitions

The proposal sets its sights on Online Communication Services (OCS) and providers of these services (OCSPs). The proposal restricts its application to services whose primary purpose is to enable communication with other users of the service over the internet. This will explicitly exclude private communications, telecommunication service providers, search engines, caching services and potentially others to be specified later (technical paper, paras 1-6). The result is that the proposal targets social media like TikTok, Facebook, and Twitter but would not include WhatsApp, review sites, or comment sections on news pages. Given the onerous obligations and sizeable administrative penalties, this perhaps makes sense. However, this approach might not help achieve the objective of reducing harmful content online.

It is unclear if only major social media platforms are targeted and what precisely that is. The discussion paper states that the intention is to target "major platforms" (discussion guide). However, the definition in the technical paper captures all social media platforms (see definitions of OCS and OCSP in paras 2 and 4). Europe's proposed *Digital Services Act* differentiates between "online platforms" and "very large online platforms" based on the number of users, imposing more onerous obligations on the large platforms. Under the Canadian proposal, the Governor in Council would have the power to add or remove categories of services from the definition of OCS. The Federal Government should consider carefully the OCSPs it wants to target. For example, JustPaste.it was started by a Polish student from his bedroom, and for a while, it was the platform of choice for ISIS. Justpaste.it has taken steps to address the challenge of terrorist content on its site, including joining the Global Internet Forum to Counter Terrorism (GIFCT). Based on the technical paper, JustPaste.it would be captured as an OCSP, but if the intent is to target major platforms, then JustPaste.it would potentially be excluded from the scope. As will be evident, the proposal does not impose obligations on a sliding scale like the *Digital Services Act*. The platform is either in or out.

Despite the title 'online harms,' the proposal more narrowly targets criminal content, specifically terrorist content, content that incites violence, hate speech, non-consensual sharing of intimate images and child sexual exploitation content. Limiting the ambit to these five harms may be a surprise to the public, as a great deal of other harmful online activities such as bullying, harassment, defamation, or invasion of privacy are out of scope. There would be unavoidable constitutional questions if the bill were to include all of these harms and the kitchen sink, but as drafted, the proposal creates an odd situation where victims of great swaths of abuse and harm cannot avail themselves of the regulator. Not to let the perfect be the enemy of the good, the content that *is* covered by the proposal certainly should be (and is already addressed criminally), but there is enormous room for improvement if this proposal is going to live up to its name.

As an example, let's examine how narrow hate speech and terrorism are from a legal perspective. Hate speech as interpreted by the Supreme Court of Canada in *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 (CanLII) (*Whatcott*), sets quite a high bar to meet: to qualify, content must communicate an expression of "detestation" or "vilification" of an individual or group on the basis of a prohibited ground of discrimination. Hurtful, humiliating, or offensive comments do not meet this threshold, even when based on a protected characteristic, such as religion, sexual orientation, gender identity or disability. Similarly, extreme, venomous abuse that isn't based on a protected characteristic while hate*ful* would not be hate *speech* (paras 41, 55-59). As one can readily observe, this leaves a tremendous grey area of abusive, arguably quite harmful content outside of the scope of this proposal.

Terrorist content is another example where many would say they have a general idea of what it includes but would struggle to produce a legal definition. Indeed, even the *Criminal Code* definition is spread across multiple sections, requiring some effort to pull together a clear understanding (see analysis of Bill C-51 by Craig Forcese and Kent Roach here). The technical paper classifies terrorist content as that which both "actively encourages terrorism and which is likely to result in terrorism" (technical paper, s 8). It is difficult to gauge whether the definition of terrorist content in this proposal expands or mirrors that of the *Criminal Code*, in particular s 83.221 that criminalizes counselling another "to commit a terrorism offence without identifying a specific terrorism offence." This determination is also made more difficult by unanswered questions about whether platforms would be required to assess the criminality of the objectionable content on the criminal standard (beyond a reasonable doubt) or under a regulatory or civil standard (balance of probabilities).

Forcese and Roach warn that "any terrorist speech prosecution, especially for speech on the internet, will be difficult to sustain." (at 215) Examining a case where criminal charges were brought relating to 85 different social media posts from 14 different accounts, they identified the difficulties in assessing whether a given instance of speech amounts to "counselling" terrorism, especially when there is more of a cumulative effect than a clear black-and-white example. There was also concern that the judge may have been too focused on individual posts, and they suspect that "a criminal jury might … have taken a more holistic approach", seeing the criminality in the forest where it was lacking in any individual tree (at 214). This is an equally pressing concern in the realm of content moderation, given that any individual piece of content lacks the context of other posts from the same user. While Forcese and Roach would support a separate form of terrorist speech offence not tied to the "counselling" offence found in the *Criminal Code* (s 83.221), we have to question whether this proposal seeks to introduce that very thing in a regulatory setting.

A danger that arises from these complex definitions is that a content moderator without legal training may resort to bias or heuristics rather than performing in-depth analysis of borderline cases where the decision to remove the content rests on a contextual analysis. Moderators are already overworked and not qualified to make legal determinations. Combining this with the virtual mountains of flagged content that they are tasked with assessing will inevitably lead to over-removal and inadequate consideration of the legal criteria. Empirical studies show a tendency by

platforms to over-remove content in notice and takedown regimes, meaning that legal content is already removed even without 24-hour time limits. This is exacerbated when automated mechanisms are used as the level of nuance and appreciation of context required to make these kinds of decisions is currently beyond their capability.

Recommendations:

- "Online Criminal Content Regulation" is a more accurate nomenclature until the proposal addresses other harmful content that doesn't rise to the level of the criminal definitions. "Beverage regulation" would not be an accurate name for a bill that only addressed alcoholic beverages.
- Consult and rescope the definitions of harmful content. In particular, examine the impact of adopting narrow definitions from the *Criminal Code* and related case law versus broader definitions in terms of the reality of content moderation practices, and harmful content sought to be reduced. The approach should be clear and justification provided.
- In evaluating harmful content, ensure that situations are captured where volume and persistence creates a situation of harm, even where any individual act or post would not violate the regulations.
- Introduce laddered obligations drawing from the *Digital Services Act*. There should be specific and more onerous obligations on major platforms, however defined, but other platforms should not be entirely out of scope.

Regulator

The proposal would create a Digital Safety Commission funded by fees levied upon industry. It would be comprised of three bodies: the Digital Safety Commissioner, Recourse Council, and Advisory Council. We broadly favour the idea of a new regulator, but it will depend on some of the finer details that are not set out in the proposal. The need to see the finer details of what is proposed is acute in this case because this will be a regulator of expression online, an unprecedented role for a regulatory body well beyond that of a human rights commission, which could have a potentially sweeping impact on day-to-day expression online. The scope and remit of the regulator needs to be fleshed out to ensure that it carefully balances rights, in particular expression, equality, and privacy, with a clear understanding of the platform economy. And like any regulatory entity, whether this is a good idea depends on the entity being suitably funded and populated with individuals with the right training and expertise.

The Digital Safety Commissioner would bear some resemblance to the role of our federal privacy commissioner. The Commissioner would oversee and enforce content moderation obligations and engage in education, research, and outreach (discussion guide, module 1, technical paper, module 1(a)). The Commissioner would receive and investigate complaints about non-compliance by regulated entities. Appeals of decisions of the Commissioner would be made to the Personal Information and Data Protection Tribunal (Tribunal), which is proposed to be created with Bill C-11, (*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*, 2nd

Sess, 43rd Parl, 2020) the proposed new consumer privacy legislation to replace the *Personal Information Protection and Electronic Documents Act*, SC 2000, c 5 (see discussion here). The Commissioner would also have the power to do such things as proactively inspect for compliance or non-collaboration, issue reports and compliance orders.

Importantly, the Commissioner would have extraordinary power to recommend or refer non-compliance to bodies with the power to impose significant fines or order website blocking and filtering. For example, the Commissioner can recommend an administrative monetary penalty of up to 10 million dollars or 3% of gross global revenue, which would be decided by the Tribunal similarly charged with administering fines for privacy breaches. Or non-compliance could be referred to prosecutors with potential fines up to 25 million dollars or 5% of gross global revenue. Further, the Commissioner could apply to the Federal Court for an order that an ISP block or filter entire websites that repeatedly fail to remove child sexual exploitation or terrorist content (technical paper, paras 102-109). The immensity of these fines or recourse, coupled with the substance of the rules proposed, creates a high-risk environment that incentivizes over-removal of content by platforms to avoid legal risk.

The Digital Recourse Council, comprised of 3-5 members, would offer a different dispute resolution mechanism. While the Commissioner would focus on whether the platform has in place the procedural safeguards mandated by legislation, the Recourse Council would be concerned with whether a platform made the correct decision about a specific item of content. A person may make a complaint to the Recourse Council concerning a decision of a platform to either remove or not remove content. However, a complainant must first exhaust all avenues of appeal within the platform. Once a complaint is made, the platforms and affected individuals would be provided with a notice of the complaint and an opportunity to make representations. If the Council decides the content is harmful, they can order the OCSP to take it down. If the Council decides the content is *not* harmful, they will issue their decision to the OCSP, and it is then in the hands of the platform whether to reinstate or remove the content subject to the platform's own terms of use (technical paper, paras 45-59).

Hearings by the regulator (either Commissioner or Council) can be held in camera if there are compelling reasons to do so. The government suggests these reasons could include privacy, national security/defence, international relations, or confidential commercial interests (technical paper, para 59). These secret hearings have attracted a great deal of critical comment, and we suggest they need to set a clear threshold or criteria for situations where these can be invoked.

The Advisory Board would be comprised of up to 7 part-time members who would provide expert advice to both the Commissioner and Recourse Council about a variety of issues such as "emerging industry trends and technologies and content-moderation standards" (discussion paper, technical paper paras 71-75). Khoo recommends that the advisors are integrated within the Commissioner's office and Recourse Council. We are not necessarily opposed to a separate advisory council, but further consultation is required about the best structure for these three bodies.

The concept of a digital regulator to address human rights or online harms issues is something that has been advocated by some scholars. Khoo recommends the creation of a centralized regulator to address technology-facilitated gender-based violence in her ground-breaking report "Deplatforming Misogyny." She envisions a body with a dual mandate to provide legal recourse and support to those impacted, and research and training. One of the authors of this post, Emily Laidlaw, has advocated for creation of a digital regulator in her scholarship. For example, the framework for a digital rights commission was detailed in chapter six of *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibilities*, (Cambridge: Cambridge University Press, 2015), with similar emphasis to the Digital Safety Commission on the need for multiple forms of support in the form of a remedial mechanism, corporate support (policies, assessment tools, audits), and education and research. This framework was developed into a specific proposal for the Law Commission of Ontario for its project Defamation Law in the Internet Age. The crib notes version is that Laidlaw recommends creation of an online tribunal for defamation disputes modelled on British Columbia's Civil Resolution Tribunal (see here and here). Further, several states have created, or are in the midst of creating, regulators to address some aspects of online harms, including in the UK and Australia.

An exploration of the benefits and drawbacks of alternative regulators for online harms is beyond the scope of this blog post. In short, the high volume of content combined with the potential devastating harms and need for speed in addressing the complaint makes courts unsuitable to adjudicate most cases. Further consultation is necessary to ensure that the regulator is narrowly scoped and achieves the goal of reducing harm and protecting rights. Some questions include:

- How will the complaints process to the Recourse Council be structured to ensure access to justice and disincentivize frivolous or abusive complaints? Examples include issues of volume of complaints, specious complaints, ease of making complaints, disparate burdens being placed on complainant/complainee.
- Who can complain?
- What is the burden of proof?
- How will this process be structured to ensure speedy resolution and due process?
- How will a complainant prove that they have exhausted all avenues of appeal via the platform? How much additional time, burden and chilling effect will this place on a complainant whose content was incorrectly removed?
- What should be the training and expertise of the Recourse Council?
- How will grey zone expression be treated?
- How will the Recourse Council intersect with the internal content moderation policies of the OCSPs?
- How will the Advisory Council interact with the other bodies? For example, what if the Advisory Council's advice is repeatedly not accepted or adopted by the Recourse Council or Digital Commissioner?

Recommendation:

- Consultation on the details of the proposed Digital Rights Commission with a focus on fleshing out the scope and remit to ensure access to justice and balancing rights.

Proactive Monitoring

The technical paper obligates OCSPs to "take all reasonable measures, which can include the use of automated systems" to identify and remove harmful content (technical paper, para 10). To comply with this obligation necessitates that a platform pro-actively monitor content hosted on their services. This is because the burden is on the OCSP to identify harmful content for removal. Thus, the platform must come up with a way to find this content, which usually entails broad surveillance, often analyzed and actioned through automated means, and at risk of function creep. The impacts of such an approach are numerous, including the privacy and freedom of expression of users, often with a greater impact on marginalized and racialized groups. This can be contrasted with a complaints-based system, such as a notice and action regime, which relies on user complaints to trigger a platform's obligation to act.

General laws that mandate proactive monitoring are controversial and criticized as a human rights infringement, and for good reason. For example, David Kaye, the former Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, stated in his 2018 report that proactive monitoring and filtering are "inconsistent with the right to privacy and likely to amount to pre-publication censorship." (paras 15-17, 67) These risks are compounded when content is prevented from being uploaded, which operates as a system of prior restraint. It creates an opaque system for users and state bodies tasked with oversight, uncertain of the rules, process, or decision-making behind moderation decisions.

This is not to say that automated systems should never be used. That would be unrealistic, and the conversation is better directed to how these systems can be designed in a way that is human rights compliant. Further, this is not to say that platforms should never prevent content from being uploaded, in particular child sexual abuse content. However, no state should mandate a general obligation to monitor. Rather, the obligation, if any, must be more narrowly carved. The proposal states that such measures should be implemented in a way that does not discriminate pursuant to the *Canadian Human Rights Act*, RSC 1985, c H-6 (technical paper, para 10). We are of the opinion that this cannot be solved at the implementation stage, and the provision must be scrapped or more narrowly carved.

There are certainly examples of legislation that impose proactive monitoring of some sort. Kaye cites both China's 2016 Cybersecurity Law and Germany's *Network Enforcement Act* (NetzDG) as laws that explicitly or implicitly force proactive monitoring that either lead to filtering or reporting to law enforcement (paras 15-16). However, Article 15 of the *E-Commerce Directive* prohibits European Member States from mandating general proactive monitoring, although specific content might be actioned to be removed and kept down. Further, the prohibition on general monitoring was maintained in the *Digital Services Act*, and exploration of such an obligation for terrorist content was softened in the final version of the EU's Regulation on preventing the dissemination of terrorist content online.

The UK's approach to proactive monitoring in the *Online Safety Bill* is illustrative as it is more narrowly circumscribed. There is no general obligation to monitor, although arguably the duty on platforms to take steps to manage the risk of harm of their platforms entails proactive monitoring. More specifically, the proposal is that the regulator Ofcom could issue a Use of Technology notice. The notice would only be available after a warning was provided and would be limited to child sexual abuse material and terrorist content, and the notice would mandate the use of "accredited technology" selected by Ofcom. We are not weighing into the strengths and weakness of the UK approach here, which is also controversial, but rather aim to highlight that even compared to the highly contentious UK model, the Canadian proposal is an outlier.

Recommendation:

- The proposal of a general obligation to monitor all harmful content should be categorically rejected. Consultation and analysis should be undertaken to explore options that are proportionate and effective to achieve the objective of reducing circulation of harmful content. Options include, without endorsement, but for discussion, more targeted measures:
    - A duty of care model requiring platforms take reasonable care in the management of their services with specific safeguards to address the risk of general monitoring;
    - Exploration of human rights safeguards that can be the central framework in content regulation and what that would entail e.g., criteria for specific versus general monitoring; and
    - Exploration of what a system of reasonable decision making might be, and how to build a cushioning system for mistakes (see this proposal by Marcelo Thompson).

24 Hour Takedown Requirements

An aspect of the proposal that has already attracted a great deal of critical attention is the requirement for content to be addressed by a platform within 24 hours of being flagged. This 24-hour requirement is separate from a platform's obligations to take all reasonable proactive measures, but instead starts a clock from the point at which a piece of content is flagged. The Governor in Council may pass regulations to adjust this timeframe for different types or subtypes of harmful content, but under this power the timeline could also be *shortened* below 24 hours (technical paper, paras 11-12).

From the point at which the content is flagged, the platform has 24 hours to either remove the content or respond to the flagger indicating that the content does not meet the definition of harmful under the Act. It is worth bearing in mind that a platform like Facebook already has systems to flag content that goes well beyond the scope of harmful content captured in this proposal. This context is needed when considering the scale and volume of content that must be moderated. A report from NYU indicates that for Facebook alone, they made moderation decisions on approximately 3.75 billion pieces of content in the first quarter of 2020, and the accuracy of those decisions falls well short of 100%. As acknowledged by Facebook founder and CEO Mark

Zuckerberg: "[t]he vast majority of mistakes we make are due to errors enforcing the nuances of our policies rather than disagreements about what those policies should actually be. Today, depending on the type of content, our review teams make the wrong call in more than 1 out of every 10 cases." (source) The NYU report goes on to recommend that Facebook double the size of its moderation team just to tackle the accuracy deficits in their existing moderation regime. Imposing a 24-hour window for reviewing and responding to specific types of content – especially if Facebook chooses not to double its moderation staff – would suggest that concerns over the quality of moderation decisions will only increase.

The proposal seems modelled on German legislation known colloquially as NetzDG, which also imposes a 24-hour time limit to remove content. However, even NetzDG, controversial in its own right, appreciates that decisions about the illegality of content sometimes require more than a few moments to make. NetzDG restricts the 24-hour rule to obviously illegal content and allows for up to a week for a platform to respond to content in circumstances where factual considerations could render the content legal, or where the poster might have a legitimate defence for posting it. Thus, NetzDG grants platforms more time to assess expression in the grey area. Canada's proposal contains none of that nuance.

This consideration of nuance is even more important here given that hate speech must be assessed in light of the Supreme Court of Canada's jurisprudence (e.g. *Whatcott, R v Keegstra*, [1990] 3 SCR 697, 1990 CanLII 24). As the NYU report highlights, the moderators for Facebook are not legally trained, and are frequently outsourced to other countries. Expecting accurate analysis of hate speech by untrained and overworked moderators is unrealistic and even less so with a liability clock ticking in the background. When the incentives on a platform weigh almost entirely in favour of removal in situations where there is the slightest doubt, and they have no reasonable motivation to consider the poster's right to free expression, the inevitable outcome is censorship.

Let us be entirely clear here – removal is the right decision for legitimately harmful content, and it is important that such content is removed quickly. The danger is that there remains a significant grey area within which entirely *legal* content could be removed solely due to platform risk-avoidance. The question the proposal wrestles with – as do all of us working in this area – is where to draw the line, and it is our view the proposal heavily errs on the side of protection from harm in a way that undermines their goals. It is notable that both the discussion and technical papers devote minimal attention to the value of freedom of expression. The 24-hour rule embodies the unbalanced analysis threaded throughout the proposal. Further, this proposal would disproportionately impact marginalized, racialized and intersectional groups (see Suzie Dunn's commentary here). For example, platforms' internal complaints systems are regularly used as mechanisms of abuse whether by a persistent individual or mobs who maliciously flag content, or because the design of the content moderation system and its practice discriminates. The examples are endless: removal of Black Lives Matters posts, LGBTQ+ posts, sexualized content and posts raising awareness of missing and murdered Indigenous women and girls (see, scholarship on this issue here and here). The result is that the voices of the very groups we seek to protect would be further silenced.

Recommendations:

- The 24-hour time limit should be abandoned in favour of a generic obligation to act expeditiously. Or, at minimum, exceptions should be drafted, which allow additional time to engage in a contextual analysis of expression in the grey zone, similar to the NetzDG model.
- Incentivize platforms to protect free expression when making moderation decisions in order to avoid banket removals. Consider addressing the prevalence of harmful content, not merely its presence (see Facebook Whitepaper at pp 9 & 13).
- Explore more creative options. For example, the *Digital Services Act* incorporates a "trusted flagger" system wherein complaints from a verified trusted flagger (person or organization) can be handled on an expedited basis. The status of the trusted flagger is contingent on the accuracy and quality of complaints made. If a trusted flagger has a certain number of "false positives" they can lose their status. (Article 19(5)-(6))

Website Blocking

In a section of the technical paper under the heading "Exceptional Recourse" are provisions dealing with granting powers to the Digital Safety Commissioner to block entire websites (paras 120-123). On its face, this seems to be an alarming power of censorship, but as drafted it would have quite narrow application. Website blocking would be restricted to child sexual exploitation content and terrorist content. Further, before this could be used, a provider must have demonstrated persistent non-compliance with orders to remove such content, *and* all other enforcement measures must have been exhausted. To reach the point of website blocking, there would have to be compliance orders made, fines issued, which may also be preceded by hearings, and each of these decisions can also be appealed before resorting to banishing a website from Canadian soil.

In the UK, for example, their proposed *Online Safety Bill* would permit the regulator, Ofcom, to enact "business disruption measures", an escalating form of sanction against non-compliant services. The measures would actually apply to ancillary parties rather than the targeted service directly, meaning ISPs, app stores, payment providers or search engines. The UK Bill frames these as levers to be used against platforms that are bucking the regulator's authority, allowing Ofcom to cut off payment processing and search results for the misbehaving service before proceeding to outright blocking. Notably, the business disruption measures can be used for various forms of non-compliance and are not strictly limited to persistent non-removal of terrorist or CSEA content.

One of the main issues with website blocking is overreach and overbreadth. The proposal here would allow the blocking of entire platforms. When website blocking has been used in a more targeted and human rights compliant manner, it has either blocked specific webpages or targeted websites that are primarily devoted to hosting illegal content (e.g. piracy websites). The proposal here does not limit website blocking in this way; rather, it relies purely on the idea of blocking as a last resort.

Recommendations:

- Maintain tightly limited scope and availability of this enforcement measure, including requiring judicial authorization.
- Add warning steps and procedural protections to ensure platforms can make representations before drastic measures are pursued.
- Examine limiting website blocking to specific webpages, or when that is not possible, to OCS that are primarily devoted to sharing illegal content.

Mandatory Reporting

The discussion paper proposes two options for consideration. First, mandatory reporting to law enforcement where "there are reasonable grounds to suspect there is an imminent risk of serious harm." (Module 2) This is a reasonable and high threshold to trigger reporting to police and something which some major platforms already do (see, for example, Twitter). The second option proposes a significantly lower threshold. It would mandate reporting to law enforcement (criminal content) and CSIS (national security content) when there are reasonable grounds to believe or be suspicious that the content is illegal within the five categories of content. This is not a good approach. It would force a platform to report content that *might* be illegal, thus targeting grey zone speech. As identified above, Black Lives Matter posts have been mistakenly labelled hate speech and removed in the content moderation process. Pursuant to this proposal, such posts would need to be forwarded to law enforcement, further harming racialized groups and undermining equality-seeking goals of online harms legislation. As Daphne Keller notes, Germany has made a similar proposal, which is being challenged by Google for violating fundamental rights. If the first option is implemented, it is critical that mandatory reporting is not coupled with a proactive monitoring obligation.

Recommendations:

- Limit mandatory reporting to circumstances where it is reasonably suspected there is an imminent risk of serious harm.
- Limit the basis for mandatory reporting, to a complaints-based approach or reasonable awareness.
- Do not impose proactive monitoring coupled with any mandatory reporting.

Transparency Obligations

To aid in monitoring and enforcing the various aspects of the proposed regulation, OCSPs will also be subject to reporting and transparency requirements (technical paper, para 14). Some of these are designed to provide insight into how moderation is being conducted behind closed doors, and others are a somewhat roundabout way of pressing platforms to address systemic concerns.

Platforms will provide data on an annual basis detailing the types and volumes of harmful content found, and separately disclosing content deemed objectionable on the basis of the platforms' own community standards, but which would not have been defined harmful under the regulations. They are also required to report on the resources and personnel allocated to content moderation, along

with their procedures, practices, and systems, including automation. Further, the proposal requires that platforms collect and sort data in a way that might be difficult to achieve. For example, the proposal would require Canada-specific data, which depending on the platform, may not be realistic to provide. For example, would the data about the "volume and type of content dealt with at each step of the content moderation process" (technical paper, para 14) be limited to posts made by Canadian-based users, or visible by Canadian users? Would any public posts visible by a Canadian be in scope?

The proposal would also require platforms to report on how they monetize harmful content, which stands out from the others as an odd inclusion. It seems designed as a form of public shaming to these companies, and therefore the data would be less reliable from the start, given that companies have an incentive to claim that they are not profiting from the hate speech or child pornography posts of their users. Also, the impetus behind this reporting requirement seems to be inspired by the EU's proposed *Digital Services Act*, wherein platforms are obliged to perform a yearly risk assessment to identify systemic risks arising from their services. However, the Federal Government's proposal is flawed because it frames the platforms as villains being unmasked by the data, rather than using the European approach enlisting platforms as collaborators seeking to solve the systemic problems.

Mandating and standardizing the content and requirements around transparency for platforms is essential, both as a method of accountability and as an avenue to better understand and address systemic problems.

Recommendations:

- Platforms should still be required to address systemic problems, but the proposal should avoid framing the requirement as a "gotcha" on platforms, rather enlisting platforms as collaborators, and using data from transparency reports as an accountability tool.
- Consult with platforms and organizations such as GIFCT and the Global Network Initiative about appropriate and achievable transparency reporting requirements.

Conclusion

Overall, while there are meritorious elements to the Federal Government's online harms proposal, the problematic areas are *very* problematic. We commend the proposal to create a Digital Safety Commission. However, the substance of the rules, although ticking all the right boxes as to topics that should be on the agenda for debate, requires a massive overhaul. When we say that it ticks the right boxes, we mean that it explores key issues of debate in internet law and policy, such as website blocking, time-limited content takedown rules, identification and actioning of content, transparency reporting and so on. However, with each of these topics, the solution proposed is rather blunt when nuance is needed to balance the various interests and rights. In the area of content regulation, to achieve the objective of an internet ecosystem that broadly strives to protect and balance various rights and interests, it is the multitude of little decisions that determine the overall effect.

Canada needs legislation to address online harms, but there must be more thorough consultation and more thoughtful consideration of every element of this proposal. There are multiple other regimes to look to that have tackled some of these very same issues. None are perfect, but the EU's *Digital Services Act* would be a good place to start, and the wealth of scholarly, industry and civil society attention that has been devoted to the issue of platform regulation and content moderation.

Summary of Recommendations:

- "Online Criminal Content Regulation" is a more accurate nomenclature until the proposal addresses other harmful content that doesn't rise to the level of the criminal definitions. "Beverage regulation" would not be an accurate name for a bill that only addressed alcoholic beverages.
- Consult and rescope the definitions of harmful content. In particular, examine the impact of adopting narrow definitions from the *Criminal Code* and related case law versus broader definitions in terms of the reality of content moderation practices and harmful content sought to be reduced. The approach should be clear and justification provided.
- In evaluating harmful content, ensure that situations are captured where volume and persistence creates a situation of harm, even where any individual act or post would not violate the regulations.
- Introduce laddered obligations drawing from the *Digital Services Act*. There should be specific and more onerous obligations on major platforms, however defined, but other platforms should not be entirely out of scope.
- Consultation on the details of the proposed Digital Rights Commission with a focus on how to structure it to balance rights and ensure access to justice.
- The proposal of a general obligation to monitor all harmful content should be categorically rejected. Consultation should be undertaken to explore options that are proportionate and effective to achieve the objective of reducing circulation of harmful content. Options include, without endorsement, but for discussion, more targeted measures:
    - A duty of care model requiring platforms take reasonable care in the management of their services with specific safeguards to address the risk of general monitoring;
    - Exploration of human rights safeguards that can be the central framework in content regulation and what that would entail e.g. criteria for specific versus general monitoring;
    - Exploration of what a system of reasonable decision making might be, and how to build a cushioning system for mistakes (see Marcelo Thompson for this proposal).
- The 24-hour time limit should be abandoned in favour of a generic obligation to act expeditiously. Or, at minimum, exceptions should be drafted, which allow additional time to engage in a contextual analysis of expression in the grey zone, similar to the NetzDG model.
- Incentivize platforms to protect free expression when making moderation decisions in order to avoid banket removals. Consider addressing the prevalence of harmful content, not merely its presence (see Facebook Whitepaper at pp 9 & 13).

- Explore more creative options. For example, the *Digital Services Act* incorporates a "trusted flagger" system wherein complaints from a verified trusted flagger (person or organization) can be handled on an expedited basis. The status of the trusted flagger is contingent on the accuracy and quality of complaints made. If a trusted flagger has a certain number of "false positives" they can lose their status. (Article 19(5)-(6))
- Maintain tightly limited scope and availability of this enforcement measure, including requiring judicial authorization.
- Add warning steps and procedural protections to ensure platforms can make representations before drastic measures are pursued.
- Examine limiting website blocking to specific webpages, or when that is not possible, to OCS that are primarily devoted to sharing illegal content.
- Limit mandatory reporting to circumstances where it is reasonably suspected there is an imminent risk of serious harm.
- Limit the basis for mandatory reporting, to a complaints-based approach or reasonable awareness.
- Do not impose proactive monitoring coupled with any mandatory reporting.
- Platforms should still be required to address systemic problems, but the proposal should avoid framing the requirement as a "gotcha" on platforms, rather enlisting Platforms as collaborators, and using data from transparency reports as an accountability tool.
- Consult with platforms and organizations such as GIFCT and the Global Network Initiative about appropriate and achievable transparency reporting requirements.

This post may be cited as: Darryl Carmichael and Emily Laidlaw, "The Federal Government's Proposal to Address Online Harms: Explanation and Critique" (September 13, 2021), online: ABlawg, http://ablawg.ca/wp-content/uploads/2021/09/Blog_DC_EL_Federal_Online_Harms_Proposal.pdf