POISONING DEMOCRACY

How Canada Can Address Harmful Speech Online

NOVEMBER 2018

PUBLIC
POLICY
FORUM

PPF

PUBLIC POLICY FORUM
FORUM DES POLITIQUES PUBLIQUES

ABOUT PPF

Good Policy. Better Canada. The Public Policy Forum builds bridges among diverse participants in the policy-making process and gives them a platform to examine issues, offer new perspectives and feed fresh ideas into critical policy discussions. We believe good policy is critical to making a better Canada—a country that's cohesive, prosperous and secure. We contribute by:

- Conducting research on critical issues
- Convening candid dialogues on research subjects
- Recognizing exceptional leaders

Our approach—called **Inclusion to Conclusion**—brings emerging and established voices to policy conversations, which informs conclusions that identify obstacles to success and pathways forward. PPF is an independent, non-partisan charity whose members are a diverse group of private, public and non-profit organizations.

1400 - 130 Albert Street
Ottawa, ON, Canada, K1P 5G4
613.238.7858

ISBN: 978-1-988886-34-3

ppforum.ca

@ppforumca

WITH THANKS TO OUR PARTNERS

PRIMARY SUPPORT

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

ADDITIONAL SUPPORT

CÉCD CSDC

CENTRE POUR L'ÉTUDE DE LA CITOYENNETÉ DÉMOCRATIQUE

CENTRE FOR THE STUDY OF DEMOCRATIC CITIZENSHIP

UNIVERSITÉ Concordia UNIVERSITY

UBC THE UNIVERSITY OF BRITISH COLUMBIA

INDIVIDUALS WHO PROVIDED FEEDBACK ON THIS REPORT

Jane Bailey, Sara Bannerman, Robyn Caplan, Gareth Chappell, Azmina Dhrodia, Tamir Israel, Satyamoorthy Kabilan, Vivek Krishnamurthy, Taylor Owen, Ben Scott, Dwayne Winseck.

Finally, thank you to all who participated in our workshop in Ottawa on May 14-15, 2018.

ABOUT THE AUTHORS

Dr. Chris Tenove, Postdoctoral Fellow, Department of Political Science, University of British Columbia

Dr. Heidi J.S. Tworek, Assistant Professor, Department History, University of British Columbia and Non-Resident Fellow at the German Marshall Fund of the United States and the Canadian Global Affairs Institute

Dr. Fenwick McKelvey, Associate Professor, Communication Studies, Concordia University

The authors of this report are independent from the Public Policy Forum. The report's views and recommendations are their own, and may not necessarily reflect those of PPF or other partner organizations.

TABLE OF CONTENTS

EXECUTIVE SUMMARY AND RECOMMENDATIONS

Social media platforms provide unprecedented opportunities for citizens, political candidates, activists and civil society groups to communicate, but they also pose new challenges for democracy. One key problem is the rise of harmful speech online, which can undermine democratic participation and debate. Harmful speech refers to a range of forms of problematic communication, including hate speech, threats of violence, defamation and harassment.

Canada has well-established policies to address the most toxic forms of harmful speech in non-digital media, and some are applicable to harmful speech online. **However, the current regulatory approaches cannot address the speed, scale and global reach of harmful speech on social media platforms.** Today, most decisions about Canadians' exposure to harmful speech are made by foreign social media companies, with little public input or accountability. As a result, there is an imbalance between the limited democratic oversight of online platforms and the significant threat that harmful speech poses to democracy.

This report explains some of the most problematic forms of harmful speech, how they affect democratic processes, and how they are currently addressed in Canada. The report draws lessons from policy responses that are being developed or implemented in other countries and at the international level. It then sets out three mutually supporting policy recommendations for the Canadian context. In brief, the report proposes that the Canadian government and key stakeholders should:

1. **Implement a multi-track policy framework to address harmful speech:**

   - An inter-agency task force should be created immediately to clarify how governments in Canada can better apply existing regulatory measures to address harmful speech online, and to examine the growing role of social media platforms in regulating free expression.

   - The federal government should set clear expectations for social media companies to provide information about harmful speech to the public and to researchers.

   - The federal government should launch a multi-stakeholder commission to examine the social and political problems posed by harmful speech online and identify solutions that fall outside of current regulatory measures. This commission would contribute to a broader Canadian discussion regarding public input and oversight of online content moderation.

2. Develop a Moderation Standards Council:

    - The multi-stakeholder commission should consider the creation of a Moderation Standards Council (MSC), analogous to the Canadian Broadcast Standards Council, but adapted for the specific context of online content.
    - The MSC would enable social media companies, civil society and stakeholders to meet public expectations and government requirements on content moderation. It would improve transparency and help internet companies develop and implement codes of conduct on addressing harmful speech. It would create an appeals process to address complaints about content moderation policies and decisions. It would also address potential jurisdictional conflicts over the regulation and standards of content moderation within Canada and contribute to international standards-making.

3. Build civil society capacity to address harmful speech online:

    - Compared to other countries, Canada lacks robust research and civil society programs to address harmful speech. Governments, universities, foundations and private companies should provide significant support for research in these areas and programs to develop, test and roll out measures to respond to them.
    - Social media companies and stakeholders should create an "election contact group" to quickly and effectively share information about threats to electoral integrity.

These policies can help government, internet companies and civil society work together to create a digital public sphere in Canada that is healthy, inclusive and democratic.

INTRODUCTION

During the first half of 2017, British member of Parliament Diane Abbott was targeted with over 8,000 abusive messages on her Twitter account.[1] Abbott, who was the United Kingdom's first black female MP, told Amnesty International:

> **Online abuse does damage you, it damages your confidence, it corrodes your self-esteem and it can make you second guess yourself as to what you talk about and what you write about ... [Online abuse] is not free speech, it's actually limiting everyone else's free speech ...[2]**

Public figures in Canada also face serious online harassment, which ranges from degrading comments to death threats.[3] As Victoria, B.C., Mayor Lisa Helps told the authors of this report, "I welcome it when people have different opinions than mine and share them, but that's very different from the torrents of abuse and misogyny that we sometimes face."

Social media platforms are used not only to target individuals with threats and harassment; they are also used to spread hate. Platforms host an alarming amount of communication that denigrates people based on their ethnicity, gender or other identities, and groups with hateful ideologies use social media to spread their messages and seek supporters.

These toxic levels of harmful speech online undermine Canada's democratic process. This report clarifies the risk that harmful speech poses to democratic engagement and democratic processes, and synthesizes what we know about the creation, circulation and political impacts of harmful speech online. It then examines current and proposed policies by other countries to respond to harmful speech and concludes with a set of policy recommendations that would enable governments, social media platforms and civil society to better understand and reduce the threat it poses to democracy.

We use the term "harmful speech" to refer to communication that is abusive, threatening, denigrating or that incites violence, and which may therefore undermine people's full, free and fair participation in politics and political debates. The term captures the wide scope of the problem and highlights the need for a complementary set of legal and regulatory responses to address different types of harmful speech. Some forms of harmful speech—such as hate propaganda and uttering violent threats—are already illegal in Canada, though enforcement policies need to be improved. Other forms of harmful speech, such as co-ordinated campaigns of harassment and denigration, may undermine democratic participation, even if

[1] Amnesty International. 2018. #ToxicTwitter: Violence and Abuse Against Women Online, p. 17.

[2] Amnesty International, p. 32.

[3] Dawson, T. 2017. Threats against Wynne range from the bizarre to the serious, documents reveal, Ottawa Citizen; Trynacity, K. 2018. 'A wake-up call': Documents detail litany of threats against Premier Rachel Notley, CBC News; Smart, A. 2018. Victoria mayor deletes Facebook because it 'rewards anger and outrage,' Canadian Press.

individual messages do not violate laws. In most cases, complaints about harmful speech are addressed (or not) according to the content moderation policies of private companies, rather than through legal systems or other processes with public oversight.

Our report calls for renewed co-ordination of the governance of free expression to meet these new challenges. Canada lacks reliable and effective processes to evaluate and influence social media policies, including those that may significantly impact democracy and human rights. Moreover, the social media companies themselves are seeking clearer expectations, guidelines and processes from governments on these issues.[4] The constant negative headlines about social media are bad for the companies' business and many people seem to be reducing their usage of platforms or abandoning them altogether. Forty-four percent of Facebook users in the United States between the ages of 18 and 27 said they deleted the Facebook app from their phone in the past year.[5] Our proposals are also meant to help social media companies help themselves. To put it simply, more co-operation on these issues with government and civil society makes good business sense.

Harmful speech is not a problem that can simply be "solved." Instead, Canada and other democracies need to engage in ongoing efforts to safeguard rights to free expression, thought and belief, as well as preventing harmful speech from disadvantaging individuals or groups based on race, national or ethnic origin, colour, religion, sex, gender expression, age or mental or physical disability. Political speech in Canada has never been completely civil, but the frequency, scale and ad-hoc enforcement of online communication have introduced serious new challenges to these longstanding issues in democratic participation.

There is urgent need to implement more accountable, more public and more robust responses to increasing levels of harmful speech online.

This report makes three mutually supporting policy recommendations that would enable government, platform companies and civil society to better pursue ongoing dialogue and action to address this pressing issue. The recommendations are for government to:

1. Create a task force to improve government enforcement of existing policies; ensure platform transparency; and launch a public process to develop responses to issues of harmful speech and online content moderation more broadly.
2. Develop a multi-stakeholder Moderation Standards Council to strengthen and co-ordinate action by internet companies and stakeholders.
3. Build civil society capacity to investigate and address harmful speech.

[4] Caplan, R. (In press). Content or Context Moderation? Artisanal, Community-Reliant and Industrial Approaches. Data & Society Research Institute.

[5] Perrin, A. 2018. Americans are changing their relationship with Facebook, Pew Research Center.

This report is written by three academic researchers who are independent of the Public Policy Forum, but it builds on PPF's report, Democracy Divided: Countering Disinformation and Hate in the Digital Public Sphere.[6] While *Democracy Divided* provided a menu of policy options to address a range of digital threats to democracy, we focus on measures to mitigate harmful speech. Our policy recommendations can also contribute to addressing disinformation, as disinformation and harmful speech often intertwine. The evidence and recommendations in this report are based on presentations at a workshop of international experts convened on May 14-15, 2018, in Ottawa (under the Chatham House Rule), background discussions and interviews with experts from Europe and the United States, as well as published research.

[6] Greenspon, E., and Owen, T. 2018. Democracy Divided: Countering Disinformation and Hate Speech in the Digital Public Sphere, Public Policy Forum.

SECTION ONE

HARMFUL SPEECH AND ITS PROBLEMS

Harmful speech takes many forms on social media platforms. Examples include targeted harassment of individual Twitter users, hateful videos posted on YouTube or Facebook, and the unwanted exposure of individuals' private information on Reddit. To be clear, we use "speech" to refer to text, video, audio, images, and other forms of communication, whether posted publicly for others to read or sent directly to people.

As Faris and colleagues observe, harmful speech "has emerged as one of the central challenges for Internet policy experts, often pitting protections for freedom of expression online against the rights and interests of those that are subject to online harassment."[7] We do not offer one static definition of harmful speech and the harms it causes. We see these definitions as contextual and the result of ongoing public discussion, since harm and harmful speech can take different forms and mean different things to different individuals and societies over time.

WHAT ARE COMMON FORMS OF HARMFUL SPEECH ONLINE?

Group-focused hate speech. Hate speech is an attack on a group or members of a group based on qualities such as race, ethnicity, religion, nationality, sexual orientation, gender or disability. For instance, during the 2016 U.S. election and 2017 German election, there were co-ordinated efforts to promote racist, nativist and anti-immigrant content on social media. There have also been widespread efforts to promote anti-Semitic, anti-LGBTQ, misogynist and anti-Muslim hatred.

Targeted threats and intimidation. Individuals may be targeted with threats of sexual assault, threats of violence to family and threats of smear campaigns, frequently against political figures, journalists and activists. People can also be threatened through the unwanted exposure of their personal information online (known as doxing), which can make them vulnerable to further harassment, threats and violence by others.

Co-ordinated harassment. People frequently encounter problematic but legal forms of harassment online, including offensive speech and memes, repeated insults, adversarial use of platform complaint processes, and the use of bots or fake accounts to flood their social media feeds. These techniques are used by individual "trolls," but also in co-ordinated campaigns by governments, partisan groups and extremist ideological movements seeking to advance political aims and hamper or discourage their opponents.

Harmful speech often involves **false or misleading information**. Disinformation techniques can be used to promote hate or distrust toward individuals and groups. One technique is to "hijack" news events by introducing false information or misleading interpretations when journalists and members of the public are

[7] Faris, R., Ashar, A., Gasser, U., and Joo, D. 2016. Understanding Harmful Speech Online, Berkman Klein Center for Internet & Society, Harvard University. p. 5.

trying to make sense of what is happening. In Canada, this has been seen in online campaigns by anti-Muslim, alt-right and other groups in the immediate aftermath of violent events such as the mass shooting in a Quebec mosque in January 2017, the sudden increase in border crossings from the U.S. to Canada in August 2017, and the vehicle attack in Toronto in April 2018.[8]

Harmful speech may also be used in **campaigns of foreign interference**. One study found that 27 governments pushed messaging on social media that targeted individuals, communities or organizations with hate speech or harassment.[9] "State-sponsored trolls" (government employees or contracted firms) frequently take actions such as making violent threats against individuals, spreading prejudiced views, and amplifying harmful speech using bots.[10]

HOW MUCH HARMFUL SPEECH IS ON SOCIAL MEDIA?

Users upload large amounts of harmful speech to social media platforms, affecting many people. In Canada, a 2016 survey found over one-quarter of people have been harassed via social media platforms, with higher rates among users who are young (under 34), or who identify as LGBTQ or a visible minority.[11] A 2017 survey of Americans found about four in ten had faced online harassment, with two in ten experiencing severe harassment such as physical threats and sexually explicit abuse.[12]

With respect to hateful speech targeting groups, the global transparency reports of social media platforms reveal the momentous challenge they face. Facebook reported that it took action on 2.5 million pieces of hateful content in Q1 2018, up from around 1.6 million in Q4 2017.[13] YouTube users flagged videos as hateful or abusive over 6.6 million times between April and June 2018.[14] Twitter does not provide similar numbers for violations of its terms of service regarding abusive or hateful messages, but research suggests it is a primary vector for harmful speech. A study by the Anti-Defamation League identified over 4.2 million anti-Semitic tweets, written in English, between January 2017 and January 2018.[15] This is not just an American phenomenon. Hateful and intolerant speech appears to be increasing dramatically on Canadian social media and websites.[16]

[8] Goldsbie, J. 2018. How the far right spun the Toronto van attack as Islamic terrorism, Canadaland; Rocha, R. 2018. Data sheds light on how Russian Twitter trolls targeted Canadians, CBC News.

[9] Bradshaw, S., and Howard, P. 2018. Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation, The Computational Propaganda Project, Oxford University.

[10] Nyst, C., and Monaco, N. 2018. State-Sponsored Trolling: How Governments Are Deploying Disinformation as Part of Broader Digital Harassment Campaigns, Institute for the Future.

[11] Angus Reid Institute. 2016. Trolls and tribulations: One-in-four Canadians say they're being harassed on social media. In this survey, 'harassment' ranged from being called offensive names to cyber-stalking and threats of violence.

[12] Pew Research Center. Online Harassment 2017. See also Lenhart, A., Ybarra, M., Zickuhr, K., and Price-Feeney, M. 2016. Online Harassment, Digital Abuse, and Cyberstalking in America, Data & Society Research Institute.

[13] Facebook. 2018. Community Standards Enforcement Preliminary Report.

[14] Google. 2018. YouTube Community Guidelines Enforcement.

[15] Anti-Defamation League. 2018. Quantifying Hate: A Year of Anti-Semitism on Twitter.

[16] Boutilier, A. 2018. Rise of right-wing extremists presents new challenge for Canadian law enforcement agencies, The Toronto Star; Naffi, N. 2017. The Trump effect in Canada: A 600 per cent increase in online hate speech, The Conversation.

There are, however, serious shortcomings regarding data on Canadian experiences of harmful speech online. Surveys use very different definitions of harassment and other terms. The data provided by social media platforms refers to decisions they have made on content moderation, which may over-report or under-report the prevalence of hate speech. For instance, it is clear that Facebook has failed to sufficiently address hateful speech in some countries.[17] Furthermore, social media platform data is often not broken down by country.

WHO IS CREATING AND SPREADING HARMFUL SPEECH?

Research shows that some governments and organizations use harmful speech to advance their aims. State-sponsored trolls (government employees or contracted firms) are widely engaged in "making death and rape threats, using bots and automated agents to amplify vitriolic attacks at scale, making accusations of treason or collusion with foreign intelligence agencies… and sowing acrimonious sexism."[18] Ideological communities and extremist groups also pursue co-ordinated campaigns of harmful speech. These include transnational alt-right groups, which promote stigmatization of people of different racial, ethnic, gender and religious identities.[19]

In addition to governments and organized groups, there are many individuals who create, target and share harmful speech. Some do so with political intent, some to settle personal scores, and some do so occasionally, in response to major news events, including elections.

Harmful speech online not only has different creators; it also has different target audiences. Hate and abuse are sometimes directed at individuals or groups to cause them distress or posted publicly for all to see. They may also be disseminated in semi-private or "dark" social media spaces, such as Facebook Groups, 4Chan sites, or WhatsApp groups. These methods are used by extremist and pro-hate groups (but also human rights activists in repressive countries) to build group identity and plan future actions.[20] Researchers and law enforcement struggle to quantify the extent of such messaging.

WHAT THREAT DOES HARMFUL SPEECH POSE TO DEMOCRACY?

In this report, we focus on those forms of harmful speech whose intent or whose effect is to exclude people from full, free and fair participation as democratic citizens.[21] Democratic participation includes opportunities

[17] Koebler, J., and Cox, J. 2018. The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People, Vice: Motherboard; Stecklow, S. 2018. Hatebook: Why Facebook Is Losing the War on Hate Speech in Myanmar, Reuters.

[18] Nyst and Monaco, State-Sponsored Trolling, p. 1. See also Bradshaw and Howard, Challenging Truth and Trust.

[19] Lewis, R. 2018. Alternative Influence: Broadcasting the Reactionary Right on YouTube, Data & Society Research Institute.

[20] Brown, A. 2018. What Is so Special about Online (as Compared to Offline) Hate Speech? Ethnicities 18(3): 297–326.

[21] For more on digital threats to democracy, see: Tenove, C., Buffie, J., McKay, S., and Moscrop, D. 2018. Digital Threats to Democratic Elections: How Foreign Actors Use Digital Techniques to Undermine Democracy, Centre for the Study of Democratic Institutions, University of British Columbia.

to vote, to run for and hold public office, to join and work for political parties and other civil society organizations, to assist electoral management, and to contribute to public discussions of political issues. Healthy democracies safeguard these opportunities from repression, threat and discrimination or disadvantage, whether these come from the actions of governments or private organizations. Harmful speech online can damage democratic participation in several ways:

Silencing or threatening individuals who seek to participate in politics. Politicians or candidates for election are frequently targeted for threat, harassment and abusive language. So, too, are activists, journalists and other individuals who seek to contribute to public debates. Online harassment and abuse can cause fear and psychological harm. To avoid attack, some people choose not to take public stands or engage in online discussions on political issues.

Creating barriers for members of certain groups to engage in politics. In addition to harmful speech that targets individuals, there is also mass dissemination of hateful, derogatory or stigmatizing messages about different social groups. Targeted groups often already face barriers to full political participation, including ethnic, racial, religious or gender minorities. In other words, harmful speech on social media may exacerbate existing cultural dynamics of exclusion. For instance, women in politics regularly face high levels of online abuse, a hazard *The New York Times* has referred to as "campaigning while female."[22]

Promoting conflict and social tension. Social media can be used to mobilize social grievances, biases and hatreds, including during elections. Such messaging can undermine the capacity for elections to channel political conflict into non-violent struggle, one of the most basic aims of electoral contests. At its most serious, harmful speech—and especially hate speech—can contribute to inter-group violence, as seen in Myanmar and Sri Lanka, or to individual attacks such as those in Quebec and Toronto.[23]

Undermining the laws, norms and trust needed for democratic communication. Actors have promoted harmful speech and disinformation online in ways that violate electoral laws, including legal limits on foreign communication or requirements for transparency regarding the purchase of public messaging.

Even when it does not violate electoral laws, harmful speech can undermine civility and other norms of communication that are needed for the productive exchange of ideas in online spaces. It can also poison people's views toward some individuals, groups or democratic institutions.

Given these threats to democracy, it is critical to understand how harmful speech works online, who is affected, and what can be done about it.

[22] Astor, M. 2018. For Female Candidates, Harassment and Threats Come Every Day, The New York Times.

[23] See, respectively, Stecklow, Hatebook; Taub, A., and Fisher, M. 2018. Where Countries Are Tinderboxes and Facebook Is a Match, The New York Times; Lamoureux, M. 2018. Here Are the Far-Right Conspiracists the Quebec City Mosque Shooter Followed, Vice; Goldsbie, How The Far Right Spun The Toronto Van Attack As Islamic Terrorism.

SECTION TWO

CURRENT POLICIES AND THEIR LIMITATIONS

Canada has several existing policies that may be used to address harmful speech online. However, the primary "regulators" on matters of harmful speech and free expression are arguably the large social media platforms. We argue that existing government policies cannot deal with the scale, speed and systemic impacts of harmful speech online and the policies of social media companies are insufficiently transparent and accountable to Canadians, given their impact on our public discourse and politics.

CANADIAN GOVERNMENT REGULATORY FRAMEWORKS

Federal and provincial governments have developed several legal frameworks that can be used to address harmful speech online.[24]

- **Criminal law.** Various forms of harmful speech fall under the Criminal Code, including criminal harassment, defamatory libel, counselling suicide, uttering threats and intimidation. There are also criminal provisions against hate speech. Specifically, sections 318 and 319 impose criminal sanctions against anyone who intentionally promotes genocide or who incites hatred against groups identifiable by factors such as race, religion and sex.
- **Human rights law.** Multiple provincial and territorial human rights provisions prohibit the incitement of hate or group discrimination through public displays, broadcasts or publications. So, too, did federal human rights law until it was repealed in 2013: Section 13 of the Canadian Human Rights Act was used in several cases of hate promoted online.
- **Civil litigation.** Individuals can pursue civil litigation to seek redress for some forms of harmful speech, including defamation and the non-consensual disclosure of intimate images.
- **Regulatory/administrative law.** Several federal, provincial, and territorial governments include offices or support units that might help address different forms of harmful speech. These include privacy commissioners, who can address complaints and contribute to requests for service providers to remove content, such as the non-consensual disclosure of private information or intimate images.

[24] For more, see Bailey, J. (2018). Abusive and Offensive Speech Online: An Overview of Canadian Legal Responses Focusing on the Criminal Law Framework, Law Commission of England and Wales, 2018; Walker, J. 2018. Hate Speech and Freedom of Expression: Legal Boundaries in Canada, Parliamentary Information and Research Service.

- **Electoral law.** The Canada Election Act has no provisions against harmful speech per se. However, harmful speech online may be used in ways that violate electoral law, such as prohibitions on foreign communications to induce Canadians to vote or not vote for particular candidates.

These legal measures provide important means for governments and individuals to address some of the most egregious forms of harmful speech. However, as several experts told us, effective use of these measures is limited by legal definitions, operational challenges and obstacles to co-operation across different agencies.

Legal Status	**Offences**
Criminal law	Criminal harassment; uttering threats; intimidation; identity fraud; extortion; false messages, indecent or harassing communications; counselling suicide; defamatory libel and hate propaganda.
Human rights law	Public displays, broadcasts, or publications that promote hate or discrimination against people or groups according to factors such as race, religion, and sex.
Civil litigation	Defamation; non-consensual disclosure of intimate images; harassment.
Election law	Publishing false statements about candidates to affect elections; foreign actions to induce people to vote or not vote for candidates.

Even if these legal measures were used more robustly, they would not adequately address harmful speech on social media. They tend to be costly and difficult to enforce (particularly if the responsible party is hard to identify or resides outside of Canada), and they are slow and reactive—making them less suitable for addressing hate and harassment in the midst of an election campaign, for example. Finally, these measures are narrow and have a high threshold for action, focused as they often are on individual or organizational culpability. This makes them inappropriate tools for addressing forms of harmful speech that are problematic and widespread but that should not lead to individual legal sanction.

We therefore argue that harmful speech online must also be addressed as an issue of media regulation and media standards*.*

This is an important regulatory approach for addressing harmful speech disseminated by other forms of mass media. Broadcasters and publishers hold significant liability for illegal content, including content appearing on their websites. Broadcasters and publishers have also developed self-regulatory measures to address issues of harmful speech and broader questions regarding their conduct and transparency. For instance, the Canadian Broadcast Standards Council has developed codes of conduct on issues including broadcasting ethics, depictions of violence and unduly negative or stereotyped portrayals of social groups. The council ensures these codes are enforced, provides templates for complaints and offers an appeals process. It also releases annual data on the number of complaints.

Social media platforms are not bound by comparable policies in Canada. Social media companies do not neatly fit into Canadian policies on telecommunications, broadcasting or publishing. Indeed, a social media company like Facebook has different functions that arguably fall under the spirit of these policy areas and more. The novelty and ambiguous status of social media platforms has challenged their regulation as media. Social media communication often cannot be said to be private, like a telephone call, nor entirely public, like the open signals of radio. Social media companies have tended to call themselves technology platforms, rather than broadcasters or publishers, to limit their responsibility for content. This question of "publisher or platform" needs to be resolved or bypassed to pursue successful regulation.[25]

Regulatory measures therefore need to be developed to address the role social media platforms play in disseminating—and responding to—harmful speech. We argue that the regulatory focus should be on issues and processes of content moderation by social media platforms, which we describe below.

REGULATION BY SOCIAL MEDIA PLATFORMS

Social media companies set and enforce policies that shape communication among their users, including on issues of harmful speech. They do so primarily through content moderation, the processes they use to filter, flag, promote and delete content. The general shape of a platform's rules on content are usually spelled out in their terms of service or community guidelines. As Data & Society researcher Robyn Caplan observes, when "policymakers within these companies try to draw lines around the kind of content they want or do not want on their platforms, they become… the *arbiters of hate, arbiters of harassment*, and *arbiters of disinformation* around the world."[26]

Social media content moderation is a critical but inadequately understood function of social media platforms. It is also a burgeoning global industry in its own right, employing tens of thousands of people.[27] Content moderation by U.S.-based platforms, even for content produced and/or consumed by individuals in other countries, is primarily governed by section 230 of the U.S. Communications Decency Act (1996). The rule exempts intermediaries from being treated as a publisher of the content they host and encourages them to undertake their own internal regulatory approaches. The rule, often credited with enabling the

[25] On this issue, see Greenspon and Owen in Democracy Divided (p. 22). See also: Digital, Culture, Media and Sport Committee. Disinformation and 'Fake News': Interim Report. House of Commons, United Kingdom, paragraphs 51-60.

[26] Caplan, Content or Context Moderation (emphasis in original).

[27] Labour conditions in this industry are frequently criticized. See Roberts, S. 2016. Commercial Content Moderation: Digital Laborers' Dirty Work, University of Western Ontario.

social media revolution, has led to active enforcement of public community standards and terms of use on platforms, albeit with little transparency or consistency across platforms.[28] These regulatory conditions have led to what have been called the "secret rules of the internet."[29] Until recently, the status of section 230 has been unclear in Canada, but the United States-Mexico-Canada Agreement (USMCA) will bring this limited liability to platforms in Canada once ratified.[30]

Social media company policies address many forms of harmful speech through content moderation policies. For instance, Facebook, Twitter and YouTube all have rules that prohibit content—including credible threats of violence or incitement to cause violence, harassment and abuse, revealing people's private information, and speech that promotes hate or violence against individuals or groups on the basis of characteristics such as race or ethnic origin, gender, religion or disability. These policies are set out in public-facing pages and are defined in greater detail in documents used to make decisions on enforcement.

Major social media platforms are devoting increasing efforts to clarifying, publicizing and enforcing their content moderation policies. However, major concerns remain.

These include:

- **Transparency.** After years of pressure, social media platforms now provide more public information about their content moderation policies and the enforcement of these policies. However, significant gaps and inconsistencies remain. Users often cannot understand why their content was removed or why problematic content they have reported has not been addressed. Researchers and civil society groups struggle to determine how quickly or reliably complaints are addressed. More broadly, researchers lack the data needed to adequately evaluate the spread and impact of harmful speech on platforms.[31]

- **Consistency versus arbitrariness**. Content moderation policies need to adapt to changing circumstances. However, social media companies often develop policies without adequately foreseeing or reacting to abuse, as seen in the case of violent mobilization in Myanmar and elsewhere.[32] In other cases, content moderation policies are implemented in an arbitrary fashion, as seen in Twitter's responses to parody accounts of Canadian politicians,[33] the inconsistent policies regarding a famous Vietnam War photograph of a child running from a napalm attack,[34] and the

[28] Gillespie, T. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press.

[29] Buni, C. 2016. The Secret Rules of the Internet, The Verge.

[30] Geist, M. 2018. Why the USMCA will enhance online free speech in Canada, Policy Options.

[31] Inconsistencies in transparency mean we have very different understandings of different platforms. For instance, researchers have significant access to data about messaging on Twitter, but the company lags far behind its competitors in offering meaningful information about content moderation policies and enforcement.

[32] Stecklow, Hatebook.

[33] Dubois, E. 2018. We've given tech companies like Twitter too much political power, Maclean's.

[34] Roberts, S. 2018. Digital Detritus: 'Error' and the Logic of Opacity in Social Media Content Moderation, First Monday 23(3).

sudden, cross-platform removal of Infowars material.[35] These inconsistencies, combined with deficient transparency regarding content moderation systems, have led many to call for better appeals mechanisms.

- **Responsiveness to cultural or national contexts.** Social media platforms with global audiences struggle to moderate content in ways that adhere to national laws, let alone local or national cultures.[36] On issues of harmful speech, as experts told us, this can lead to failures to recognize hateful comments disguised as euphemisms or images.
- **Effectiveness and fairness in the use of artificial intelligence (AI).** Facebook and other companies have admitted that while their content moderation algorithms are very effective at identifying porn, spam and copyright violations, they are less effective at identifying hate speech. In its most recent report, Facebook stated its algorithms detected 38 percent of the hateful content it ultimately acted on (which, as noted previously, was 2.5 million pieces of hateful content in Q1), compared to 99.5 percent of terrorist content and 95.8 percent of adult nudity and sexual activity.[37] Identifying hate speech through keyword-based algorithms is difficult because hate and harassment often require very contextualized interpretation, and hate communities and target communities frequently discuss similar topics, although for very different reasons. Users thus identify hate speech more often than the platform companies themselves do. This places a significant burden on individual users. Further, plans to implement AI to solve content moderation have raised serious limitations about the state of technology—straightforward declarations of race and sexual identity, such as "I am a black gay woman," have been ranked highly toxic in one review of a content moderation AI system.[38]

Social media policies are setting the terms of free expression for much of our online communication, and these policies generate concern and confusion. The aim for Canadian regulation should not be to further restrict freedom of expression, but to disclose how platforms are already moderating freedom of expression and ensure more robust governance in line with Canadian laws and norms.

[35] Romano, A. Apple Banned Alex Jones's Infowars. Then the Dominoes Started to Fall, Vox.

[36] Caplan, Content or Context Moderation.

[37] Facebook, Community Standards Enforcement Preliminary Report.

[38] Blue, V. 2017. Google's comment-ranking system will be a hit with the alt-right, Engadget.

SECTION THREE

INTERNATIONAL RESPONSES TO HARMFUL SPEECH

Canada is one of many democracies grappling with these complex questions of online content. As a comparatively small market in the online world, the international context is important for several reasons. First, Canada can co-ordinate efforts with other countries and international actors. Second, Canada can learn from specific existing and proposed policy responses. Third, Canada can consider international legal frameworks that already address some of these issues.

1. **Germany**

   The best-known policy to address issues of harmful speech is Germany's Network Enforcement Act (NetzDG), known colloquially as a hate speech law, which came into full force in January 2018. Contrary to popular misconceptions, the law *did not* introduce new definitions of hate speech. Its purpose is to enforce 22 statutes in the online world that already existed in the German criminal code (which is much stricter than U.S. law). The law's novelty lies in placing a targeted form of liability on the platforms for content. Platforms are required to respond to complaints within 24 hours or face a fine of up to 50 million Euros. This was the first law to push social media companies that have more than two million users in Germany to adapt their content moderation to a national context. The platforms protested but responded accordingly. Facebook, for example, now employs around 65 people solely to process complaints under NetzDG.[39] There were concerns, however, that the law outsourced decisions over speech to private companies, undermined due process, and that it could enable government censorship in less democratic countries, which might emulate the idea.[40]

   The law is comparatively new and its effects still hard to assess. We can, however, learn three lessons. The first is that real financial liability commanded platform companies' attention. Germany had tried a voluntary compliance system with the companies since 2015 but found it ineffective. The German government chose the path of law only after it deemed the companies insufficiently compliant. A government study in early 2017 found that YouTube had deleted 90 percent of criminal content, Facebook 39 percent, and Twitter only one percent within the requested 24 hours.[41] Since the introduction of the law, compliance rates have hovered above 95 percent. The second lesson is that the law's focus on individual complaints places the burden on individual users or complaint bodies to report instances of hate speech. (Facebook received 1,704 complaints in the

[39] Busvine, D. 2018. Facebook deletes hundreds of posts under German hate-speech law, Reuters.

[40] Tworek, H. 2017. How Germany Is Tackling Hate Speech. Foreign Affairs.

[41] Federal Ministry of Family Affairs and Federal Ministry of Justice and Consumer Protection. 2017. Löschung von strafbaren Hasskommentaren durch soziale Netzwerke weiterhin nicht aussreichend.

first half of 2018 and removed 362 posts; YouTube addressed 214,827 flagged items and removed or blocked 27 percent; and Twitter received 264,818 complaints and removed just under 11 percent.[42]) Given the massive flows of online information, the question is whether it is sustainable to expect individual users to police platforms. The third lesson is that it is hard to predict and measure the full effects of legal policies. The narrow focus on the number of complaints and this narrow category of illegal speech tell us very little about any potential larger effects on Germany's information ecosystem or political discourse.

2. United States

Until very recently, social media companies had one set of rules for the world, extrapolated from American laws and norms on speech. Complaints about the sometimes arbitrary deletions and curation decisions were countered with one of a few objections: sheer volume made it impossible to apply different standards in different places; companies did not want to open the door to censorship demands from non-democracies; regulation might stifle innovation; and the companies were trying their best in a new and constantly evolving environment.

U.S. responses have mainly placed the burden on the individual, whether to report problematic content or to demonstrate harm. First Amendment protections are the cornerstone of American discussions about content. The delivery of news by platforms has, however, been politicized. Right-wing pro-free-speech groups in particular have used the call of "anti-conservative bias" to push platforms to remove human moderators and, in many cases it seems, change algorithms to promote conservative content. Facebook responded to these criticisms in part by removing human moderators for its news feed in late summer 2016, which helped to create vulnerabilities that massively increased the spread of disinformation in the fall of 2016.[43] Social and political pressure in the U.S. has prompted further changes to platform companies' content moderation policies. Many of these changes have been improvements, but they have largely happened behind the scenes and without oversight.

The current U.S. administration has, for the first time, chipped away at the idea that platforms hold no liability for content. Under Section 230 of the Communications Act, online companies bear no legal liability for content published on their platform. But in early 2018, the U.S. adopted two bills that made it possible to hold websites liable for promoting or facilitating prostitution and sex trafficking. The introduction of specific exceptions to Section 230 may open the door to placing further content liability on platforms in the future, which has led to court challenges and predictions that these laws are ushering in the "next big battle over internet freedom."[44]

[42] See Facebook. 2018. NetzDG Transparency Report; Google. 2018. Removals under the Network Enforcement Law; Twitter. 2018. Netzwerkdurchsetzungsgesetzbericht: Januar - Juni 2018.

[43] Bell, E. and Owen, T. 2017. The Platform Press: How Silicon Valley Reengineered Journalism, Tow Center for Digital Journalism, Columbia University.

[44] Stewart, E. 2018. The next big battle over internet freedom is here, Vox.

Looking forward, some Democratic senators have proposed bills and potential regulation to change the landscape. The Honest Ads Act[45] is a bipartisan bill that would require online ads to include disclosures on who purchased them and how the targeted audience was chosen; this is an extension of the requirements for political advertisements in print, radio and television. Senator Mark Warner (D-VA) released a white paper in August 2018 that detailed 20 social media regulation proposals, ranging from increasing platform liability for content to a public interest data access bill, to creating an inter-agency task force focused on asymmetric threats to democratic institutions.[46]

3. European Commission

The European Commission has pursued a range of strategies to address harmful speech and disinformation online. In May 2016, the EC and several major internet companies (Facebook, Microsoft, Twitter and YouTube) announced a Code of Conduct to counter illegal hate speech online.[47] The companies agreed to clarify their terms of use to prohibit illegal incitement to hatred, and to put in place clear and effective processes to review and remove or disable access to such content within 24 hours. The code, which has since been adopted by Instagram and Google, is not legally binding. However, in early 2018 the EC declared that the companies on average had "removed 70% of all the illegal hate speech notified to them by the NGOs and public bodies participating in the evaluation."[48]

The EC has also attempted to address disinformation in ways that might apply to policy action on harmful speech. In an April 2018 communication to the European Council and Parliament, the EC outlined a "European Approach" to combatting disinformation.[49] It proposed overarching principles and objectives and helped create a multi-stakeholder forum to develop key performance indicators to assess whether its objectives are being met. The EC added a threat of consequences for non-compliance: "Should the results prove unsatisfactory, the Commission may propose further actions, including actions of a regulatory nature." The Multi-stakeholder Forum on Disinformation, which includes Facebook and Google, issued its self-regulatory Code of Practice in September 2018; this is the first set of voluntary standards for social networks and advertisers to address disinformation.[50]

These efforts are important for several reasons. First, Europe is much more active on the issue of harmful speech than the U.S., and is therefore likely to be a better partner for pursuing solutions at the international level and in negotiations with internet companies.[51] Second, the EC strategy is to

[45] Klobuchar, A. 2018. S.1989 - 115th Congress (2017-2018): Honest Ads Act.

[46] Warner, M. 2018. Potential Policy Proposals for Regulation of Social Media and Technology Firms.

[47] European Commission. 2016. Code of Conduct on Countering Illegal Hate Speech Online.

[48] European Commission. 2018. Countering illegal hate speech online – Commission initiative shows continued improvement, further platforms join.

[49] European Commission. 2018. Tackling Online Disinformation: A European Approach.

[50] Multi-stakeholder Forum on Disinformation. 2018. Code of Practice on Disinformation, European Commission.

[51] Tenove, C., and Tworek, H. 2018. What Europe can teach Canada about protecting democracy, The Conversation.

engage and work with internet companies, via codes of conduct, stakeholder forums and industry self-regulation, while threatening stricter regulation if clear measures of success are not established and met. The effectiveness of this strategy warrants close examination. Third, the EC approach includes strong support for policies to empower civil society actors (especially research, journalism and fact-checking organizations), to improve the quality of online communication and to hold to account social media platforms and actors who misuse them.

4. **United Nations and International Human Rights Advocates**

International human rights law requires states to protect rights to free expression and access to information online, and to provide guarantees against discrimination.[52] These obligations flow from the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights, as well as pronouncements and decisions by international and regional bodies. In addition, the Guiding Principles on Business and Human Rights, adopted by the UN Human Rights Council, set out non-binding obligations for private corporations, and thus internet companies. **An international human rights framework is important for policy responses to harmful speech, as it can help structure principled, legal and institutional co-operation among democratic countries**. It also helps clarify the problem of unequal exclusion faced by disadvantaged groups, such as the UN Special Rapporteur on violence against women's recent report on the causes, consequences and responses to online violence against women and girls.[53]

The UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, has proposed a framework for applying international human rights to content moderation by social media companies.[54] This report, based in part on consultations with states, civil society organizations and social media companies, suggests how to balance freedom of expression with concerns about violence, discrimination and other consequences of harmful speech online. "A human rights framework enables forceful normative responses against undue State restrictions," he argued, provided that internet companies "establish principles of due diligence, transparency, accountability and remediation that limit platform interference with human rights" (para 42). More specifically, Kaye joined a number of international rights organizations in calling for global and/or national social media councils to help social media companies meet human rights obligations and achieve greater transparency.[55]

[52] Kaye, D. 2018. Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression; UN Human Rights Council. 2016. The Promotion, Protection and Enjoyment of Human Rights on the Internet. See also the Manila Principles on Intermediary Liability.

[53] Šimonović, D. 2018. Advanced Edition: Report of the Special Rapporteur on Violence against Women, Its Causes and Consequences on Online Violence against Women and Girls from a Human Rights Perspective, UN Human Rights Council.

[54] Kaye, Report of the Special Rapporteur.

[55] ARTICLE 19. 2018. Self-Regulation and 'Hate Speech' on Social Media Platforms.

INTERNATIONAL CHALLENGES FOR POLICY RESPONSES TO HARMFUL SPEECH ONLINE

From these international lessons, we see five major challenges to addressing harmful speech:

1. **Attempts to regulate online communication can provoke paralyzing concerns about threats to free expression.** This is an important issue, but non-intervention is too often justified by legitimate concerns about how to "regulate" user-generated content, or by an unproven belief in a "marketplace of ideas" where the good ideas win out over bad or toxic ones. In reality, harmful speech and disinformation can be faster, cheaper and more enticing to spread. Furthermore, harmful speech can stifle free expression, particularly for members of vulnerable or targeted groups. These considerations require a more thoughtful definition of free expression. Indeed, addressing harmful speech continues a long-standing discussion of how Canada can promote tolerance and full political expression in a multicultural society, one of many reasons why the problems of harmful speech require a Canadian solution.

2. **Regulation of online communication can quickly become an issue of political conflict**, in which parties take positions to achieve incremental political gains with different factions, while allowing the quality of democracy to decline. The failure of the U.S. government to find measures to counteract foreign interference and illegal activity in the 2016 election is a prime example.

3. **Policies to address harmful speech on social media need to address a complex global context.** The major social media platforms are global corporations with a global impact, and that impact varies considerably across countries and groups. Policy responses in any one country need to be attentive to possible knock-on effects—as made clear when concerns about fake news in the U.S. in 2016 rapidly led to fake news policies with troubling consequences for free expression in Brazil, Egypt, Turkey and Zimbabwe. (For this reason, the United Nations, the European Union and many other countries and organizations call on states to respond to harmful speech and disinformation in ways that adhere to international human rights law.) At the same time as being a global issue, social media regulation is also an American issue—since the major platforms outside of China are owned by U.S. corporations. However, other countries are developing strategies to ensure the enforcement of their national or regional laws and policy aims. Canada should too.

4. **Policy responses to harmful speech need to strike a fair balance between public oversight and regulation, and the need to protect innovation and financial health in the industry.** Policies need to be flexible and responsive to different platforms and operating environments, but also fair, reliable and non-arbitrary. Social media companies themselves have shown an admirable willingness to design and implement changes to content moderation but, too often, these seem to be generated in a reactive manner to address the latest high-profile case.

5. **Social media companies like Facebook are struggling with what has been called "the impossible job" of content moderation at the scale of millions or billions of users.**[56] The scale and speed of online communications are a different order of magnitude compared to previous communications technologies. But every new communications technology has posed difficulties. Social media pose new regulatory challenges that resemble past policy controversies, including the use of radio for international propaganda, the perceived threats to "decency" of music and video games, and tensions between participatory culture and copyright protection. In these and other cases, democratic governments and civil society took action and achieved imperfect but workable solutions.

Overall, there are major gaps in our understanding of harmful speech online and its impacts on individuals, groups and democratic processes. These gaps exist due to rapid and ongoing changes in social media platforms' content moderation approaches, insufficient access to data by researchers, and major deficits in research funding and capacity.[57] This list is meant as an introduction to the scope of these challenges.

[56] Koebler and Cox, The Impossible Job.

[57] Matias, J. 2017. Ensuring Beneficial Outcomes of Platform Governance by Massively Scaling Research and Accountability. Perspectives on Harmful Speech Online, Berkman Klein Center for Internet & Society, Harvard University.

SECTION FOUR

POLICY RECOMMENDATIONS FOR CANADA

Governments, private companies and civil society must pursue extensive and sustained co-operation to counteract the toxic levels of harmful speech online. Policy responses to address harmful speech—and the online ecosystem in which it thrives—will necessarily be multi-faceted and complex. We make three recommendations.

RECOMMENDATION 1: IMPLEMENT A MULTI-TRACK POLICY FRAMEWORK TO ADDRESS HARMFUL SPEECH

The Canadian government has several important roles to play in addressing harmful speech online, acting as a maker and enforcer of policies and as a facilitator to ensure dialogue and co-ordination among stakeholders. We propose three tracks of near-term action for the federal government.

First, government should create a high-level task force or working group to ensure federal, provincial and territorial agencies are effectively pursuing and enforcing existing policies. These policies include criminal, electoral, human rights and media laws, as well as intelligence and security responses to foreign interference. Clarity is needed about which agencies are responsible for addressing different forms of harmful speech online, such as co-ordinated hate campaigns or targeted harassment of public figures. In the area of human rights, the task force should address the obstacles posed by the repeal of Section 13 of Canada's Human Rights Act. The task force should include departments of Justice, Public Safety and Emergency Preparedness, and Privy Council / Democratic Institutions, as well as their provincial and territorial counterparts. One possible model is the Cybercrime Working Group, convened in 2013 to address issues of cyberbullying and the non-consensual distribution of intimate images online.[58]

Second, the task force would determine which department in government would take responsibility for setting clear expectations for transparency by social media companies. These should include best practices on public reporting regarding the definition and implementation of companies' terms of service policies and creating public databases of political advertising.

Third, government should launch a multi-stakeholder process to create a roadmap toward addressing issues of harmful speech, better defining their existing legal and regulatory measures, and improving public

[58] Co-ordinating Committee of Senior Officials, Criminal Justice, Cybercrime Working Group. 2013. Cyberbullying and the Non-Consensual Distribution of Intimate Images: Report to the Federal/Provincial/Territorial Ministers Responsible for Justice and Public Safety.

understanding, input and oversight of online content moderation. This process will require research and analysis to gain greater clarity about the extent of harmful speech online and its impacts in Canada.

The eventual roadmap should propose a Canadian approach to harmful speech online, building on existing Canadian media regulation principles and approaches. Critically, the framework should use a layered approach to governance, considering how best to co-ordinate overlapping law as well as different responsibilities of information intermediaries in the dissemination of harmful speech. The framework should protect the neutrality of common carriage services, and it should impose higher standards on applications and platforms that select, moderate and otherwise shape online communication.[59]

The roadmap should also clarify how Canadian initiatives on content moderation can contribute to global efforts to safeguard democracy and comply with international human rights law, including freedom of expression. This might include enforcing criminal law and human rights commitments; clarifying platform liability; and providing oversight and support to address societal harms of harmful speech without undue risk to free speech and economic viability. **The goal, in short, is not simply to address harmful speech, but to develop goals and aspirations for an internet that betters democracy.**

The roadmap should include timelines for achieving goals, including actions by government and by platform companies, and consequences of non-participation or non-compliance.

Several models exist. One option is for government to create and fund an independent high-level commission, such as the Special Committee on Hate Propaganda in Canada (Cohen Committee), which contributed to Canada's criminal laws on hate speech. Another option is for the government to oversee a multi-stakeholder consultation, which could include a high-level advisory group or an independent stakeholder forum, to inform government policies. The European Commission pursued this route to develop a "European Approach" to tackling online disinformation.

For all three tracks, Canada can learn lessons from other countries' processes. Germany, the United States and the European Commission have shown that clear statements about timelines and the consequences of non-compliance can help foster dialogue with platform companies. We suggest that the Canadian government set expectations with high enough costs for failure to induce action by social media companies. Germany's approach shows that, currently, the only way countries outside the U.S. receive sustained attention from social media companies is if they are a massive market (like China or the European Union) or they threaten companies with significant fines. Canada cannot become a bigger market, but it can consider real enforcement mechanisms that will give it a hearing at social media companies and the ability to consult with them concerning solutions. Furthermore, Canada should co-ordinate with international partners when Canadian policy aims overlap with theirs, such as with key aspects of EU approaches to hate speech and

[59] Bridy, A. 2018. Remediating Social Media: A Layer-Conscious Approach. Boston University Journal of Science and Technology Law, 24(2): 193–229. Klonick, K. 2018. The New Governors: The People, Rules, and Processes Governing Online Speech, Harvard Law Review 131: 1598–1670.

disinformation or joint summons to hearings, such as that issued on October 31, 2018 by Damian Collins MP from the United Kingdom and Bob Zimmer from Canada to Facebook CEO Mark Zuckerberg.[60]

RECOMMENDATION 2: CREATE A MODERATION STANDARDS COUNCIL

As part of the policy process, we recommend that multi-stakeholder consultations should create a Moderation Standards Council (MSC) as a specific response to the changing media industry.

Content moderation by social media companies and other app providers is currently in a regulatory grey zone in Canada. The creation of a standards council, modelled after the Canadian Broadcast Standards Council (CBSC), would be an important step forward. The MSC—and the process leading to its creation—could help clarify public expectations and legal responsibilities regarding content moderation in Canada. The new body would help social media platforms address these expectations and responsibilities in a significantly self-regulatory manner, while involving stakeholder input and government oversight. Once established, the MSC could help social media platforms co-ordinate their actions to mitigate harmful speech, disinformation and other threats to democratic discourse online.[61]

Consultations to create the council must include major national and international internet companies, and government must offer appropriate incentives—and threats of more intrusive regulation—to secure their participation. Building on the lessons from Germany, the MSC should have meaningful regulatory tools, such as fines, in its mandate. To ensure the composition of the MSC is not lopsided, the participation of other key stakeholders should be encouraged and supported, ranging from Indigenous communities to human rights organizations to political parties.

The mandate of the MSC would be to work with social media companies, civil society and government to:

1. Create a code of conduct regarding policies to address harmful speech and other important content moderation issues.
2. Develop standards and shared best practices to meet public expectations regarding transparency and accountability, especially in the use of AI.
3. Develop processes to share information with researchers, including means to address companies' concerns about sharing proprietary information or individuals' data.
4. Create and run an appeals process to address complaints about content moderation policies and decisions.

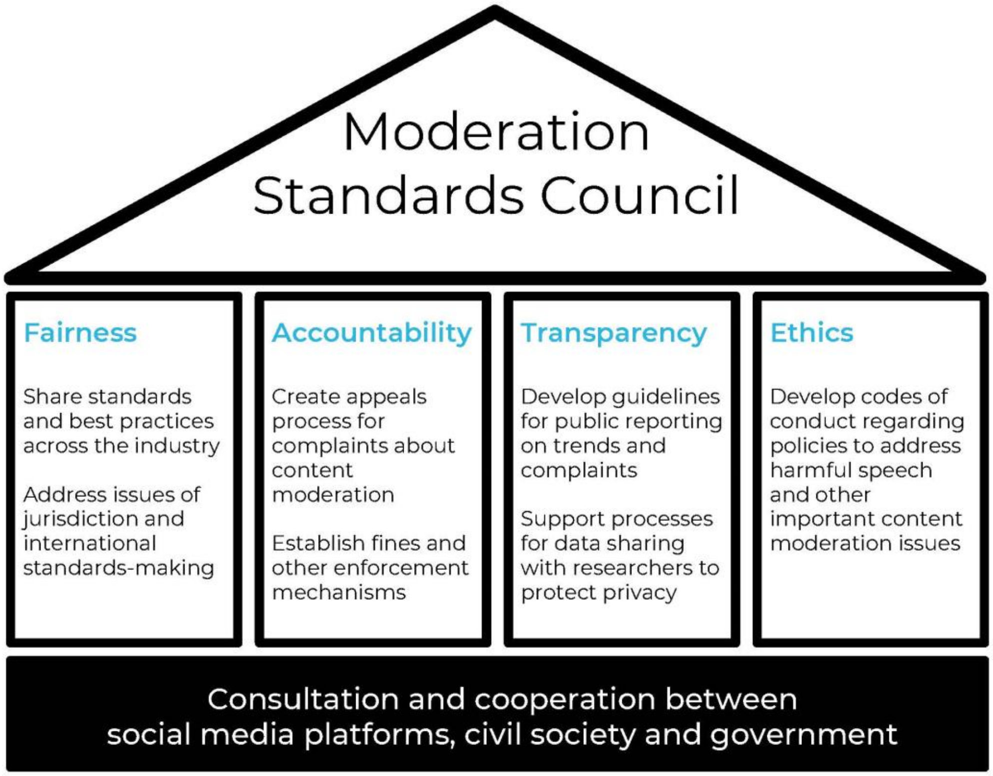
[60] Waterson, J. 2018. UK and Canada Unite to Demand Mark Zuckerberg Answers Questions. The Guardian.

[61] For another proposal to create a standards council for social media, see ARTICLE 19, Self-Regulation and 'Hate Speech' on Social Media Platforms.

5. Address issues of jurisdiction and international standards-making.
6. Establish fines and other powers for government or other enforcement agencies to support its authority and decisions.

We do not propose here what standards should be for these functions, because we recognize that they must arise from stakeholder consultations. However, we expect that international human rights standards and emerging international best practices would provide a helpful baseline for Canadian regulation and would ensure easier co-ordination among countries and social media companies. These include recommendations by the UN Special Rapporteur,[62] as well as the "Santa Clara Principles" on content moderation.[63]

Unlike other suggestions that place burdens on individual users to flag content, a standards council addresses these problems on a societal/ecosystemic level. Addressing ecosystemic questions does not equal censorship. Rather, it enables deliberative and ongoing reflection about how free expression can be protected and encouraged. **Public input and oversight of content moderation is an important means to achieve more accountable, democratic and ultimately more robust responses to increasing levels of harmful speech online.**



Moderation Standards Council

Fairness

Share standards and best practices across the industry

Address issues of jurisdiction and international standards-making

Accountability

Create appeals process for complaints about content moderation

Establish fines and other enforcement mechanisms

Transparency

Develop guidelines for public reporting on trends and complaints

Support processes for data sharing with researchers to protect privacy

Ethics

Develop codes of conduct regarding policies to address harmful speech and other important content moderation issues

Consultation and cooperation between social media platforms, civil society and government

[62] Kaye, Report of the Special Rapporteur.

[63] The Santa Clara Principles on Transparency and Accountability in Content Moderation.

On issues of harmful speech, the MSC would work complementarily with government agencies that hold individuals and institutions responsible for violating criminal laws (e.g. hate speech, uttering threats, or libel) and electoral laws (e.g. foreign inducement or campaign spending).

The MSC would build on the legacy of proactive Canadian cultural policy and work to advance well-understood government commitments. The government should review and offer guidance as a working paper on the implications of the Multiculturalism Act on content moderation. Further, a task force of the council could develop best practices for dealing with content from Canada's many nations, part of the government's ongoing commitment to Truth and Reconciliation. Training materials could be developed for moderators to help understand the inequitable and implicit biases, including through complaint-driven enforcement and advances in automation.

The MSC would signal that significant responsibilities come with greater editorial management of content and facilitation of mass communication. Higher standards would be applied to companies like Facebook, Twitter and YouTube, which actively intervene in the visibility of content and mediate access to large audiences. By contrast, less demanding standards may apply to platforms that perform a function similar to traditional common carriers, or those that only support one-to-few communications. Doing so may encourage applications like WhatsApp or Signal, which provide valuable opportunities for encrypted communication, to limit their capacities for viral dissemination of content to large audiences, since this is increasingly being used to foment violence and disinformation.

The MSC could be implemented and enforced in several ways, with more or less government involvement.

We recommend that the Canadian Radio-television and Telecommunications Commission (CRTC) require large social media platforms to be members and address a range of forms of illegal speech according to set criteria. In this scenario, the CRTC would apply a new group-based approach, as proposed in its recent report to government. The CRTC suggests that: "broadly based agreements tailored to and established with a few dozen specific companies or affiliated groups of companies, individually or collectively offering a variety of services (service groups)" might be used to provide "public scrutiny and should set out specific binding commitments applicable to the service group."[64] While we caution against using this experimental framework for online broadcasting or digital common carriers, the content moderation industry could fit as a group in this new "nimble regulatory" approach. The MSC would then be responsible for enforcing the group's service agreements.

The MSC could also be established through new legislation as a dedicated regulator, with the ability to enforce its own decisions independent of the CRTC.

Several further options exist, which entail less government oversight. The MSC could function like the Internet Corporation for Assigned Names and Numbers (ICANN) or Better Business Bureaus, as an

[64] Canadian Radio-television and Telecommunications Commission. 2018. Harnessing Change: The Future of Programming Distribution in Canada.

independent non-profit body that develops and enforces standards on companies that choose to be members. Membership would be voluntary, but platforms would be encouraged by threats of government regulation if they do not take sufficient action. Finally, the MSC could exist as a purely advisory body, which would develop best practices that could be adopted on a voluntary basis.[65]

Will social media companies agree to help develop and be bound by a Moderation Standards Council in Canada? That will ultimately be up to the companies themselves, but their participation may offer a range of benefits. These include reduced exposure to regulation and liability enforcement by the government, provided they adhere to the council's recommendations and processes. The MSC would provide social media companies with better access to resources, best practices, and input from civil society organizations and researchers. It would also provide platforms with clearer guidelines and procedures that they could follow when addressing challenging and contentious questions of content moderation, including the creation of a legal and legitimate appeals body to handle difficult cases. Senior staff at leading social media companies have expressed a desire for governments to develop these processes, or at least a willingness to accept them.[66] For instance, Facebook CEO Mark Zuckerberg publicly declared:

> **"[O]ver the long term, what I'd really like to get to is an independent appeal [body]. So maybe folks at Facebook make the first decision based on the community standards that are outlined, and then people can get a second opinion. You can imagine some sort of structure, almost like a Supreme Court, that is made up of independent folks who don't work for Facebook, who ultimately make the final judgment call on what should be acceptable speech in a community that reflects the social norms and values of people all around the world."[67]**

While we propose a Canadian body to play this function, rather than a global body, the MSC captures the essence of Zuckerberg's vision.

[65] For a well-developed set of policies that companies may voluntarily adopt, see Change the Terms - Reducing Hate Online, developed by U.S. civil society organizations. These policies are recommended for social media platforms as well as companies engaged in activities such as online advertising, financing, or website hosting, but not for Internet Service Providers.

[66] Caplan, Content or Context Moderation.

[67] Klein, E. 2018. Mark Zuckerberg on Facebook's hardest year, and what comes next, Vox.

RECOMMENDATION #3:
BUILD CAPACITY IN CIVIL SOCIETY

Governments, foundations and internet companies in Canada should support and empower civil society to identify, investigate and respond to harmful speech online. Doing so requires financial investment, as well as increased transparency from government and private companies (particularly social media platforms). Ultimately, if our aim is evidence-based policy, researchers and civil society need evidence to investigate and capacity to evaluate that evidence.

Governments can assist in myriad ways, such as direct funding, indirect funding via research institutes and academic granting agencies, pressuring social media platforms to share data or provide reports to civil society, and addressing liability concerns of platform companies when they share data with researchers.

Build capacity for research and monitoring campaigns of disinformation, computational propaganda and harmful speech.

Canada lacks sufficient research and monitoring capacity in its academic and civil society sectors. Several models exist to support major projects in these areas without compromising research independence. Core funding can come through:

- **Research granting agencies.** For example, the Computational Propaganda Project at the University of Oxford is significantly funded by the European Research Council.
- **Issue-specific governmental program funding.** For example, the Media Against Hate project is a Europe-wide campaign led by the European Federation of Journalists and a coalition of civil society organizations, primarily funded by the EU's Rights, Equality and Citizenship Programme. Canadian examples include Public Safety's Community Resilience Fund and Kanishka Project.
- **Private foundations and private corporations.** For example, First Draft News at Harvard University, which is funded by foundations including the Ford Foundation and the John S. and James L. Knight Foundation, as well as by internet companies including the Google News Initiative and Facebook Journalism Project.

Support or encourage the creation of an "election contact group" in Canada to improve communication between civil society and social media platforms before and during elections.

During elections there is often need for quick and trusted exchanges of information between social media platforms and stakeholders, including political parties, civic tech organizations, election management and monitoring organizations, and journalism outlets. Such exchanges can improve the capacity for actors to address issues quickly as they arise, and to avoid the appearance of undue or opaque influence. Canada could also support civil society engagement in a global election integrity contact group. The U.S.-based Democratic National Institute is developing a similar idea to operate at a global level, called the Design 4 Democracy Coalition.

Build civil society and political party capacity to address online abuse and cyber security risks

Political candidates, members of human rights organizations, journalists and other publicly engaged individuals need more assistance and resources to address online abuse. Governments, political parties, donors and private companies should support research on this abuse and its consequences, and should develop and share resources and best practices, especially for members of groups facing particularly high incidents of hate speech and harassment.

Harmful speech online is often paired with attempts at cyber-security breaches. The University of Toronto's Citizen Lab is a world leader in investigating possible breaches, but it can only investigate a fraction of potential cases. In addition to support for university-based research teams like the Citizen Lab and non-government organizations that provide security assistance (such as Access Now's Digital Security Helpline), government and private industry could develop pro bono programs to enable private security researchers to provide part-time assistance to civil society groups.

Develop a Canadian research network to study how AI may promote or counteract hate speech and discrimination.

As social media companies and other tech companies increasingly rely on AI to moderate and share content, they generate new risks for propagating harmful speech or group discrimination. The lack of training data sets in Canada, and the possible biases in existing data sets, contribute to such risks. In particular, it is unlikely that conventional machine learning strategies can address harmful speech, given that it is often highly contextual, euphemistic, and tailored to particular target groups. Civil society groups and researchers can help identity and address these problems. They can do so by increasing the quantity and quality of research of the algorithms, or the design and impact of algorithms, provided they can have greater transparency from social media platforms. Civil society and researchers can also help develop new approaches to supervised machine learning, and new data sets for training, that can do a better job of identifying harmful speech. One example of this in the U.S. is the Online Hate Index, a research partnership between the Anti-Defamation League and the University of California, Berkeley. While we recommend caution regarding AI tools to automatically filter harmful speech, such tools can be used to better monitor and measure it.

In a broader context, Canada needs one or more major research institutes to address the ethical, human rights and democracy issues of AI, such as the U.K. is pursuing via its new Ada Lovelace Institute.

Support organizations that develop civic tech in Canada.

To address harmful speech and support democratic participation and deliberation, Canada needs a robust "civic tech" sector to develop and promote technological and social innovation. Examples in the U.S. include Online SOS, a platform that helps the targets of harassment receive immediate online assistance to document and respond to the abuse, and HeartMob, which helps the targets of harassment to crowdsource assistance from its online community. Civic tech efforts often need more seed funding, more assistance to grow to scale, and more research on their efficacy.

Support the creation and use of social media companies' research repositories.

Government should encourage and support the creation of research repositories, so that researchers have adequate access to social media platforms' data, processes and/or algorithms in order to evaluate their impact on Canadian democracy. Along these lines, U.S. Senator Mark Warner has proposed a public interest data access bill in a recent white paper.[68] Government should also clarify how it will address social media companies' legitimate concerns about sharing proprietary information or users' private data. This could follow similar access procedures to researching confidential medical records, as currently happens with research supported by Canadian federal funding agencies like the Canadian Institutes of Health Research (CIHR).

Companies have taken different approaches regarding accessibility to researchers. One important development is Facebook's new industry-academy partnership, Social Science One. The project uses multiple committees of top-notch researchers to adjudicate researchers' proposals and enable unprecedented access to Facebook data, while being responsive to legitimate obstacles that Facebook faces. This is a step in the right direction, but we don't yet know how well it will work and there remain questions about access and sustainability. For Canada specifically, it is worth noting that there does not seem to be a single academic from a Canadian university sitting on any committee of Social Science One.

Ultimately, researchers need sufficient information to understand how social media are shaping how Canadians participate in politics and communicate with one another. This requires either that all social media companies operating in Canada individually provide much greater and more responsive access to information for researchers, or that government—in partnership with major federal research agencies—create a national repository containing information from all platform companies.

[68] Warner, Potential Policy Proposals.

CONCLUSION

In response to increasing online hate speech, the Chief Commissioner of the Canadian Human Rights Commission observed:

> **"It threatens our public safety, it threatens our democracy, it threatens our diversity… Canada needs reasonable protections that evolve at the same pace as our technology and social media."**[69]

We agree. Harmful speech, including but not limited to hate speech, can undermine the quality of democratic participation and processes. This report synthesizes current knowledge, but it is clear that much remains to be learned—about how harmful speech is propagated, about how it affects political inclusion and deliberation, and about how different policies might address it.

This report's recommendations are primarily oriented toward creating legitimate and effective *processes* for developing meaningful policies in these areas, rather than proposing the final policies ourselves. Furthermore, any policies in this area will need to consider developments in other policies that seek to protect democratic processes—including on issues of advertising transparency, campaign spending, limitations on foreign interference, and security measures to address foreign interference. Similar recommendations may also be used to address another problem for content moderation—the spread of disinformation.

Our recommendations try to strike an appropriate balance between protection of free expression and other rights, industry viability, and the necessary capacity for a democratic country to foster fair, inclusive and robust debate among citizens.

Given the serious challenges that harmful speech poses to democracy in Canada, including the upcoming 2019 federal election, bold and swift action is necessary.

[69] Landry, M. 2018. Statement: The Internet shouldn't be a safe space for hate. Canadian Human Rights Commission.

PUBLIC
POLICY
FORUM