Rebuilding Canada's Public Square

Response to the Government of Canada's Proposed Approach to Address Harmful Content Online

September 2021

Sam Andrey | Alexander Rand | M.J. Masoodi | Karim Bardeesy

cybersecure policy exchange

Powered by RBC®

Cybersecure Policy Exchange

The Cybersecure Policy Exchange (CPX) is an initiative dedicated to advancing effective and innovative public policy in cybersecurity and digital privacy, powered by RBC through Rogers Cybersecure Catalyst and the Ryerson Leadership Lab. Our goal is to broaden and deepen the debate and discussion of cybersecurity and digital privacy policy in Canada, and to create and advance innovative policy responses, from idea generation to implementation. This initiative is sponsored by the Royal Bank of Canada; we are committed to publishing independent and objective findings and ensuring transparency by declaring the sponsors of our work.

Rogers Cybersecure Catalyst

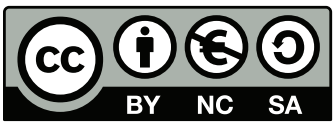
Rogers Cybersecure Catalyst is Ryerson University's national centre for innovation and collaboration in cybersecurity. The Catalyst works closely with the private and public sectors and academic institutions to help Canadians and Canadian businesses tackle the challenges and seize the opportunities of cybersecurity. Based in Brampton, the Catalyst delivers training; commercial acceleration programming; support for applied R&D; and public education and policy development, all in cybersecurity.

Ryerson Leadership Lab

The Ryerson Leadership Lab is an action-oriented think tank at Ryerson University dedicated to developing new leaders and solutions to today's most pressing civic challenges. Through public policy activation and leadership development, the Leadership Lab's mission is to build a new generation of skilled and adaptive leaders committed to a more trustworthy, inclusive society.

How to Cite this Report
Andrey, S., Rand, A., Masoodi, M.J., and Bardeesy, K. (2021, September). *Rebuilding Canada's Public Square.* Retrieved from https://www.cybersecurepolicy.ca/public-square.

350 Victoria St, Toronto, ON M5B 2K3



Contributors
Nour Abdelaal, Policy Analyst, Cybersecure Policy Exchange
Sam Andrey, Director of Policy & Research, Ryerson Leadership Lab
Karim Bardeesy, Executive Director, Ryerson Leadership Lab
Sumit Bhatia, Director of Innovation and Policy, Rogers Cybersecure Catalyst
Zaynab Choudhry, Design Lead
Charles Finlay, Executive Director, Rogers Cybersecure Catalyst
Mohammed (Joe) Masoodi, Senior Policy Analyst, Cybersecure Policy Exchange
Alexander Rand, Research and Policy Assistant, Cybersecure Policy Exchange
Stephanie Tran, Research and Policy Assistant, Cybersecure Policy Exchange
Yuan Stevens, Policy Lead, Cybersecure Policy Exchange

For more information, visit: https://www.cybersecurepolicy.ca/

@cyberpolicyx @cyberpolicyx Cybersecure Policy Exchange

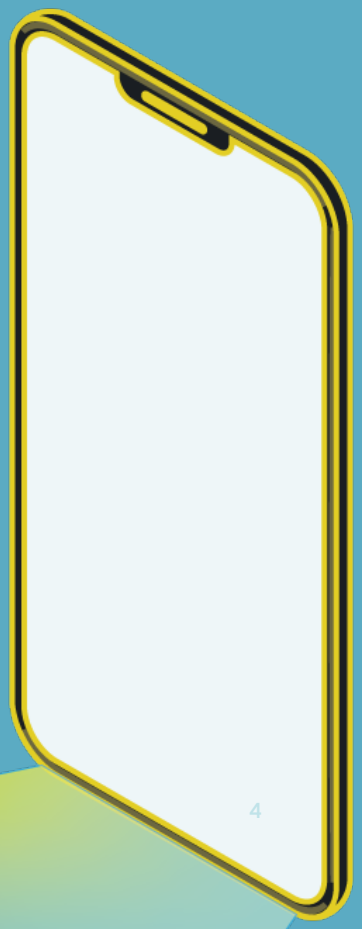
Executive Summary

Social media is in many ways the new public square — where most Canadians now connect with friends and family, and engage in civic discourse. It has become increasingly clear that this new square is having a toxic influence on our society and democracy: hate speech and harassment targeting marginalized people; disinformation enabling extremism and conspiracy theories to flourish; and online activities fueling real-world violence and exploitation.

Over the past three years, we conducted national representative surveys with Canadian residents on these important issues. Key findings include:

- **More than one in three Canadian residents report encountering harmful content online at least weekly**, such as hate speech and violent material.
- That figure rises to about half of those who regularly use social media for news and current events.
- **Racialized Canadians are 50% more likely than non-racialized Canadians to encounter racist content online** and report content to platforms for being hateful.
- **Canadians do not trust social media platforms to act in the public's best interest.** In fact, they are less trusted than oil companies, telecommunication providers and news media.
- **71% of Canadians want the government to intervene in social media companies in 2021** — up from 60% in 2019.
- **75% of Canadians support requirements for platforms to delete illegal content in a timely manner** such as hate speech, harassment and incitement to violence.

These results underscore that Canadians are concerned about what they experience on social media, and are looking for action to help address the harms produced. The unique reach and speed of social media platforms call for unique regulatory solutions aimed at countering the spread of online harms, while at the same time protecting Canadians' rights and freedoms, including our right to free expression. The Government of Canada has **announced its intention** to introduce new legislation to address some of these harms, namely: hate speech; terrorist and violent content; child sexual exploitation; and non-consensual sharing of intimate images.



Intent of Report

This report is intended to provide our best advice on how to begin to genuinely rebuild this new public square in a manner that protects and advances Canadians' fundamental rights and freedoms, and furthers efforts at international platform governance alongside allied jurisdictions. Our recommendations to improve the Government's proposal include:

1. Clarify the online platforms in scope to exclude journalism platforms and platforms where user communication is a minor ancillary feature of a platform (e.g., fitness, shopping, travel).
2. Establish platform size thresholds to place fewer obligations on smaller and non-profit platforms to avoid entrenching incumbents.
3. Require minimum standards of user reporting features and transparency for private platforms with very large user reach.
4. Clarify the definitions of harmful content as it relates to online content moderation and consider adding identity fraud to the list of harmful content in scope.
5. Narrow the requirement for platforms to take "all reasonable measures" to identify harmful content, to avoid over-censorship and ensure wrongful takedown is appealable.
6. Ensure the length of time provided for content moderation decisions can evolve through regulatory changes.
7. Limit any requirements for mandatory platform reporting to law enforcement to cases where imminent risk of serious harm is reasonably suspected, and consider narrowing to only child sexual exploitation and terrorist content.
8. Ensure platform transparency requirements are publicly accessible in a manner that respects individual privacy, and work with international allies to ensure data comparability.
9. Require larger platforms to cooperate with independent researchers, and annually review and mitigate their systemic risks.
10. Remove or significantly narrow the ability to block access to platforms for non-compliance.

01 Introduction

Social media is increasingly used by Canadians to stay up-to-date with the news, connect with friends and family, and engage in civic and political discourse. It is no wonder that many have referred to social media as our new public square. But over the past decade, it has become increasingly clear that this public square is also responsible for producing negative effects on society: hate speech and harassment that target racialized communities and other marginalized groups are rampant; disinformation abounds while helping extremist content and conspiracy theories to flourish; and real-world violence, including sexual abuse and child exploitation, is unfortunately an increasing reality.[1,2,3,4]

The impacts of this digital transformation on Canada's communications ecosystem are continuing to take shape and are still not fully understood. Indeed, new and emerging platforms continue to develop and rise, often blending public and private communication in ever-changing ways, creating a dynamic digital environment. However, what is becoming clear at this moment is that the spread of online harms through social media is real and poses significant risks to Canada's social cohesion, public safety and democracy.[5] As a result, there have been growing calls for technical and regulatory changes to mitigate these harms and rebuild our "public square."[6,7,8,9,10] At the same time, legitimate concerns have been raised regarding over-censorship of online content and that any changes may unreasonably limit our rights and freedoms, particularly the right to free expression.[11]

The Government of Canada has laid out in commendable detail what its intentions are with respect to addressing some of these online harms. The goal of this report is to respond to the Government's proposed approach to address harmful content online, and share the results of representative surveys that we conducted in Canada over the course of the last three years on these important questions.

We believe the results of regular surveys such as these, while imperfect, are important tools as we know so little about these platforms, in part because of the lack of meaningful transparency, cooperation with independent research, and regulatory action to date. Any action in this area should be informed by evidence about Canadians' experience with those harms, as well as Canadians' views on the appropriate role of government in addressing those harms. This report is intended to reflect and provide our best advice on how to do so in a manner that protects and advances Canadians' fundamental rights and freedoms.

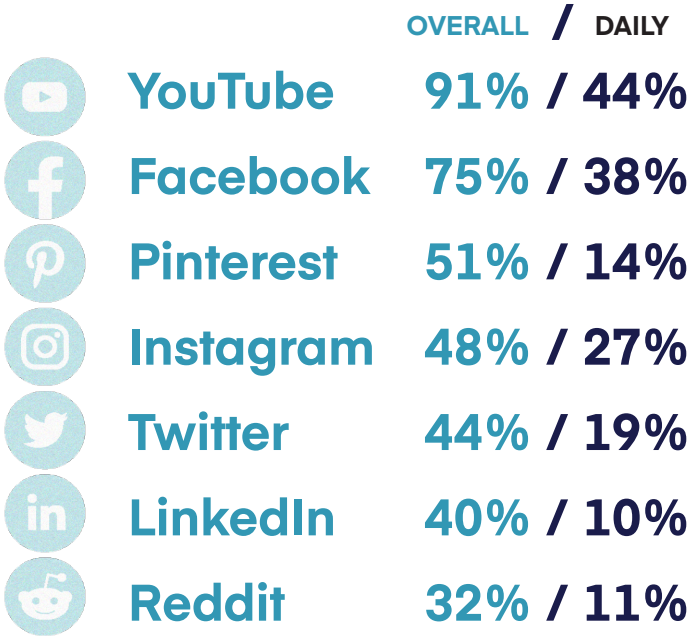
Canadians' Experience on Social Media Platforms

Three national representative surveys conducted by our team over the course of three years (2019, 2020 and 2021) provide a comprehensive picture of the social media landscape in Canada. We provide a summary here, as we believe a clear understanding of the significant and growing role of social media in Canada is foundational to designing solutions to the online harms facilitated through those platforms.

Overall Use of Platforms

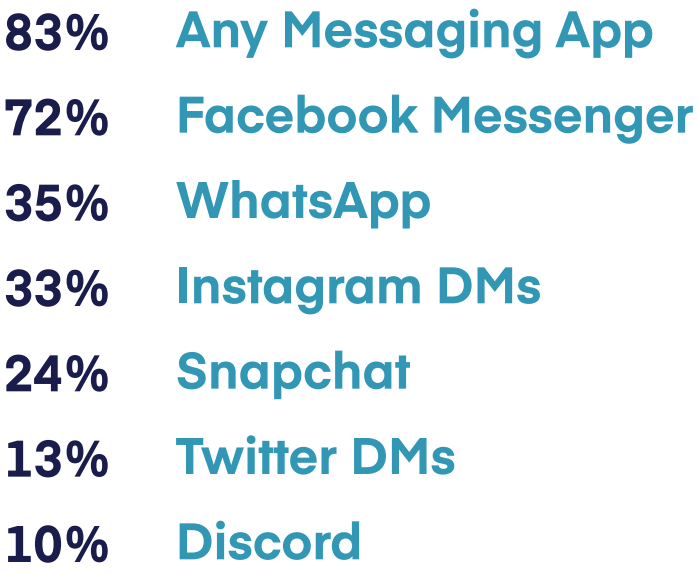
Most Canadians are using social media platforms — many every day (**Figure 1**). In fact, more than half of Canadians aged 18-29 report using YouTube (65%), Instagram (52%) and Facebook (51%) at least every day.

Canadians are also increasingly using private messaging apps to connect and share content. More than 8 in 10 report using private messaging apps in 2021, with Facebook Messenger, WhatsApp and Instagram direct messages as the most used platforms (**Figure 2**). As with public platforms, there were significant differences in the use of platforms across age groups: the majority of those aged 16-29 used direct messaging on Instagram (72%) and Snapchat (65%), compared to 15% and 8%, respectively, among those aged 45 and older.



n=3,000

Figure 1: Canadians' Use of Social Media Platforms Overall and Daily (2019)

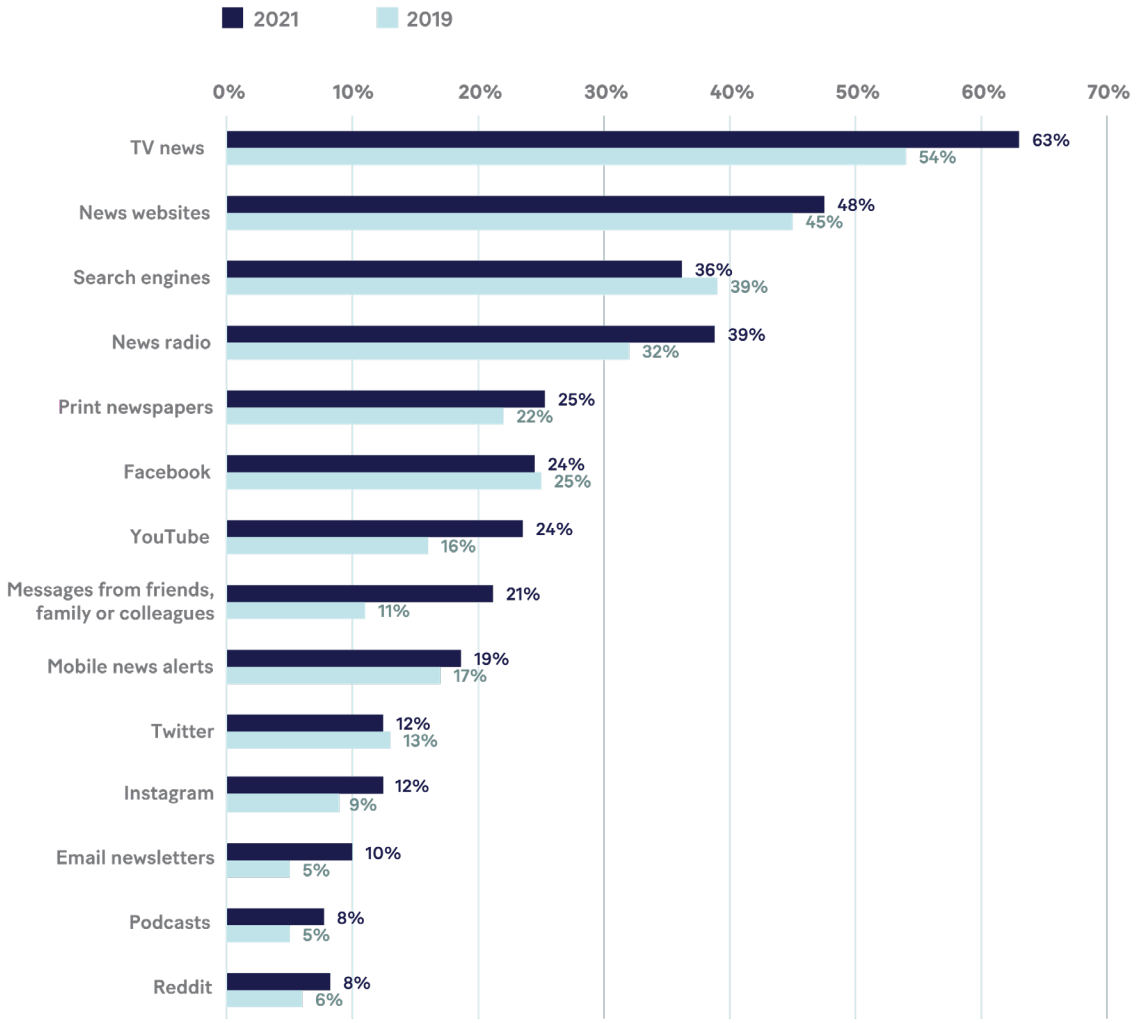


n=2,451

Figure 2: Canadians' Use of Private Messaging Apps Overall (2021)

Platforms as a News Source

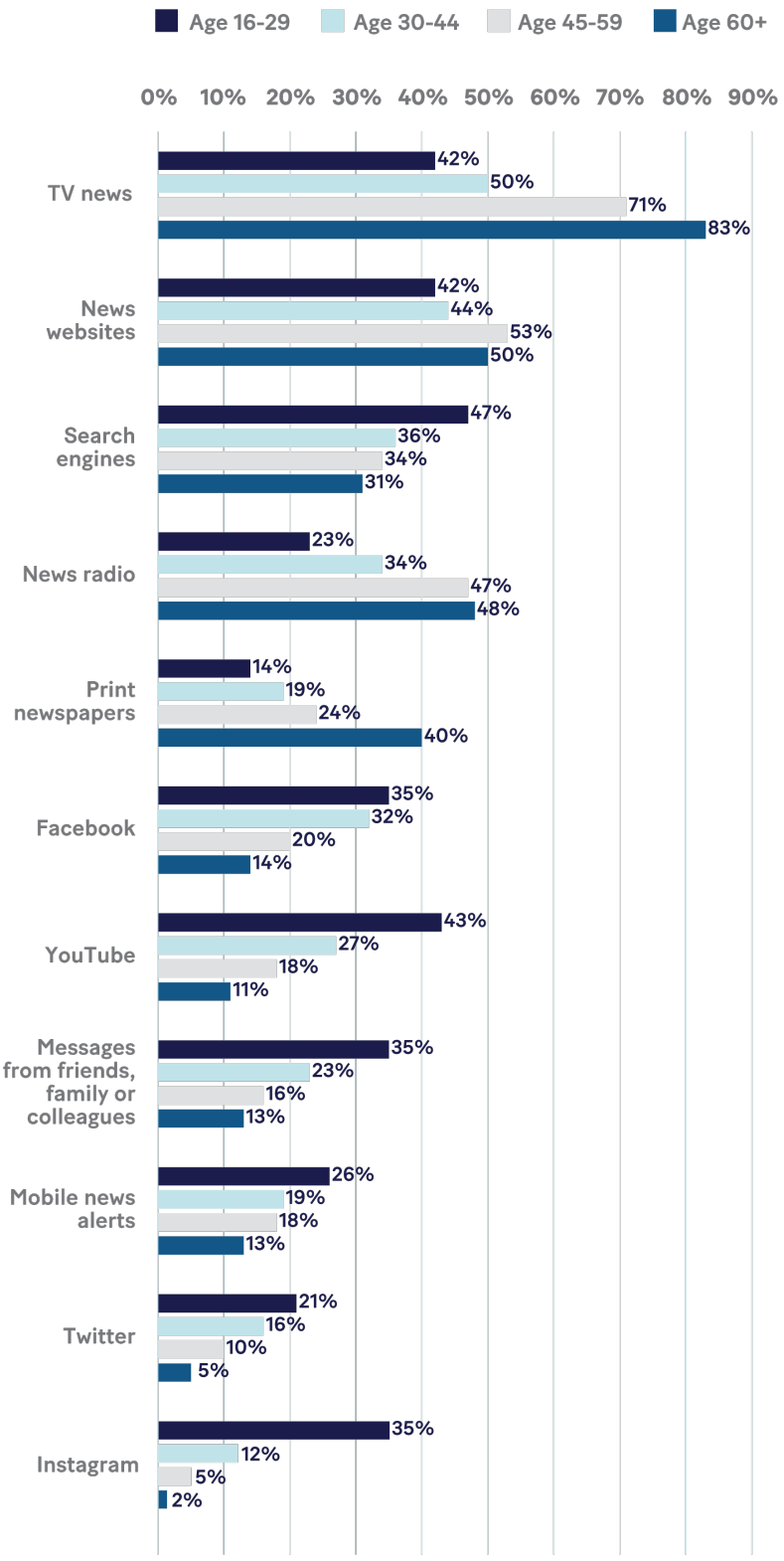
While traditional media, such as television, radio and newspapers, continue to play a large role in how Canadians consume news, **one in four report using Facebook and YouTube to stay up-to-date with the news and current events**, with 21% using messages from friends, family and colleagues (**Figure 3**). Differences by age are again significant — those aged 16-29 use YouTube (43%), Facebook (35%), Instagram (35%) and private messaging (35%) for news at greater or comparable rates than news websites (42%) or traditional media such as TV (42%) and radio (23%)(**Figure 4**).



n=2,451 (2021); 3,000 (2019)

Figure 3: Canadians' Reported Sources for News and Current Events

In addition to consuming news, a significant proportion of Canadian residents actively engage with news and politics on these platforms. According to our 2019 survey, 43% of respondents 'like' a news or political post or story on social media at least once per week, 40% join social media groups about an issue or cause, 33% share news/political stories on social media at least weekly, and 30% comment on a news/political post in their own words at least weekly.



n=436 (16-29); 636 (30-44); 654 (45-59); 725 (60+)

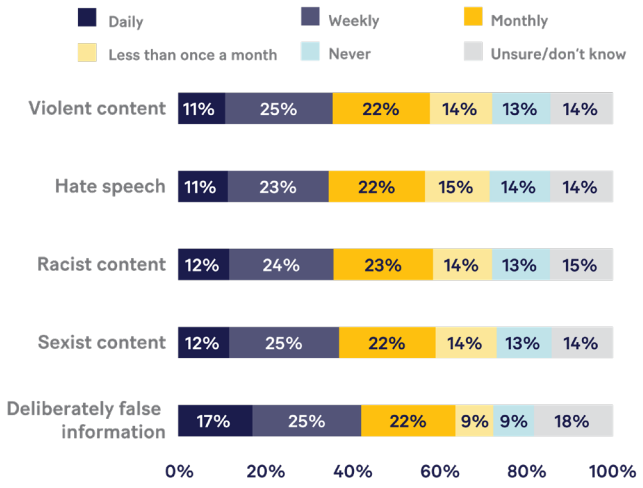
Figure 4: Canadians' Reported News Sources by Age Group (2021)

Exposure to Online Harms

Amidst this increasing use of social media platforms is a significant degree of reported exposure to harmful content.

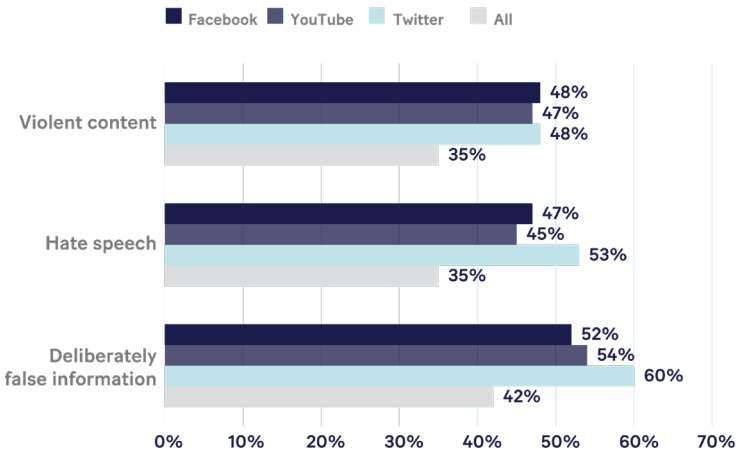
In our 2019 survey, 42% of Canadian residents reported seeing deliberately false information on online news sources, including social media platforms, at least once per week (**Figure 5**). More than one-third of respondents reported encountering other types of harmful content at least once per week, including sexist content, racist content, hate speech, and violent content, with nearly 60% reporting seeing such content at least monthly.

Further, those who used Facebook, Twitter and YouTube to stay up-to-date on news and current events were significantly more likely to report encountering online harms at least weekly (**Figure 6**).



n=3,000

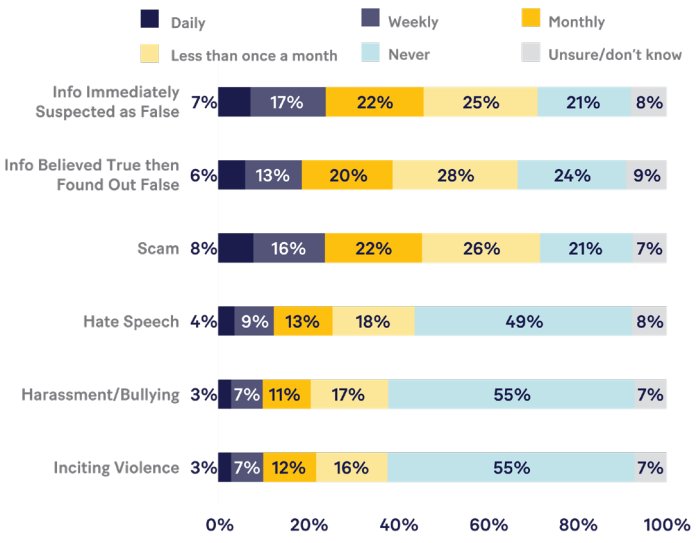
Figure 5: Canadians' Reported Exposure to Online Harms (2019)



n=754 (Facebook); 490 (YouTube); 405 (Twitter); 3,000 (all)

Figure 6: Canadians Using Social Media for News Report More Frequent Exposure to Online Harms (2019)

In our 2021 survey, we asked respondents how frequently they encountered a range of online harms specifically through private messaging apps. About half (46%) reported seeing information that they immediately suspected was false at least a few times a month; while 39% reported seeing information that they initially believed was true, but later found was at least partially false, with the same frequency. Scam or phishing messages were also reported as a relatively frequent occurrence, with 46% reporting receiving these messages at least a few times a month. Hate speech was identified by 26% of respondents at least a few times a month, with one in five encountering content that promoted or encouraged violence and harassment or bullying at least a few times a month (**Figure 7**).



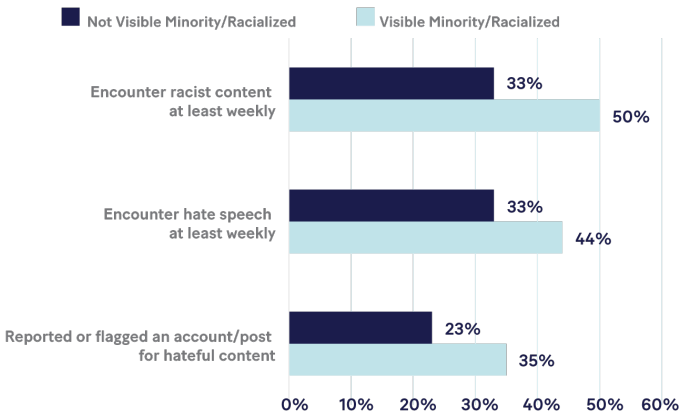
n=2.044

Figure 7: Canadians' Reported Exposure to Online Harms through Private Messaging Apps (2021)

Respondents who used private messaging apps as a regular news source were also more likely to believe a number of common false conspiracy theories about COVID-19 (see *Private Messaging, Public Harms* for more information). Sixty-three percent of believers in COVID-19 conspiracy theories received news through Facebook Messenger at least a few times a week, compared to an overall average of 47%. In turn, compared to the average Canadian, COVID-19 conspiracy believers are 34% more likely to get their news regularly from Facebook Messenger. Similarly, 39% of believers in COVID-19 conspiracy theories received news through WhatsApp at least a few times per week, compared to 22% overall, making them 77% more likely to receive their news in this way. This echoes the findings from previous research that found a correlation between consuming news on social media platforms and the likelihood to believe in COVID-19 conspiracy theories.[12]

Racialized respondents also report more frequent exposure to online harms on public and private platforms. Our 2019 data showed that those who identified as racialized were 33% more likely to report encountering hate speech and 52% more likely to report encountering racist content at least weekly, compared to non-racialized Canadians (**Figure 8**). In our 2021 survey, hate speech was reportedly received through private messaging apps by about one-quarter (26%) of respondents at least a few times a month. However, reported rates were significantly higher among Latin American (58%), Middle Eastern (44%), Southeast Asian (44%) and Black (40%) respondents. These findings strongly indicate that exposure to online harms on social media platforms is experienced more by marginalized communities.

One in four respondents in 2019 had reported harmful or fake posts or accounts. Again, those who identified as racialized were also 52% more likely to report an account or post for hateful content (35% of racialized individuals, compared to 23% of non-racialized). Likewise, 22% of respondents in 2021 reported someone for sending illegal, hateful or harassing content on a messaging app, with rates significantly higher among people of colour. Of those who did make reports about hateful content on social media, 38% ranked its effectiveness (from 1 to 9) as 7-9, 39% ranked 4-6 and 23% ranked as 1-3. These numbers were very similar for private messaging apps: when asked a similar question in 2021, 35% gave 7-9, 39% ranked 4-6 and 21% said 1-3. These assessments indicate that harmful content reporting to platforms can be an effective mechanism to mitigate harms.



n=2,450 (not); 540 (visible minority/racialized)

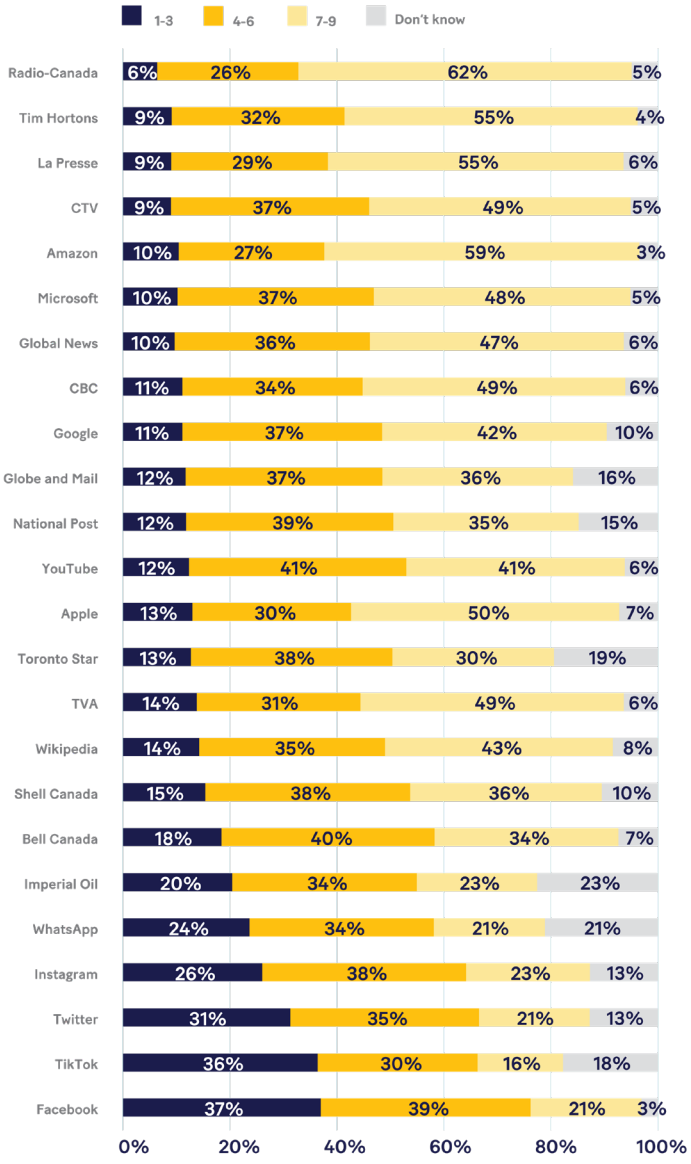
Figure 8: Racialized Canadians Report Online Harms More Frequently (2019)

Canadians' Views on Platform Regulation

Low Trust in Platforms

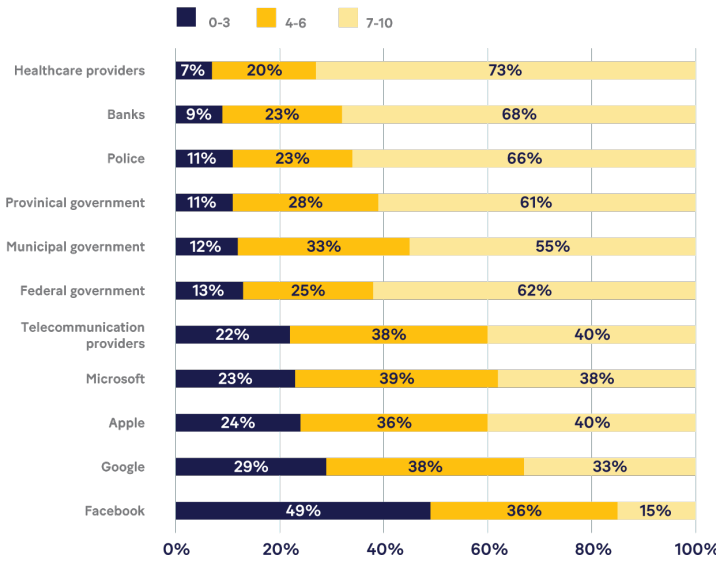
A consistent finding across all three surveys is that **Canadian residents do not trust social media platforms.** Specifically, Canadians do not believe that these companies, including Facebook, TikTok, Twitter and Instagram, make decisions in accordance with the best interest of the public. When asked to rate how much they trust various organizations on a scale from 1 to 9, respondents were less trusting of social media platforms than oil companies, telecommunication providers and news media (**Figure 9**). Our surveys found that trust in Facebook, including the other services and apps it owns, declines moderately with age, particularly among men.

We also found that big tech companies are less trusted than governments and other public and private institutions to keep personal data secure (**Figure 10**).



n=2,451

Figure 9: Canadians' Trust to Act in Public's Best Interest (on a scale from 1-9) (2021)

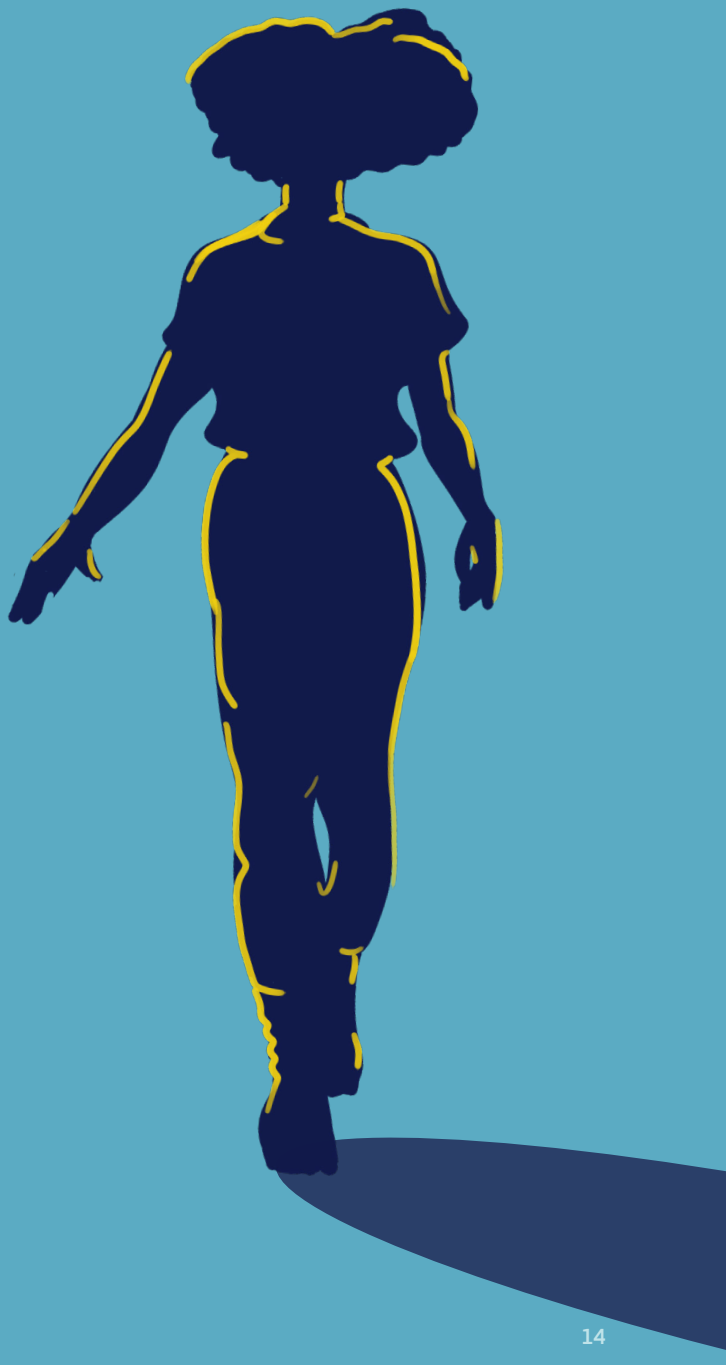


n=2,000

Figure 10: Canadians' Trust to Keep Personal Data Secure (on a scale from 0-10)(2020)

Trust levels in major social media platforms have remained relatively consistent from 2019 to 2021. For example, feelings of low trust (1-3) in Facebook increased only slightly, from 36% to 37%, and high trust (7-9) fell from 26% to 21%. Likewise, feelings of low trust in Twitter increased by two percentage points and high trust fell by one point, while reports of low trust in Instagram increased by one point and high trust fell by two points. This pattern in Canadians' degree of trust in social media companies is striking considering the rapid changes in events between these two surveys. For example, our 2019 survey was conducted prior to the COVID-19 pandemic, which saw a proliferation of false health information on social media, called by some "the biggest challenge fact-checkers have ever faced."[13] The widely recognized role played by social media in amplifying conspiracy theories and undermining public health efforts could have further damaged public trust in those companies. Moreover, the pandemic and misinformation during the 2020 U.S. election also saw unprecedented collaboration between social media platforms and fact-checking groups, with social media companies applying greater levels of content moderation and warning labels.[14] Our 2021 survey, conducted a year into the pandemic, shows that neither of these events exerted considerable influence on public perceptions of social media companies, or perhaps that the effects merely cancelled one another.

The consistent low level of public trust in social media companies among Canadians, despite the turbulent years that filled the gap between these two surveys, contributes to the overall impression indicated by the data that Canadians have a strong appetite for greater intervention.



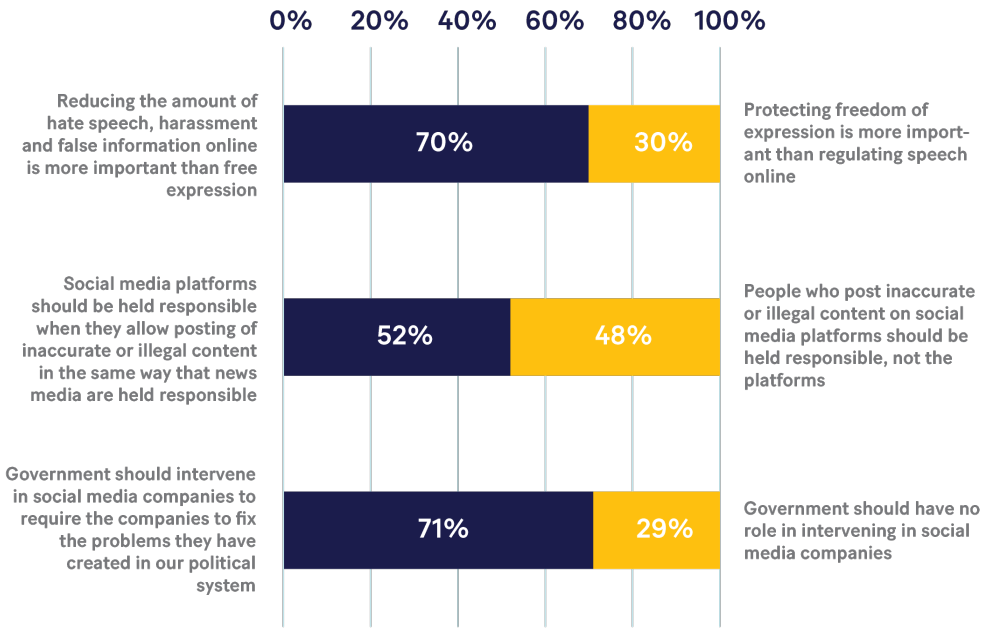
A Role for Government

Survey results from both 2019 and 2021 indicate that most Canadian residents are prepared for government intervention to address online harms. When asked in 2019, 80% of Canadians said that an increase in deliberately spread false information was a problem affecting Canadians and society in general, while 70% of Canadians said they thought the role social media plays in our political system was a similar problem.

We asked Canadian residents to choose among a series of statements which best described their perspective, and each indicated a growing willingness for platform intervention. In 2019, 63% of respondents said that reducing the amount of hate speech, harassment and false information online was more important than protecting freedom of expression.

When asked again in 2021, this number had increased to 70% (**Figure 11**). There was also a small increase in the proportion of respondents that believe that social media platforms should be held responsible when they allow posting of illegal or inaccurate content in the same way that traditional news media are held responsible, from 47% to 52%. Most strikingly, the proportion that said that the government should intervene to require social media companies to fix the problems they have created in our political system increased from 60% to 71% between the two surveys.

While Canadians' desire to see government action on this issue appears to have increased between 2019 and 2021, their opinions with respect to specific policy interventions — such as requiring platforms to delete harmful content in a timely manner or delete the accounts of users intentionally spreading false information — have remained stable during the same time period.

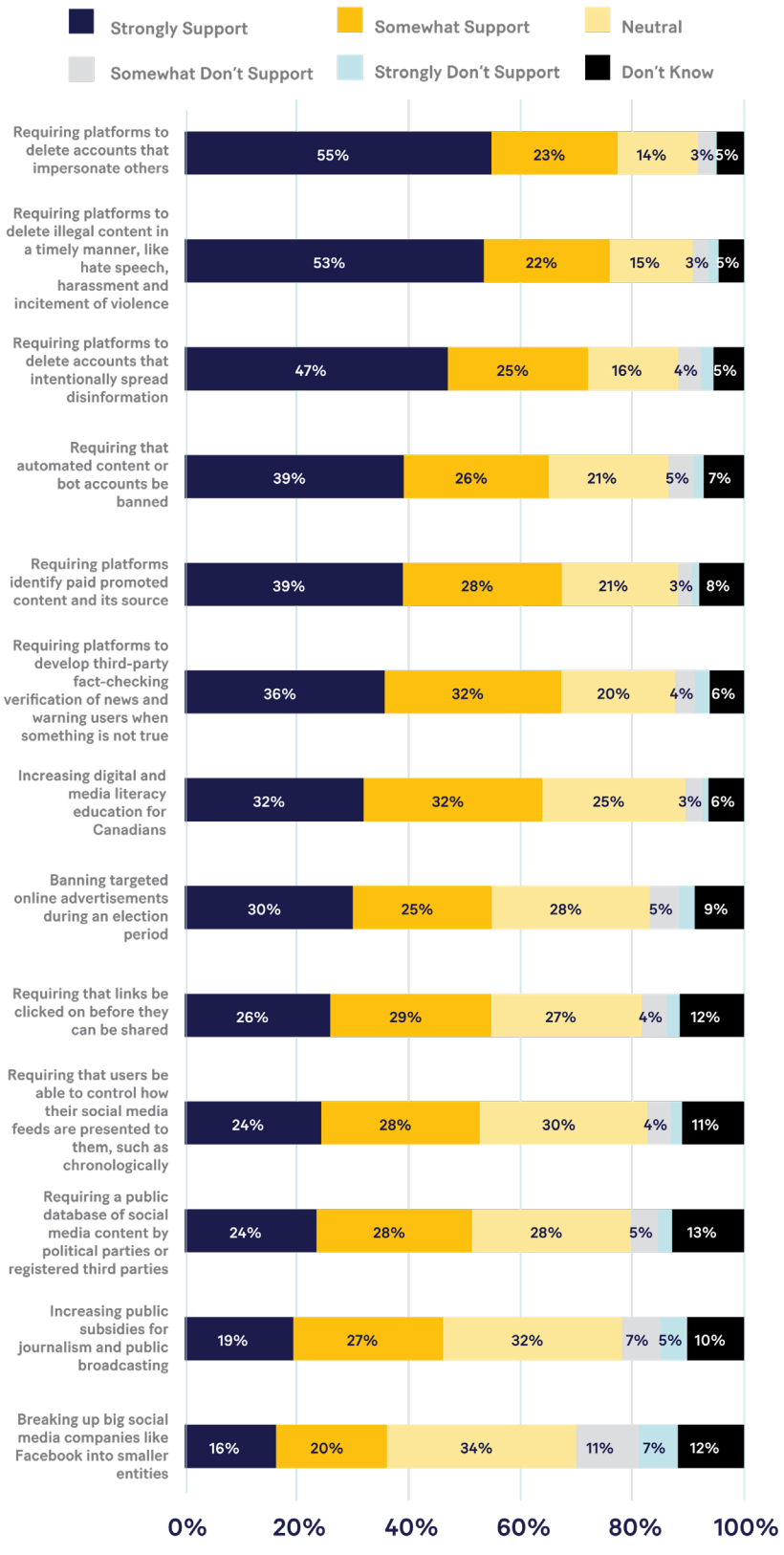


n=2,122; 2,162; 2,018

Figure 11: Canadians' Views on the Role of Government Regulation (2021)

One key takeaway from Canadians' opinions around specific policy interventions is that the policies that were most supported by Canadians were those that imposed new responsibilities for content moderation on the platforms themselves, with opposition to these policies never exceeding 10% of respondents (**Figure 12**). Policies that would address these issues in an indirect way — such as by funding digital literacy programs for Canadians or by supporting traditional media outlets as an alternative to social media — had generally lower levels of support across both surveys.

Another key point is that, while Canadians are generally supportive of various approaches that place responsibilities on platforms to moderate the content that they host, that support diminishes when the approach would result in significant changes to the service being provided. For example, when asked in 2021, 45% of respondents were less supportive of imposing new responsibilities on Facebook if those measures would cause Facebook to shut down operations in Canada; and 54% were less supportive if it would require Facebook to charge a $5 monthly fee to users (their approximate revenue per user). However, there was more willingness to impose content moderation responsibilities if it would result in Facebook needing to delay posts by a few minutes in order to carry out content moderation — only 18% of Canadians were less supportive in this case, whereas 43% were more supportive.

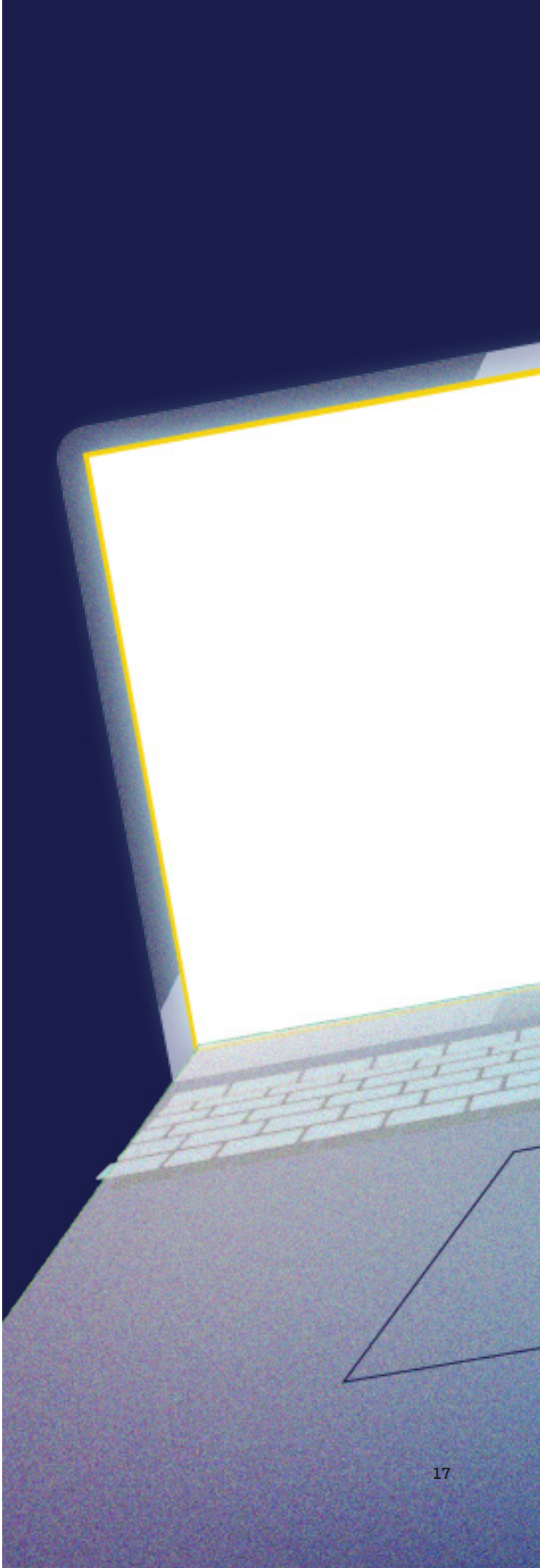


n=2,451

Figure 12: Canadians' Support for Policy Interventions (2021)

Somewhat surprisingly, the survey data did not indicate any strong relationship between trust in social media companies to act in the public's best interests and support for more stringent requirements for those companies. This is in part explained by the broad levels of support for greater action, where even those with high trust in social media companies are still supportive of intervention. Less surprisingly, those who report being victims of various online harms, such as privacy breaches and account hacks, express significantly greater support for intervention than those who have not been victims.

We believe that these results collectively paint a clear picture: **Canadians are ready for new action to address online harms while maintaining access to services that enable them to connect and share with others.** It is worth noting that, when Canadians were asked who they trusted the most to address the issue of disinformation, hateful speech and extreme views on social media, no clear consensus emerged. Twenty-eight percent indicated trust in the handling of the issue by social media platforms themselves; 22% by a government agency; 19% by the people who use social media; and 23% were not sure. We believe a takeaway from this could be that while Canadians are prepared for action, they are not sure who is best positioned to lead this work. We believe that approaches that promote direct platform responsibility while maintaining democratic and sovereign oversight and accountability for action are most likely to meet the expectations of Canadians.



Global Regulatory Approaches to Online Harms

02

Designing regulatory interventions into the communication and information ecosystem of Canadians must be done with great care, and in a manner that protects fundamental rights and freedoms. Learning the lessons from other jurisdictions around the world tackling these same issues should be top-of-mind.

In addition, an exclusive focus on content moderation can miss broader structural issues with modern online platforms, such as platform competition, personal data use by companies, and recommendation algorithms. However, it should also be acknowledged that some of these structural issues with the largest platforms are outside of meaningful influence from Canada alone. As such, Canada should try to align its regulatory efforts in the broader global context to the extent possible, and support and coordinate efforts at international governance. A theme throughout our advice that follows is to align with other jurisdictions' definitions or approaches, to enable Canada to enhance, rather than detract from, a growing democratic force on these global platforms. To this end, we provide here a summary of the most relevant efforts by allied jurisdictions to govern online platforms and harms that we will reference in our advice.

The United Kingdom

The governance model chosen by the UK, and developed in a series of consultations with stakeholder groups, is known as the 'duty of care'.[15] Under this model, the UK's communications regulator Ofcom would oversee and enforce compliance with a standards framework designed by the government, in order to "ensure that companies continue to take consistent and transparent action to keep their users safe."[16] Some commentators have argued that this duty of care required of tech platforms for online spaces is analogous to the duty of care required of property owners for their physical spaces.[17]

The scope of the UK's duty of care framework is quite broad. The framework applies to all companies whose services host user-generated content that can be accessed by users in the UK; and/or facilitate public or private online interaction between service users, one or more of whom is in the UK, as well as search engines.[18] However, this breadth is restrained by certain specific exceptions. For example, services that play a mostly 'functional' role in enabling online activity, such as ISPs, would not be subject to the framework.[19] Perhaps most notably, journalistic content, as well as user comments on that content, would be specifically exempted in an effort to protect freedom of the press.[20]

The harms proposed to be addressed by the duty of care framework are also quite broad. The framework targets criminal offences and harmful content affecting children, as well as content that can be harmful to adults even if legal.[21] Disinformation and misinformation are also included in the framework, but only in situations in which that information could cause harm to individuals.[22] Specifically out of scope are violations of intellectual property rights, data protection, fraud, consumer protection law, and cybersecurity breaches or hacking.[23]

In terms of the specific actions that companies would need to take, any company that falls within the scope of the framework would be responsible for taking action to prevent user-generated content on their platforms from causing physical or psychological harm to individuals.[24] This would involve carrying out assessments of the risks associated with their services and taking action to reduce those risks.[25]

If a user were to encounter harmful content on a platform which had an obligation under the framework to address that harm, then the user can report that harm and seek redress, such as content removal or sanctions against offending users, among other possibilities.[26]

There would also be different obligations imposed on different 'classes' of companies. These classes would be determined by their degree of reach in the public media landscape, and therefore their potential to contribute to online harms.[27] Such companies would have additional responsibilities under the framework, particularly with respect to the regulation of harmful content, even when that content is not illegal.[28] Among other differences, these companies would be required to regularly publish transparency reports in order to detail the approaches they had adopted to address online harms.[29] The government explicitly stated that it would reserve the right to impose personal liability on the managers of tech companies in the event of failure to attain the standards of care specified by the regulator.[30]

The European Union

Like the UK, the EU has been moving toward a model of regulating online harms based largely on the idea that platforms should bear more responsibility when it comes to monitoring and addressing those harms. The EU's new approach to regulating online harms began with a 2018 recommendation document published by the European Commission. Building on the feedback from these recommendations, in December 2020 the European Parliament and European Council received a legislative proposal from the European Commission titled the *Digital Services Act* (DSA).[31] It outlines a broad set of measures to regulate online platforms. What follows is a direct quotation of the stated intent of the legislation:[32]

- **measures to counter illegal goods, services or content online**, such as a mechanism for users to flag such content and for platforms to cooperate with "trusted flaggers";
- **new obligations on traceability of business users in online market places**, to help identify sellers of illegal goods;
- **effective safeguards for users**, including the possibility to challenge platforms' content moderation decisions;
- **transparency measures for online platforms on a variety of issues**, including on the algorithms used for recommendations;
- **obligations for very large platforms to prevent the misuse of their systems** by taking risk-based action and by independent audits of their risk management systems;
- **access for researchers to key data of the largest platforms**, in order to understand how online risks evolve;
- **oversight structure to address the complexity of the online space:** EU countries will have the primary role, supported by a new European Board for Digital Services; for very large platforms, enhanced supervision and enforcement by the Commission.

While the new law upholds existing legal protections for platforms in terms of not being liable for the content they host in the EU, it also introduces a new responsibility to remove illegal content in a "timely, diligent and objective manner" once identified.[33] As with the UK approach, the proposed EU model would operate using a tiered system, with larger platforms being subject to more stringent requirements.[34] For example, platforms with over 45 million users would be required to abide by a range of new restrictions, such as:

- Risk management obligations;
- External audits to assess the degree of risk for harm posed by the platform's activities;
- Transparency around recommendation systems related to user content;
- Obligations to share data with researchers to help understand online harms; and
- Cooperation with authorities in the event of crises.[35]

For the first time in the EU, the law would specify that companies that fail to comply with these obligations would be subject to fines of up to 6% on their annual profits.[36]

Germany

In June 2017, the German Federal Parliament adopted the *Network Enforcement Act* or the NetzDG, which came into effect in October 2017.[37] It should be noted that the law was adopted in a fast-tracked legislative process and was subject to significant criticism from civil society organizations.[38] The law aimed to reduce hate speech, criminally punishable disinformation and other harmful content on social media. Under the *Act*, social networks with at least two million members in Germany are subject to multiple obligations.[39] Most notably, the law requires social networks to remove or block access to content that is "manifestly unlawful" within 24 hours of receiving complaints unless provided otherwise by law enforcement.[40] Social networks must also remove or block access to all other simply "unlawful" content generally within seven days of receiving a complaint, with certain exceptions involving whether the factual allegation is true or false or if the decision will be decided upon by an approved self-regulatory institution.[41,42] The law also requires social networks to maintain effective and transparent organizational procedures for handling complaints about unlawful content available to users.[43] Platforms designed to enable "individual communication or the dissemination of specific content" are specifically exempt from the law.[44]

Australia

Australia's eSafety Commissioner is dedicated exclusively to promoting online safety and enforcing compliance with online content moderation requirements under the *Enhancing Online Safety Act* (EOSA).[45] The eSafety Commissioner was initially focused on promoting online safety for children; however, in 2017, the *Act* was amended to expand the scope of its functions to include safeguarding against risks of online harm for all Australians.[46]

Under the EOSA, the eSafety Commissioner is responsible for monitoring online platforms' compliance with safety requirements related to the cyber bullying of children[47] and non-consensual sharing of intimate images.[48] The EOSA requires social media service providers to include a provision that "prohibits end-users from posting cyber bullying material" in its terms of use and a complaints framework under which users can report and request the removal of harmful material.[49] Under the EOSA, if a material is considered a cyber bullying act targeting an Australian child, and the social media service does not remove the material within 48 hours of a complaint, the eSafety Commissioner has the power to request the removal of the material within 48 hours of a written notice.[50] Moreover, the Commissioner has the power to issue an "end-user notice," under which the person posting the cyber-bullying material is required to remove it and refrain from posting harmful content in the future.[51] Civil penalties are enforced for failure to comply with the removal notice.[52] The Commissioner can also invoke these regulatory powers to enforce the removal of intimate images shared without the subject's consent.[53]

The eSafety Commissioner also has powers under the *Broadcasting Services Act* (BSA) and the *Criminal Code Act* (CCA). Under the BSA, the Commissioner can investigate complaints and enforce the removal of "prohibited content" as defined by the Classification Board—the government body responsible for classifying films, publications and online content, issuing age restrictions, and implementing censorship guidelines.[54, 55] The Commissioner can issue a "removal notice" to a host of the illegal content in Australia,[56] or a "blocking notice" to a local Internet Service Provider to prevent or restrict access to illegal content hosted outside of Australia.[57] The Classification Board's definition of illegal content includes child abuse material, content promoting terrorism, and incitements of violence.[58] Under the CCA, the Commissioner can request the removal of "abhorrent violent material," defined as content that records or streams terrorist acts, violence or kidnapping,[59] and requires the internet, content, or hosting service provider to inform the Australian Federal Police "within a reasonable time after becoming aware of the existence of the material."[60]

In June 2021, the Australian government enacted the *Online Safety Act* to once again expand the Commissioner's powers.[61] The new legislation, which will come into effect in January 2022, expands the Commissioner's cyberbullying regulations to include adult-targeted cyber harms[62] and requires the removal of cyberbullying material from a wide range of online services, not just social media sites.[63] The new *Act* also grants the Commissioner enhanced powers to rapidly block websites that host abhorrent violent material in real time[64] and reduces the time frame required for service providers to comply with removal notices from 48 to 24 hours.[65]

Our Advice on the Government's Current Proposal

03

We appreciate the opportunity to respond to the Government's proposed approach in detail. The following provides our recommendations to strengthen and clarify the proposal, to ensure it best meets its objective of supporting a safe, inclusive and open online environment while protecting and advancing fundamental rights and freedoms.

Platforms in Scope

The Government's proposed definition of an "Online Communication Service" (OCS) to be in scope for this new law is "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" and excludes "services that enable persons to engage only in private communications." The proposal provides regulatory power to the federal government to further specify the definition of an OCS, such as including or excluding a category of services, and the meaning of the term "private communications". The proposal's briefing material provides examples of in-scope platforms, such as Facebook, YouTube, TikTok, Instagram and Twitter, while also providing examples of what it intends to exempt, including telecommunications providers, as well as private messaging, fitness, ridesharing and travel platforms.

Platforms' Primary Purpose

The Government should consider clarifying its intentions by further defining what is meant by a service's "primary purpose" to ensure the very broad definition of user communication does not capture what it does not intend and that regulatory exemptions are not applied inconsistently. The Government may consider adopting language from the EU's proposed *Digital Services Act* (DSA), which clarifies platforms should not be in scope "where the dissemination to the public is merely a minor and purely ancillary feature of another service and that feature cannot, for objective technical reasons, be used without that other, principal service, and the integration of that feature is not a means to circumvent the applicability of the rules of this Regulation applicable to online platforms."[66] Such language would clarify intentions with respect to services such as fitness, shopping or travel platforms.

Likewise, language from Germany's NetzDG and the UK's online harms bill aiming to protect freedom of the press and platforms exclusively dedicated to journalism could be adopted to specifically exclude "platforms offering journalistic or editorial content, the responsibility for which lies with the service provider itself."[67] The EU's DSA preamble also specifies "the comments section of an online newspaper" as being exempt as an ancillary feature.[68] The UK and Australia also both specifically exclude closed internal business platforms, which could be considered.

Platform Size

As currently drafted, it appears that no size or user reach thresholds are proposed to exempt smaller platforms from the law. The Government should consider mirroring the platform size thresholds established in other jurisdictions that have been carefully crafted to prevent only entrenched incumbents with the resources to meet sophisticated regulatory requirements, as well as mitigate the risk of smaller platforms withdrawing their services, which could undermine freedom of expression and access to information.

The EU's DSA requires platforms of all sizes to have the basic ability for users to report illegal content, but exempts small platforms without significant reach from recourse and appeal mechanisms for such content, as well as transparency requirements.[69]

These are currently defined as enterprises employing fewer than 50 people with an annual balance sheet below EUR 10 million ($15 million CAD) and fewer than 45 million average monthly active users in the EU (approx. 10% of population).[70] Germany's NetzDG has a threshold of two million registered users in Germany (approx. 2% of population), and is also limited to platforms that have "profit-making purposes" to exempt non-profit and public enterprises.[71] Australia's eSafety Commissioner can designate "large" platforms with legally-binding requirements while enabling others to participate on a cooperative basis; it has designated only three to date: Facebook, Instagram and YouTube.[72]

Canada could potentially model this after similar size thresholds it established in the *Canada Elections Act* for online advertising transparency, which map closely to the EU's DSA thresholds, and defines platforms in scope as those visited or used by Canadian users over the prior 12 months by an average of 3 million per month in English; 1 million per month in French; or 100,000 times per month in another language.

Private Communication

The proposed exemption for "services that enable persons to engage only in private communications" captures an important and extremely complex element of this proposed law that potentially requires further clarification.

Many platforms offer both public and private communication functions, and clarification that blended platforms will have their different functions treated differently would help clarify scope. For example, Instagram is in scope, but its direct message functions are not intended to be. Wording similar to the EU's DSA could be adopted: "Where some of the services provided by a provider are covered by this Regulation whilst others are not, or where the services provided by a provider are covered by different sections of this Regulation, the relevant provisions of this Regulation should apply only in respect of those services that fall within their scope."[73]

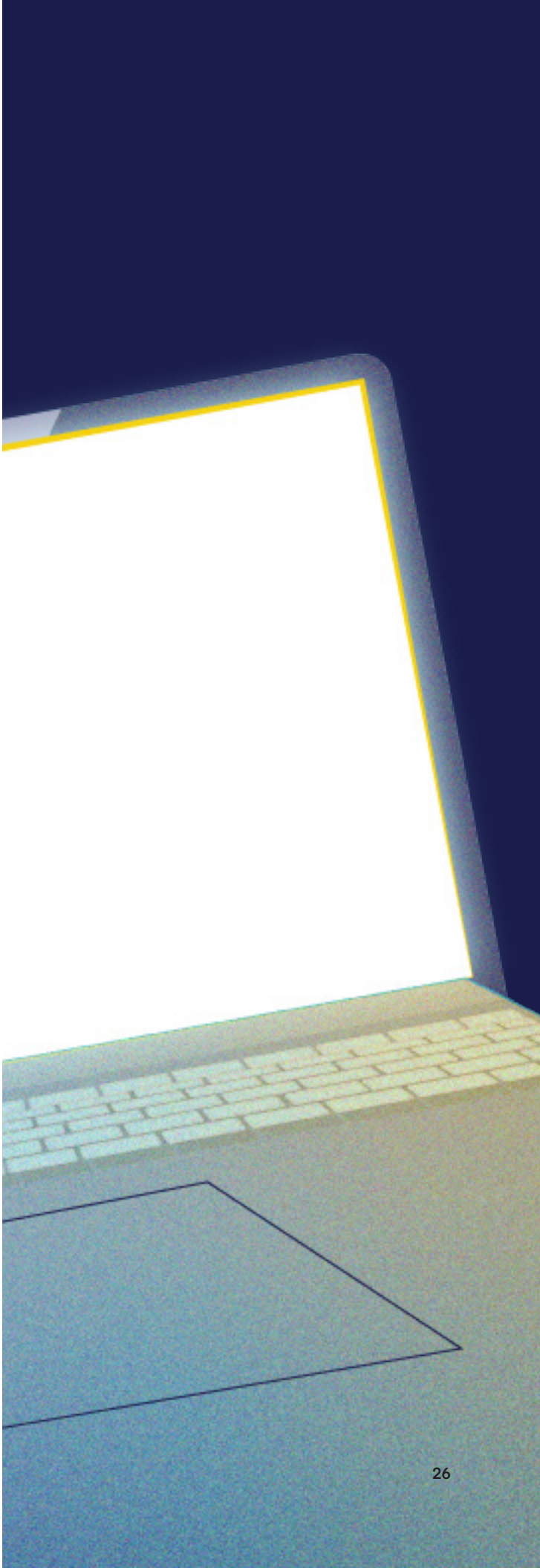
However, the distinction between public and private communications on many online platforms is not always clear. For example, is the proposal's intention to capture posts on social media that are private to only its followers (e.g., a private Facebook or Instagram profile or group)? If not, is it rational that regulatory action would be prioritized for content viewed by say a dozen people on a public profile over content viewed by thousands or millions on a private profile or group? To use another example, if a public Instagram profile posts a story to its close friends (a feature that limits access to a user-defined list of followers), is that post now private communication?

The EU's DSA attempts to make this public/private distinction through its definition of "dissemination to the public" as "making information available, at the request of the recipient of the service who provided the information, to a potentially unlimited number of third parties", thereby exempting private profiles or groups.[74] This has come under some scrutiny from experts; for example, Caroline Cauffman and Catalina Goanta ask: "should there not be a critical number of 'friends' or 'group members' that leads to the loss of confidentiality protection and to the same treatment as offers to or information shared with the public in general?"[75] The EU's *DSA* is, however, not a consensus approach. Germany

exempts only "individual communication;"[76] the UK includes private profiles and messaging, but excludes emails and SMS messages;[77] while Australia's approach includes all private communication.[78]

While thresholds at the individual level may be problematic, there may be no way to avoid establishing a threshold by what is considered private. For example, closed groups on Telegram can have up to 200,000 users, which surely stretches the meaning of "private" communication; however, one could envision all iMessage or Instagram message groups (each capped at 32 users) being considered private. However, we think it makes sense that this be left to regulations to evolve over time, in consultation with experts and Canadians. One could also imagine the thresholds being different for different types of harms, for example a lower threshold for intimate images than other content.

Under the EU's DSA, however, large private platforms that do not meet the "dissemination to the public" requirement are still required to have user-friendly mechanisms to electronically report content that users consider illegal, as well as provide notice to users if it removes or disables content, including the reasons for its decision and available redress possibilities. The law also still requires annual reports outlining their content moderation activities, including the number of user reports by type of alleged illegal content, action taken, and average time needed for taking action, as well as proactive measures taken as a result of the application and enforcement of their terms and conditions. Finally, when enabled by national laws, EU member states would also be able to order hosting services to remove illegal content.



The Government should craft the legislation to enable a similar approach, in which private platforms of a significant size are still subject to minimum requirements, such as user notice-and-action mechanisms and transparency requirements. This would better enable harm reduction, promote greater understanding of online harms, and would mitigate the risk of an incentive for companies to create more closed or private platforms as a means of sidestepping content moderation obligations. For a more detailed examination of potential regulatory mechanisms for online harms on private messaging apps, see *Private Messaging, Public Harms*.

Key Recommendations:

1. **Clarify the online platforms in scope** to exclude journalism platforms and platforms where user communication is a minor ancillary feature of a platform (e.g., fitness, shopping, travel).

2. **Establish platform size thresholds** to place fewer obligations on smaller and non-profit platforms, to avoid entrenching incumbents.

3. **Require minimum standards** of user reporting features and transparency for private platforms with very large user reach.

Harmful Content in Scope

The Government's proposal specifies five types of harmful platform content for which moderation will be regulated:

1. Terrorist content;
2. Content that incites violence;
3. Hate speech;
4. Non-consensual sharing of intimate images; and
5. Child sexual exploitation content.

These five categories are all worthy of regulatory action, though each is also very different, and the new regulator will need to develop expertise in each to meaningfully understand and implement the distinct categories of content.

The proposal refers to using Criminal Code definitions of this content "adapted to a regulatory context." The Government should engage experts and stakeholders further in these definitions, given the very different contexts. For example, the proposed definition of content that incites violence is that which "actively encourages or threatens violence and which is likely to result in violence"; clarification may be needed as to whether coordination or recruitment to violence in the absence of encouragement or threat is in scope, and whether this includes self-harm. Darryl Carmichael and Emily Laidlaw also raise important questions about the definition of terrorist content in their submission.[79]

We also think a sixth category of harmful content is worthy of consideration: identity fraud. Online impersonation is among the most common online harms, is often a poor fit for the criminal justice system given the scale and speed of the platforms, and also has a clear Criminal Code definition ("fraudulently personates another person, living or dead, with intent to: gain advantage for themselves or another person; obtain any property or an interest in any property; or cause disadvantage to the person being personated or another person").[80] As an example, Facebook and Instagram reported in their most recent global transparency report that it actioned 1.7 billion fake accounts, compared to a combined total of 143 million accounts for hate speech, violent content, child endangerment and terrorism.[81] YouTube also reports impersonation as a more frequent reason for channel removal than promotion of violence or terrorism.[82] This could also enable the regulator to address an emerging threat to our democracy: synthetic media and deepfakes.

Key Recommendations:

4. **Clarify the definitions of harmful content** as they relate to online content moderation, and consider adding identity fraud to the list of harmful content in scope.

Content Moderation Requirements

The Government's proposal places obligations on platforms to "take all reasonable measures, which can include the use of automated systems, to identify harmful content that is communicated on its OCS and that is accessible to persons in Canada, and to make that harmful content inaccessible to persons in Canada." It also provides that platforms must take measures to ensure that the implementation and operation of the content moderation procedures, practices, rules and systems put in place do not result in differential treatment of any group based on a prohibited ground of discrimination within the meaning of the *Canadian Human Rights Act* and in accordance with regulations. It also requires that content flagged by any person in Canada as harmful be addressed "expeditiously," which it indicates will be defined as 24 hours from the content being flagged or another period prescribed in regulations, including the ability to set different times for different types of harmful content. It requires a notice of decision to the user, the ability to compel a prompt review of the decision, and user notice of the reconsideration, including the ability to appeal to the new Digital Recourse Council.

This proposed wording regarding "all reasonable measures" may be construed by platforms as a requirement to proactively monitor or filter all content accessible to persons in Canada, even from non-Canadians. This would have far-reaching implications. The UN's Special Rapporteur for Freedom of Opinion and Expression has criticized such general monitoring obligations as "inconsistent with the right to privacy and likely to amount to pre-publication censorship."[83] We believe this provision needs to be reworked to be more narrow in scope, or at the very least, provisions in the EU's DSA should also be adapted, such as: "Nothing in this Regulation should be construed as an imposition of a general monitoring obligation or active fact-finding obligation, or as a general obligation for providers to take proactive measures to relation to illegal content"[84] and "The removal or disabling of access should be undertaken in the observance of the principle of freedom of expression."[85] Proposals have been advanced in the EU to clarify that monitoring obligations should only be enabled in specific cases, such as blocking content that is identical to content that has previously been declared unlawful.[86] The UK's proposal also moves to limit proactive monitoring only to child sexual abuse and terrorist content; and requires all platforms to protect users' right to freedom of expression within the law when deciding on, and implementing, safety policies and procedures.[87]

The proposed measures to ensure that monitoring obligations do not result in differential treatment or discrimination are positive features that somewhat mitigate risks. Consideration could be given to provide explicit authority to the new regulator to conduct independent audits of differential treatment. Cynthia Khoo's *Deplatforming Misogyny* provides excellent insights into ways to achieve substantive equality with respect to content moderation, or the notion that people in different positions may have to be treated differently to achieve true equality, that should also be considered.[88]

The current proposal is also asymmetrical with respect to user content wrongfully removed compared to harmful content that remains accessible; there is no regulated

ability to appeal content removed or service suspended incorrectly under the platform's terms and conditions. The ability to appeal decisions to remove content through platform measures or automated systems is left at the discretion of the platforms, whereas illegal content that remains accessible is subject to a series of reporting and appeal mechanisms. This asymmetry is likely to incentivize more aggressive proactive filtering, with implications for freedom of expression. To rebalance these incentives, the Government should also consider a complementary platform user notice and appeal mechanism for wrongful takedown or suspension of service and timely redress as is articulated in EU's DSA Article 17.3. It could also consider requiring that users receive notices regarding when their content has been filtered or moderated through automated means, and the right to request that the platform's review of this decision be conducted through non-automated means.

Based on evidence to date, the 24-hour requirement for content decisions is likely to lead to over-censorship of non-harmful content.[89] Even the Germany model only requires 24 hours for "manifestly unlawful" content and up to seven days to review other content.[90] The EU's DSA also has a mechanism for "trusted flaggers" to have the content flagged prioritized for moderation, which Canada may wish to model.[91] We acknowledge the proposal already allows for regulatory flexibility in this regard, though we would advise explicit reference to 24 hours be removed. Instead, we would suggest the new regulator develop more precise requirements around the meaning of "expeditiously" in consultation with experts and stakeholders, and reflecting the reality of how this new law is implemented in Canada, including the effectiveness of the Recourse Council in providing guidance to platforms and improving democratic oversight of takedown decisions over time. The current proposal's structure may enable this, but the Government may also wish to consider focusing timely removal on content with more reach for certain types of harmful content, or setting standards that aim to reduce the overall number of Canadians who see illegal content.

Finally, we would advise that a provision be explicitly added to ensure the user reporting and appeal mechanisms for illegal content are free of charge to the user throughout the process.

Key Recommendations:

5. **Narrow the requirement for platforms to take "all reasonable measures"** to identify harmful content, to avoid over-censorship and ensure wrongful takedown is appealable.

6. **Ensure the length of time provided for content moderation decisions can evolve through regulatory changes.**

Law Enforcement Reporting Requirements

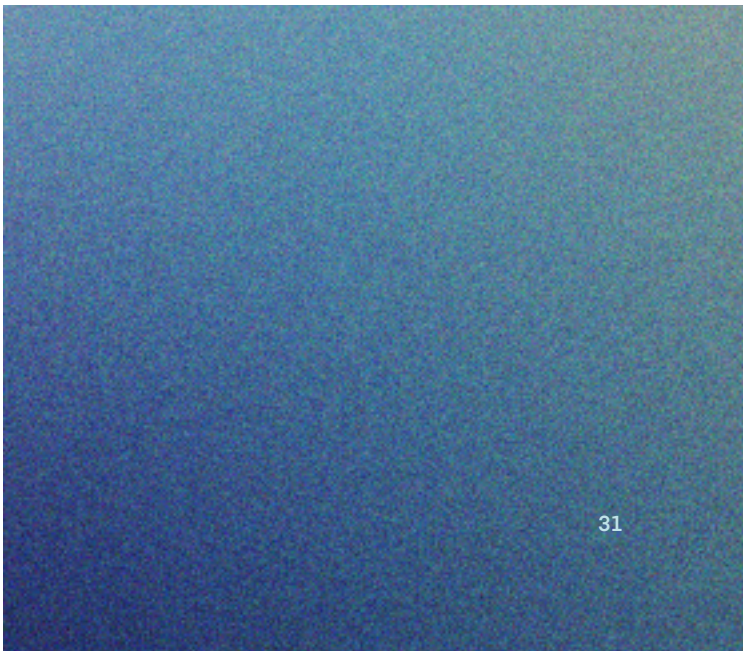
The Government describes its proposal for mandatory law enforcement reporting as an 'interplay' between law enforcement and CSIS to identify public safety threats and prevent violence. The discussion guide acknowledges the limitations of content removal, suggesting that it may be counterproductive by potentially pushing threat actors to encrypted platforms and away from the visibility and reach of law enforcement, thus producing more unmoderated harmful content. Although the potential of user migration to encrypted services is certainly a real phenomenon, discussed further in our report *Private Messaging, Public Harms*, the Government's proposal does not give due credence to the challenges and potential harms of mandatory reporting to law enforcement operating in conjunction with automated content monitoring and removal.

The Government proposes two potential models for requiring platforms to report harmful content to law enforcement:

a. when the platform has reasonable grounds to suspect the content reflects an imminent risk of serious harm to any person or to property; or
b. when the platform believes content is illegal within the prescribed criminal offences of the five harmful content categories.

The first approach is consistent with the EU's DSA and many platforms' existing practices. The second approach intertwines content moderation with mandatory reporting, is too discretionary for platforms to meaningfully carry out without creating additional harm and should be abandoned. This approach risks disproportionately impacting racialized, religious minorities, LGBTQ2S+ people and other marginalized groups who, as Suzy Dunn has identified, are particularly at risk of having their content removed either deliberately through individuals who maliciously flag content, or through content moderation systems that discriminate.[92] Such groups could increasingly find themselves caught in a content removal-policing nexus where their posts would be forwarded to law enforcement or CSIS for investigation, potentially unbeknownst even to the users themselves. The unintended consequences to free expression are not merely hypothetical. Google has challenged Germany's recent and similar proposal for violating fundamental human rights.[93] In addition, such an approach could undermine the equality-driven purpose of this legislation, causing more harm to racialized and marginalized groups.

Even under the Government's first more limited proposal, regulatory clarity should be provided regarding the definitions of "reasonable grounds to suspect" and "serious harm" or else this proposal still risks undermining freedom of expression and the right to be secure against unreasonable search and seizure.



For example, the Electronic Frontier Foundation suggested in the EU context that user reports alone should not be sufficient to trigger obligations for reporting.[94] Further, the subtext of this section seems focused on child sexual exploitation and terrorist content. We cannot foresee a scenario where automated filtering and reporting to law enforcement without victim consent of potential acts of hate speech or intimate images does not create more harm than good. We would urge considering limiting this section to be specific to the harmful content it intends to capture.

Key Recommendations:

7. **Limit any requirements for mandatory platform reporting to law enforcement** to cases where imminent risk of serious harm is reasonably suspected, and consider narrowing to only child sexual exploitation and terrorist content.

Platform Transparency Requirements

The Government's proposal requires platforms to produce reports on a scheduled basis to the new regulator, providing Canada-specific data about several important elements, including:

- the volume and type of harmful content;
- the volume and type of content moderated;
- the volume and type of content that was accessible to persons in Canada in violation of their community guidelines; and
- platform's content moderation procedures, systems, resources and activities.

These are important transparency provisions, and we would recommend that the legislation clarify these reports should be publicly accessible in a manner that respects individual privacy. The proposed provision regarding content "in violation of their community guidelines" is well-intentioned, though we think it would be clearer to replace "community guidelines" with "terms and conditions", as "community guidelines" is a term only used by some platforms. Splitting "the volume and type of content moderated" between automated and human moderation would also strengthen this provision.

Mandated and audited transparency is among the most powerful platform governance tools that governments have. It would also be beneficial for these requirements to be built in cooperation with international allies, to ensure data can be compared to other countries to the extent possible, as well as leave regulatory flexibility for the new regulator to add additional transparency requirements that advance their overall mandate in consultation with experts, allies and stakeholders.

Key Recommendations:

8. **Ensure platform transparency requirements are publicly accessible** in a manner that respects individual privacy and work with international allies to ensure data comparability.

New Regulators

The Government proposes to create a new regulatory body in the Digital Safety Commission to administer and enforce these requirements, as well as engage in partnerships, education outreach activities and research. It also proposes the establishment of the Digital Recourse Council to review and issue content moderation decisions, as well as an Advisory Board to support and advise the Commission and the Recourse Council. The Commissioner will have broad inspection and enforcement powers, including the ability to recommend fines of up to the greater of 3% of global revenue or $10 million to the body responsible for administering privacy violations, or to refer fines to prosecutors of up to 5% of global revenue or $25 million.

The design of the regulatory and oversight bodies seems fit for purpose, though of course the devil will be in the details of how these new bodies are implemented, adequately resourced, and use their authorities. For example, there may be considerable complaint volume at the Recourse Council, so we wonder if it would be best to leave the maximum number of members (currently prescribed as five) as flexible in the regulations.

It is worth noting that the functions of the Digital Safety Commission seem deliberately broader than just the five prescribed types of harmful content, which is positive and will hopefully allow the Commission to engage in partnerships and research on broader issues of digital safety not yet in scope for regulatory action (e.g., disinformation harmful to public safety, synthetic media or automated/bot content labelling, ad transparency, doxing, algorithmic transparency, etc.). It would also seem to enable the Digital Safety Commissioner to engage in partnerships with civil society and international allies; one could envision investigations or partnerships with European and Australian digital commissioners on matters of joint interest.

The broad inspection powers proposed for the Commission may satisfy this, but the Government may consider adopting the more specific provisions in the EU's DSA Article 31 to compel very large platforms (defined as more than 10% of the population or 450 million monthly active users) to cooperate with independent research, including providing data to vetted academic researchers, and specific data security and confidentiality requirements, including provisions relating to trade secrets.[95] These EU provisions are world-leading and the Government should ensure Canadian researchers can similarly engage in a better understanding of online platforms.

The Government should also consider mirroring the EU's DSA Articles 26 and 27, that requires very large platforms to annually review and put in place mitigation measures for their systemic risks in: the dissemination of illegal content; any negative effects for the exercise of the fundamental rights and freedoms; and intentional manipulation of their service with effects on the protection of public health, minors, civic discourse, electoral processes and public security.[96] These provisions enable their Commission to produce an annual report with the most prominent and recurrent systemic risks and best practices for mitigation. Like in the financial services industry, compelling companies to review their potential risks to society can be a powerful tool for mitigation.

Key Recommendations:

9. **Require larger platforms to cooperate** with independent researchers, and annually review and mitigate their systemic risks.

Website Blocking

The Government's proposal also enables the Commissioner to apply to the Federal Court for an order to block access to a platform, in whole or in part, that demonstrates persistent non-compliance with orders regarding child sexual exploitation or terrorist content. Site-blocking powers have understandably been met with significant criticism by internet service providers and civil society organizations for censorship, impairing individual liberty, and potentially exacerbating harm against the marginalized populations that the law in part seeks to protect.[97] This proposed power requiring judicial authorization is quite prescribed, though it is worth noting that Germany and the EU's approach do not contain this power relying on monetary penalties;[98, 99] and Australia only has site-blocking powers for time-limited viral distribution of terrorist content in response to the Christchurch Call.[100] The Government may also wish to review the UK's proposed approach, which enables blocking of 'ancillary' services, such as payment processing, advertising services and search results for a site, as a means to pressure compliance before outright blocking.[101, 102]

It is not clear that this measure is necessary, effective and proportionate, given that major platforms increasingly appear to be in compliance with removal requirements for unlawful content. For example, in an evaluation of the European Commission's Code of Conduct on countering illegal hate speech online, companies removed on average 70% of illegal hate speech notified to them, with companies meeting the set target of reviewing the majority of notifications within 24 hours, reaching an average of more than 81% (and figures for both have steadily increased with each evaluation).[103] Recognizing that the existing provision allows for site blocking to be "in part", many of the platforms proposed to be in scope host far more legal expression than illegal, so enabling site-blocking only of platforms where the majority or significant proportion of content is non-compliant could also be a way to narrow scope and mitigate Charter scrutiny.

Key Recommendations:

10. **Remove or significantly narrow the ability** to block access to platforms for non-compliance.

About the Authors

Sam Andrey is the Director of Policy & Research at the Ryerson Leadership Lab. Sam has led applied research and public policy development for the past decade, including the design, execution and knowledge mobilization of surveys, focus groups, interviews, randomized controlled trials and cross-sectional observational studies. He also teaches about public leadership and advocacy at Ryerson University and George Brown College. He previously served as Chief of Staff and Director of Policy to Ontario's Minister of Education, in the Ontario Public Service, and in not-for-profit organizations advancing equity in education and student financial assistance reform. Sam has an Executive Certificate in Public Leadership from Harvard's John F. Kennedy School of Government and a BSc from the University of Waterloo.

Alexander Rand is interested in disinformation and the ways in which new technologies influence online political discourse. He has worked as a Public Policy Researcher at the Centre for the Future of Democracy, and at the London-based AI think tank Future Advocacy. He holds a Master of Public Policy from Cambridge University, where he conducted statistical analyses of online partisanship and disinformation in the Canadian context, as well as a BA from McGill University in Economics and Music Technology.

Mohammed (Joe) Masoodi is a Senior Policy Analyst in the Ryerson Leadership and Cybersecure Policy Exchange. Joe has been conducting research and policy analysis at the intersections of surveillance, digital technologies, security and human rights for over six years. He has conducted research at the Surveillance Studies Centre at Queen's University and the Canadian Forces College. He holds an MA in war studies from the Royal Military College of Canada, an MA in sociology from Queen's University, and has studied sociology as a PhD candidate from Queen's University, specializing in digital media, information and surveillance.

Karim Bardeesy is the Co-Founder and Executive Director of the Ryerson Leadership Lab. Karim is a public service leader who has worked in progressively senior roles in public policy, politics, journalism and academia in Toronto and the United States since 2001. He is also a board member of The Atmospheric Fund and Corporate Knights, Inc., a member of the Banff Forum, and a founding faculty member of Maytree Policy School. Karim was previously Deputy Principal Secretary for the Premier of Ontario, the Honourable Kathleen Wynne, and served as Executive Director of Policy for Premiers Wynne and Dalton McGuinty. He has worked as a journalist, an editorial writer at *The Globe and Mail*, and as an editorial assistant at *Slate* magazine. Karim holds a Master in Public Policy from Harvard's John F. Kennedy School of Government.

Methodology

Three anonymous online surveys were conducted with random samples of research study panelists in Canada to better understand Canadians' views on online harms and regulation:

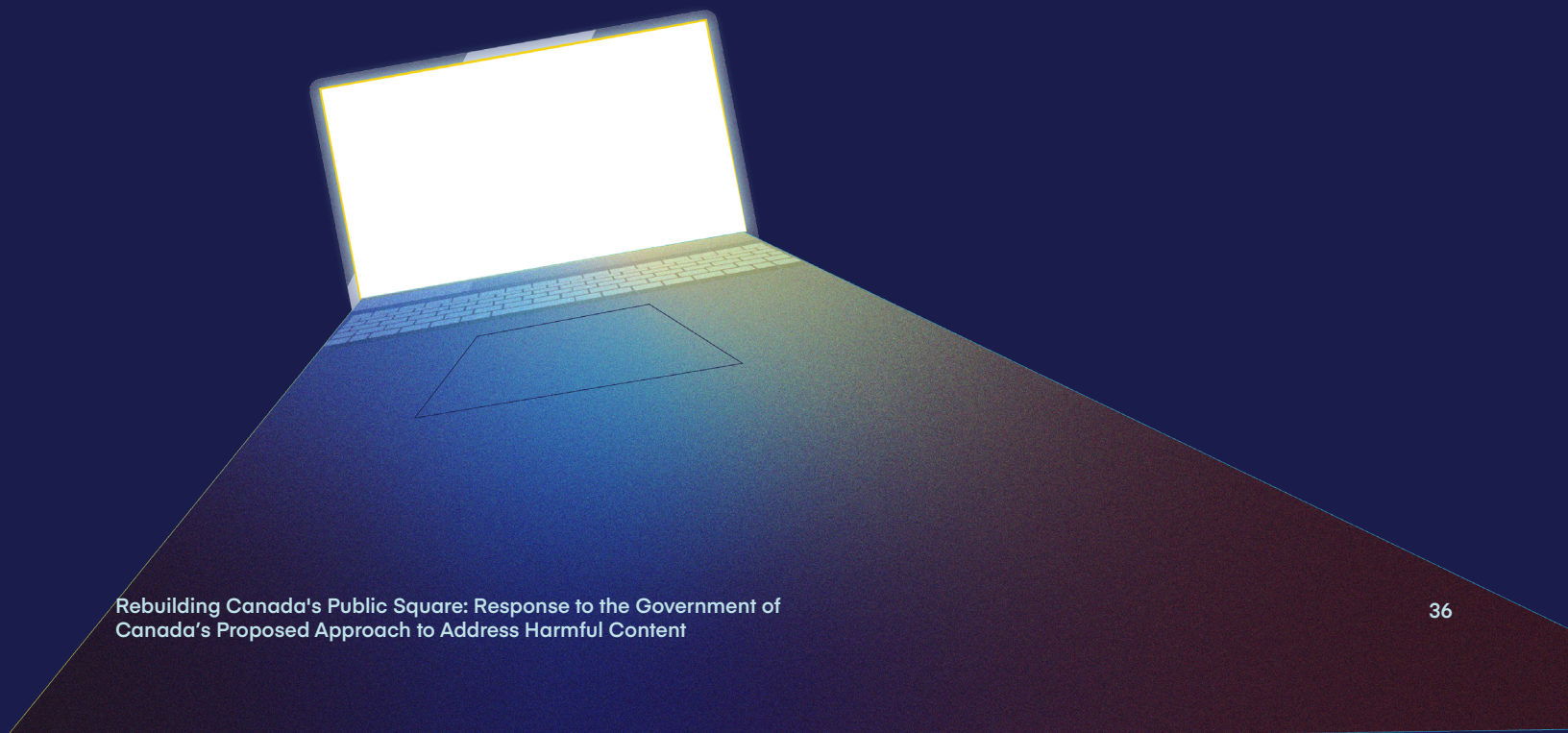
2019: 3,000 Canadian residents aged 18 and over from August 1-7, 2019 conducted by Abacus Data from a set of panels based on the Lucid exchange platform and Leger panel.

2020: 2,000 Canadian residents aged 18 and over from May 14-22, 2020 conducted by Pollara Strategic Insights using the AskingCanadians panel.

2021: 2,500 Canadian residents aged 16 and over from March 17-22, 2021 conducted by Abacus Data from a set of panels based on the Lucid exchange platform.

Response quotas were set and the data were weighted according to the latest Canadian census data to ensure that the sample matched Canada's population according to age, gender, educational attainment and region. Totals may not add up to 100 due to rounding. As a guideline, a probability sample of this size would yield results accurate to ± 2 percentage points, 19 times out of 20.

The 2019 survey was supported by the Governments of Canada and Ontario. The 2020 survey was supported by RBC. The 2021 survey was supported by RBC and the Government of Canada.



Survey Questions

Figure 1: Which best describes how often do you do the following?

- About once an hour
- A few times a day
- Daily
- A couple times a week
- Once a week
- Once every few weeks
- A few times a year
- I don't do this/use this service
- Unsure/don't know

a. Watch news on TV
b. Listen to the news on the radio
c. Listen to a podcast
d. Visit a news website
e. Open a news app on your mobile device
f. Read something on Wikipedia
g. Read a print newspaper
h. Read a print magazine
i. Use Google Search
j. Use Google News
k. Use Facebook Newsfeed
l. Use Facebook Messenger
m. Use LinkedIn
n. Use Instagram
o. Use Pinterest
p. Use Reddit
q. Use Snapchat
r. Use Tumblr
s. Use Twitter
t. Use WeChat
u. Use WhatsApp
v. Watch something on YouTube

Figure 2: Have you used any of the following messaging apps in the last year?

- Yes
- No
- Don't know or prefer not to say

a. WhatsApp
b. Facebook Messenger
c. WeChat/Weixin
d. Telegram
e. Signal
f. Snapchat
g. Direct messages on Instagram

[Viber/imo/Weibo]*
[LINE/Discord/Clubhouse]*
[QQ/Direct messages on Twitter/ Direct messages on TikTok]*
* Survey respondents split into three and each asked one of each

Figures 3 and 4: Which of the following do you use to stay up-to-date with the news or current events? (select all that apply)

a. An email newsletter
b. Messages from friends, family or colleagues (e.g., text, WhatsApp, Facebook Messenger)
c. TV
d. Radio
e. Podcasts
f. Print newspapers
g. Print magazines
h. News websites
i. News alerts on my mobile device
j. Search engine (e.g., Google, Bing, etc.)
k. Facebook
l. Instagram
m. Reddit
n. LinkedIn
o. Twitter
p. YouTube

Figure 5: Thinking of any online sources for news or political information (websites, Facebook, Twitter, Instagram, news apps, etc.), how often do you think you encounter the following?

- Every day
- A few times a week
- Once a week
- A few times a month
- Once a month
- Less than once a month
- Never
- Unsure/don't know

a. Deliberately false information
b. Accidentally false information
c. Deliberately misleading or biased information
d. Accidentally misleading or biased information
e. Deliberately inflammatory or divisive content
f. Accidentally inflammatory or divisive content
g. Something you would consider hate speech
h. Something you would consider racist content
i. Something you would consider sexist content
j. Something you would consider violent content

Figure 6: Proportion of respondents to question in Figure 5 who chose Facebook/YouTube/Twitter in question to Figures 3 and 4

Figure 7: [only asked to those who indicated using at least one private messaging app] Thinking about all the messaging apps you use, how often do you think you receive messages, including links, images or videos, that contain what you would consider:

- Every day
- A few times a week
- A few times a month
- A few times a year
- Never
- Don't know or prefer not to say

a. Information about the news or current events that you immediately suspect to be false
b. Information about the news or current events that you believe to be true and later find out is at least partly false
c. Hate speech that wilfully promotes hatred against an identifiable group
d. Harassment or bullying
e. A scam (e.g., phishing to provide personal information or to download malware)
f. Promoting or encouraging violence

Figure 8: Question from Figure 5, in addition to: Which of the following actions have you done?

- Yes
- No
- I think so
- Unsure

a. Fact checked a post about the news on a different site
b. Blocked or muted an account or phrase
c. Reported or flagged an account or post for hateful content
d. Reported or flagged an account for being fake/ automated
e. Reported or flagged a post for being false
f. Downloaded an ad-blocker or privacy app to track data being shared with third parties
g. Change the settings on each app/platform so that your profile is less public

Do you consider yourself a member of a visible minority / racialized community?

Figure 9: Below we have a list of specific companies or services. We want you to think about whether each of these make decisions that you consider to be in the best interest of the public, and others that you consider to care less about what is in the best interest of the public.
On a scale of 1-9, where 1 means you have no trust at all and 9 means you have a high degree of trust, how do you feel about each of the following when it comes to trusting them to act in the best interest of the public:

a. Amazon
b. Apple
c. Bell Canada
d. CBC / Radio-Canada [split outside/inside of Quebec; n=1,854/597]
e. National Post / La Presse [split outside/inside of Quebec]
f. CTV/TVA [split outside/inside of Quebec]
g. Toronto Star / Le Journal de Montreal [split outside/inside of Quebec]
h. Facebook
i. Globe and Mail
j. Global News
k. Google (Alphabet Inc.)
l. Imperial Oil / Shell Canada [split sample; n=1,168/1,283]
m. Instagram
n. Microsoft
o. Tim Hortons
p. TikTok
q. Twitter
r. Wikipedia
s. WhatsApp
t. YouTube

Figure 10: Below is a list of organizations that often handle data about Canadians. How much do you trust these organizations to keep your personal data secure? Rate on a scale of 0 to 10, with 0 being "Do not trust at all" and 10 being "Completely trust".

a. The federal government
b. Your provincial government
c. Your municipal government
d. Health care providers (e.g., hospitals, doctors)
e. The police
f. Banks
g. Telecommunication providers (e.g., Bell, Rogers, Telus)
h. Apple
i. Facebook
j. Google
k. Microsoft

Figure 11: Please indicate which of the following best describes your perspective:

a. Protecting freedom of expression is more important than regulating speech online.
b. Reducing the amount of hate speech, harassment and false information online is more important than free expression.

a. Social media platforms should be held responsible when they allow posting of inaccurate or illegal content in the same way that news media are held responsible.
b. People who post inaccurate or illegal content on social media platforms should be held responsible, not the platforms.

a. Government should intervene in social media companies to require the companies to fix the problems they have created in our political system.
b. Government should have no role in intervening in social media companies.

Figure 12: There have been a number of actions proposed to address some of the challenges with social media today. For each of the following, would you say you strongly support, somewhat support, are neutral, somewhat don't support or strongly don't support:

a. Requiring platforms to delete accounts that intentionally spread disinformation
b. Requiring platforms to delete accounts that impersonate others
c. Requiring platforms to delete illegal content in a timely manner, like hate speech, harassment and incitement of violence
d. Requiring platforms to develop third-party fact-checking verification of news and warning users when something is not true
e. Requiring that users be able to control how their social media feeds are presented to them, such as chronologically
f. Increasing digital and media literacy education for Canadians
g. Increasing public subsidies for journalism and public broadcasting
h. Requiring that automated content or bot accounts be banned
i. Requiring platforms identify paid promoted content and its source
j. Requiring a public database of social media content by political parties or registered third parties
k. Banning targeted online advertisements during an election period
l. Requiring that links be clicked on before they can be shared
m. Breaking up big social media companies like Facebook into smaller entities

If the Canadian government were to introduce some of these actions and they had the following impacts on Facebook's operations (which includes Facebook, Messenger, Instagram and WhatsApp), would this make you much more, somewhat more, somewhat less, or much less supportive of the government getting involved? If it would have no impact, please say so.

- Much more supportive
- Somewhat more supportive
- No impact
- Somewhat less supportive
- Much less supportive
- Don't know or prefer not to say

a. Facebook shutting down operations in Canada
b. Facebook charging a monthly $5 fee in order to operate in Canada
c. Facebook delaying your posts by a few minutes to review the content

References

1 Canadian Race Relations Foundation. (2021, January 25). *Poll demonstrates support for strong social media regulations to prevent online hate and racism*. https://www.crrf-fcrr.ca/en/news-a-events/media-releases/item/27349-poll-demonstrates-support-for-strong-social-media-regulations-to-prevent-online-hate-and-racism

2 Garneau, K. & Zossou, C. (2021, February 2). Misinformation during the COVID-19 pandemic. *Statistics Canada*. https://www150.statcan.gc.ca/n1/pub/45-28-0001/2021001/article/00003-eng.htm

3 Owen, T. et al. (2021, January 5). Understanding Vaccine Hesitancy in Canada: attitudes, beliefs, and the information ecosystem. *Media Ecosystem Observatory*. https://files.cargocollective.com/c745315/meo_vaccine_hesistancy.pdf

4 Humphreys, A. (2020, July 3). Man who allegedly crashed truck through Rideau Hall's gate with four guns is soldier troubled by COVID conspiracies. *National Post*. https://nationalpost.com/news/man-who-allegedly-crashed-truck-through-rideau-halls-gate-with-four-guns-is-soldier-troubled-by-covid-conspiracies

5 Canadian Commission on Democratic Expression. Harms Reduction: A Six-Step Program to Protect Democratic Expression Online. (2021, January). *Public Policy Forum*. https://ppforum.ca/wp-content/uploads/2021/01/CanadianCommissionOnDemocraticExpression-PPF-JAN2021-EN.pdf

6 It's time to block hate online. (n.d.). *Blockhate*. https://blockhate.ca

7 Canadian Coalition to End Online Hate. (n.d.). *Centre for Israel and Jewish Affairs*. https://www.cija.ca

8 Housefather, A. (2019, June). Taking action to end online hate. Report of the Standing Committee on Justice and Human Rights. 42nd Parliament, 1st Session. House of Commons. https://www.ourcommons.ca/Content/Committee/421/JUST/Reports/RP10581008/justrp29/justrp29-e.pdf

9 Laidlaw, E. (2015, August). Regulating Speech in Cyberspace: Gatekeepers, Human Rights, and Corporate Responsibility. *Cambridge University Press*. https://www.cambridge.org/core/books/regulating-speech-in-cyberspace/7A1E83C71D0D67D13756594BE3726687

10 Examining the impact of digital technologies on Canadian society. (n.d.). *Democratic Expression*. https://www.commissioncanada.ca

11 Carmichael, D. & Laidlaw, E. (2021, September 13). The Federal Government's Proposal to Address Online Harms: Explanation and Critique. *University of Calgary Faculty of Law*. https://ablawg.ca/2021/09/13/the-federal-governments-proposal-to-address-online-harms-explanation-and-critique/

12 Stecula, D., Pickup, M., & van der Linden, C. (2020, July 6). Who believes in COVID-19 conspiracies and why it matters. *Policy Options*. https://policyoptions.irpp.org/magazines/july-2020/who-believes-in-covid-19-%20conspiracies-and-why-it-matters/

13 Suárez, E. (2020, March 31). How fact-checkers are fighting coronavirus misinformation worldwide. *Reuters Institute*. https://reutersinstitute.politics.ox.ac.uk/risj-review/how-fact-checkers-are-fighting-coronavirus-misinformation-worldwide

14 Edelman, G. (2020, December 27). Better Than Nothing: A Look at Content Moderation in 2020. *Wired*. https://www.wired.com/story/content-moderation-2020-better-than-nothing/

15 *Draft Online Harms Bill 2021*, pt. 2. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf

16 Department for Digital, Culture, Media & Sport. (2020, December 15). Online Harms White Paper: Full government response to the consultation. *Government of the United Kingdom*. https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response

17 Woods, L., & Perrin, W. (2019, April). Online harm reduction – a statutory duty of care and regulator. *Carnegie UK Trust*. https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2019/04/06084627/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf

18 Online Harms White Paper, 2020

19 Ibid.

20 Ibid.

21 *Draft Online Harms Bill 2021*, pt. 2 c. 6.

22 Online Harms White Paper, 2020

23 Ibid

24 Ibid

25 *Draft Online Harms Bill 2021*, pt. 2 c. 2

26 Ibid

27 Online Harms White Paper, 2020

28 Ibid

29 *Draft Online Harms Bill 2021*, pt. 3 c. 1.

30 Online Harms White Paper, 2020

31 Europe fit for the Digital Age: Commission proposes new rules for digital platforms. (2020, December 15). Press Release. *European Commission*. https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2347

32 Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. COM/2020/825 final. https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN

33 *Digital Services Act 2020*, c. III s. 1 a. 11. https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN

34 *Digital Services Act 2020*, c. III s. 4 a. 25.

35 *Digital Services Act 2020*, c. III s. 4 a. 26-33

36 *Digital Services Act 2020*, c. IV s. 3 a. 59

37 Germany: Flawed Social Media Law. (2018, February 14). *Human Rights Watch*. https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

38 Kettemann, M. (2019, May). Follow-up to the comparative study on "blocking, filtering and take-down of illegal internet content". Leibniz-Institute for *Media Research & Hans-Bredow-Institut*. https://rm.coe.int/dgi-2019-update-chapter-germany-study-on-blocking-and-filtering/168097ac51

39 *Network Enforcement Act 2017*, a. 1 s. 1. https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=BBE250F3A09040DF3193FEC171E78E06.2_cid297?__blob=publicationFile&v=2

40 *Network Enforcement Act 2017*, s. 3.

41 *Network Enforcement Act 2017*, s. 3.

42 Kettemann, M. (2019, May). Follow-up to the comparative study on "blocking, filtering and take-down of illegal internet content". Leibniz-Institute for *Media Research & Hans-Bredow-Institute*, 4. https://rm.coe.int/dgi-2019-update-chapter-germany-study-on-blocking-and-filtering/168097ac51

43 *Network Enforcement Act 2017*, a. 1 s. 3

44 *Network Enforcement Act 2017*, a. 1 s. 1

45 eSafety Commissioner. (n.d.) Government of Australia. *Our Legislative Functions*. https://www.esafety.gov.au/about-us/who-we-are/our-legislative-functions

46 Ibid.

47 *Enhancing Online Safety Act 2015* (Cth) pt 3. https://www.legislation.gov.au/Details/C2018C00356/Html/Text#_Toc524097331

48 *Enhancing Online Safety Act 2015* (Cth) pt 5A. https://www.legislation.gov.au/Details/C2018C00356/Html/Text#_Toc524097331

49 *Enhancing Online Safety Act 2015* (Cth) pt 4. https://www.legislation.gov.au/Details/C2018C00356/Html/Text#_Toc524097331

50 *Enhancing Online Safety Act 2015* (Cth) pt 4 div 2. https://www.legislation.gov.au/Details/C2018C00356/Html/Text#_Toc524097331

51 *Enhancing Online Safety Act 2015* (Cth) pt 5. https://www.legislation.gov.au/Details/C2018C00356/Html/Text#_Toc524097331

52 *Enhancing Online Safety Act 2015* (Cth) pt 4 div 2-3. https://www.legislation.gov.au/Details/C2018C00356/Html/Text#_Toc524097331

53 *Enhancing Online Safety Act 2015* (Cth) pt 5A – Non-Consensual Sharing of Intimate Images. https://www.legislation.gov.au/Details/C2018C00356/Html/Text#_Toc524097331

54 *Broadcasting Services Act 1992* (Cth) pt 4 div 1-2. https://www.legislation.gov.au/Details/C2021C00042/Html/Volume_2#_Toc62734656

55 Department of Infrastructure, Transport, Regional Development, and Communications. Government of Australia. *Classification Ratings.* https://www.classification.gov.au/classification-ratings/what-do-ratings-mean

56 *Online Safety Act 2021* (Cth) pt 9. https://parlinfo.aph.gov.au/parlInfo/download/legislation/bills/r6680_first-reps/toc_pdf/21022b01.pdf;fileType=application%2Fpdf

57 *Online Safety Act 2021* (Cth) pt 8 div 3. https://parlinfo.aph.gov.au/parlInfo/download/legislation/bills/r6680_first-reps/toc_pdf/21022b01.pdf;fileType=application%2Fpdf

58 eSafety Commissioner. (n.d.) Government of Australia. *Illegal Harmful Content*. https://www.esafety.gov.au/key-issues/Illegal-harmful-content

59 *Criminal Code Act 1995* (Cth) pt 10.6 div 474.32 https://www.legislation.gov.au/Details/C2021C00066/Html/Volume_2#_Toc63237533

60 *Criminal Code Act 1995* (Cth) pt 10.6 div 474.33 https://www.legislation.gov.au/Details/C2021C00066/Html/Volume_2#_Toc63237533

61 eSafety Commissioner. (n.d.) Government of Australia. *Our Legislative Functions*. https://www.esafety.gov.au/about-us/who-we-are/our-legislative-functions

62 *Online Safety Act 2021* (Cth) pt 3 div. 4. https://parlinfo.aph.gov.au/parlInfo/download/legislation/bills/r6680_first-reps/toc_pdf/21022b01.pdf;fileType=application%2Fpdf

63 eSafety Commissioner. (n.d.) Government of Australia. *Our Legislative Functions*. https://www.esafety.gov.au/about-us/who-we-are/our-legislative-functions

64 *Online Safety Act 2021* (Cth) pt 8 div 3. https://parlinfo.aph.gov.au/parlInfo/download/legislation/bills/r6680_first-reps/toc_pdf/21022b01.pdf;fileType=application%2Fpdf

65 *Online Safety Act 2021* (Cth) pt 9. https://parlinfo.aph.gov.au/parlInfo/download/legislation/bills/r6680_first-reps/toc_pdf/21022b01.pdf;fileType=application%2Fpdf

66 *Digital Services Act 2020*, s.13

67 *Network Enforcement Act 2017*, a. 1 s. 1.

68 *Digital Services Act 2020*, s.13

69 *Digital Services Act 2020*, Explanatory Memorandum, s.2

70 Ibid

71 *Network Enforcement Act 2017*, a. 1 s. 1 -2.

72 eSafety Commissioner. (n.d.). Government of Australia. *Working with social media.* https://www.esafety.gov.au/about-us/consultation-cooperation/working-with-social-media

73 *Digital Services Act 2020*, s.15

74 *Digital Services Act 2020*, c.1 a.2 s.(i)

75 Cauffman, C. & Giants, C. (2021, April 15). A New Order: The Digital Services Act and Consumer Protection. *Cambridge University Press.* https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/new-order-the-digital-services-act-and-consumer-protection/8E34BA8A209C61C42A1E7ADB6BB904B1

76 *Network Enforcement Act 2017*, a. 1 s. 1.

77 *Draft Online Harms Bill 2021*, p. 2 c. 6

78 *Online Safety Act 2021* (Cth) pt 1 s.13A

79 Carmichael, D. & Laidlaw, E. (2021, September 13). The Federal Government's Proposal to Address Online Harms: Explanation and Critique. *University of Calgary Faculty of Law.* http://ablawg.ca/wp-content/uploads/2021/09/Blog_DC_EL_Federal_Online_Harms_Proposal.pdf

80 Criminal Code, RSC 1985, c C-46, s. 403(1)

81 Community Standards Enforcement Report. (n.d.). Transparency Center (Q2 2021). *Facebook* https://transparency.fb.com/data/community-standards-enforcement/

82 YouTube Community Guidelines Enforcement. (n.d.) Transparency Report. *Google*. https://transparencyreport.google.com/youtube-policy/removals?hl=en

83 Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. (2018, April 6). Presented to HRC, 38th session. *United Nations Human Rights Office of the High Commissioner*. https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ContentRegulation.aspx

84 *Digital Services Act 2020*, s. 28

85 *Digital Services Act 2020*, s. 22

86 European Parliament. (2021, June 21). *Committee on Culture and Education*. https://www.europarl.europa.eu/doceo/document/CULT-PA-693943_EN.pdf

87 *Draft Online Safety Bill 2021*, p.4 c.4

88 Khoo, C. (2021). Deplatforming Misogyny: Report on Platform Liability for Technology-Facilitated Gender-Based Violence. *Women's Legal Education and Action Fund*. https://www.leaf.ca/publication/deplatforming-misogyny/

89 Keller, D. (2021, February 8). Empirical evidence of over-removal by internet companies under intermediary liability laws: an updated list. *The Centre for Internet and Society.* http://cyberlaw.stanford.edu/blog/2021/02/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws

90 *Network Enforcement Act 2017*, a.1 s.3

91 *Digital Services Act 2020*, s.3 a.19

92 Geist, M. (Host). (2021, August 23). "They Just Seemed Not to Listen to Any of Us" – Cynthia Khoo on the Canadian Government's Online Harms Consultation (No. 99). [Audio podcast episode]. *In Law Bytes*. https://www.michaelgeist.ca/2021/08/law-bytes-podcast-episode-99/

93 Google takes legal action over Germany's expanded hate-speech law. (2021, July 27). *Reuters*. https://www.reuters.com/technology/google-takes-legal-action-over-germanys-expanded-hate-speech-law-2021-07-27/

94 Digital Services Act Proposal: Recommendations for the EU Parliament and Council. (2021). *Electronic Frontier Foundation*. https://www.eff.org/files/2021/05/07/dsa_recommendations_parliament_council.pdf

95 *Digital Services Act 2020*, s.4 a.31

96 *Digital Services Act 2020*, s.4 a.26-27

97 "Specific groups or persons may be vulnerable or disadvantaged in their use of online services because of their gender, race or ethnic origin, religion or belief, disability, age or sexual orientation. They can be disproportionately affected by restrictions and removal measures following from (unconscious or conscious) biases potentially embedded in the notification systems by users and third parties, as well as replicated in automated content moderation tools used by platforms." *Digital Services Act 2020*, s. 3. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM:2020:825:FIN

98 *Network Enforcement Act 2017*, a. 1 s. 4.

99 *Enhancing Online Safety Act 2015* (Cth) pt 4 div 2-3.

100 Barbaschow, A. (2020, March 24). ISPs to continue blocking graphic violent content in Australia. *ZDNet*. https://www.zdnet.com/article/isps-to-continue-blocking-graphic-violent-content-in-australia/

101 *Online Safety Bill 2021* (Cth). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf

102 Countering illegal hate speech online - Commision initiative shows continued improvement, further platforms join. (2018, January 19). Press Release. *European Commission*. https://ec.europa.eu/commission/presscorner/detail/en/IP_18_261