Research Report - December 2020

Processes, People, and Public Accountability: How to Understand and Address Harmful Communication Online

Chris Tenove, Heidi Tworek

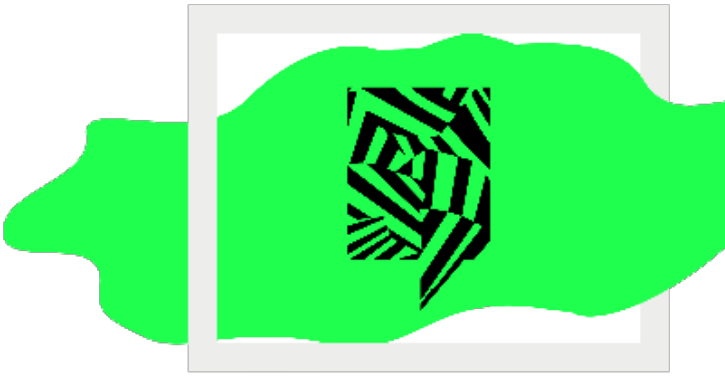
PUBLIC POLICY FORUM
FORUM DES POLITIQUES PUBLIQUES
PPF

Centre for MEDIA, TECHNOLOGY and DEMOCRACY

McGill

MAX BELL SCHOOL of PUBLIC POLICY

Table of Contents



About the Initiative

DEMOCRATIC EXPRESSION DÉMOCRATIQUE

The Canadian Commission on Democratic Expression is a three-year initiative, led by the Public Policy Forum that aims to bring a concerted and disciplined review of the state of Canadian democracy and how it can be strengthened. The centerpiece is a small, deliberative Commission which will draw on available and original research, the insights of experts and the deliberations of a representative Citizen's Assembly to assess what to do about online harms and how to buttress the public good. The Commission is designed to offer insights and policy options on an annual basis that support the cause of Canada's democracy and social cohesion. The Commission is supported by national citizen assemblies as well as by an independent research program.

This initiative grew out of earlier insights about the relationship of digital technologies to Canada's democracy covered by the Public Policy Forum's ground-breaking report, The Shattered Mirror and its subsequent interdisciplinary research outlined in the Democracy Divided report (with UBC) and through the Digital Democracy Project partnership with McGill university.

The initiative is stewarded by Executive Director, Michel Cormier and delivered in partnership with MASS LBP and the Centre for Media, Technology and Democracy at McGill University's Max Bell School of Public Policy, who are executing the national citizen assemblies and research program, respectively.

To learn more about the initiative and how you can become involved, please visit www.ppforum.ca. The initiative will run from April 2020 to March 2023.

This project has been made possible in part by the Government of Canada.
PPF would also like to thank the McConnell Foundation for their support.

Canada La Fondation McConnell Foundation

About the Authors

Chris Tenove

Postdoctoral Fellow, University of British Columbia (UBC)

Dr. Chris Tenove is a Postdoctoral Fellow in the Department of Political Science at the University of British Columbia. He researches political theory and international relations, with a focus on democratic participation, public policy and digital politics. He has several published and forthcoming peer-reviewed articles and chapters on the impact of disinformation and harmful speech on democracy. Together with Heidi Tworek, he leads an in-depth study of the online abuse of Canadian politicians. Tenove previously worked as an award-winning journalist and broadcaster.

Heidi Tworek

Associate Professor of International History at the University of British Columbia (UBC)

Dr. Heidi Tworek is Associate Professor of Public Policy and International History at the University of British Columbia, Vancouver. She is a non-resident fellow at the German Marshall Fund of the United States and the Canadian Global Affairs Institute. Her latest book, News from Germany: The Competition to Control World Communications, 1900-1945 (Harvard University Press, 2019), was awarded the Wiener Holocaust Library Fraenkel Prize and the Ralph Gomory Prize from the Business History Conference. Alongside her academic work, she has written multiple policy reports on hate speech, communications policy and digital democracy in Europe and North America. In May 2019, she testified on hate speech before the Canadian House of Commons Standing Committee on Justice and Human Rights. She was a steering committee member of the Transatlantic High-Level Working Group on Content Moderation and Freedom of Expression and she is a term member of the Council on Foreign Relations.

Introduction

Researchers and reporters documented three forms of harmful online communication during Canada's 2019 federal election campaign:

- *Abuse of individuals*: Minister Catherine McKenna received thousands of negative messages on social media during the campaign period, including threats of violence, which culminated in the defacement of her constituency office with misogynistic slurs.[1]

- *Intolerance and hate toward marginalized groups in public online spaces*: Significant volumes of intolerant content, ranging from casual use of dismissive terms to racist slurs and conspiracy theories, were directed toward Muslims and other social groups on Twitter, Facebook, Reddit and YouTube.[2,3]

- *Building support for hate in private online spaces*: White supremacist, ethnonationalist and anti-government networks, which included members of Canada's military, shared ideas and coordinated activities in private Facebook groups and online chatrooms.[4,5]

These cases show some of the myriad forms of online communication that may be considered harmful. These include forms of speech that are *already illegal* in Canada (e.g., uttering threats), instances of *harmful but not illegal* communication (e.g., anti-Muslim posts that don't reach the threshold of criminal hate propaganda) and harmful *patterns* of communication that contribute to systemic discrimination (e.g., large volumes of dismissive and disrespectful communication toward women).

In this report we propose a framework to distinguish key dimensions of harmful online communication in Canada. We summarize initial findings from our study of online abuse of political candidates in the 2019 federal election, which emphasizes how patterns of discourse and interactions between online and offline experiences may be harmful. We then analyze international policy responses by governments and social media companies. We conclude with several principles to guide policy development in Canada: 1) focus on systemic *processes* rather than individual pieces of content; 2) pay attention to the *people* who perpetrate, suffer and address harm, and not just to online spaces; and 3) promote *public accountability* – including, but not limited to, transparency – of regulators and platform companies. Although our report focuses on the negative impacts of online communications, we should not forget the potential benefits, including how social groups and political actors leverage these spaces for democratic purposes. Indeed, addressing harmful communication can promote a more just distribution of the benefits of internet use.

Dimensions of harmful communication online

Harmful online communication is a complex phenomenon: it includes factors such as algorithms and interface designs, rapidly changing usage patterns and interactions between different digital services. These new factors complicate the longstanding challenge of determining how messages and media affect people's beliefs and behaviors. It is difficult to conceptualize and measure harms from communication, but doing so is crucial for mitigating those harms and balancing any interventions against restrictions to free expression and other goods. We focus on potential harms from hosting and disseminating user-generated content, but recognize that there may also be harms from data collection and exploitation, tax avoidance and market dominance.

General categories of harmful communication

Digital media researchers have focused on several categories of harmful political communication. The broadest category is incivility. While *incivility* may be understood as violating expectations of politeness, in democratic societies we are often more concerned about violations of speech norms necessary for deliberation. These violations can include unjustified expressions of disrespect toward individuals and groups, or attempts to shut down, intimidate or otherwise silence people with different views. What counts as uncivil can vary significantly according to context and incivility may sometimes be appropriate

or justifiable, such as when protestors interrupt a public event to bring attention to injustice or when speakers angrily express their moral outrage at wrongdoing.

Abuse or *harassment* are more extreme forms of incivility, which go beyond issues of politeness and include insults and threats toward individuals. From the perspective of democratic inclusion, abuse and harassment can function as *intolerant discourse*, that which attacks or characterizes individuals or groups in ways that may damage their fair participation in a democratic society.[6]

Hate speech is generally understood as discourse that aims to denigrate, threaten, or deeply insult people according to their identity or social group affiliations, though its exact definition varies widely in research and legislation.[7] Facebook defines hate speech as a "direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability."[8] Hate speech may arguably be used against any social group, including white men (as Facebook has decided), but it is more damaging to political participation or wellbeing when directed at marginalized groups and identities. Hate speech can harm individuals *and* democratic society more broadly. It frequently does so by using terms (such as racial slurs) or claims (such as that women or gender minorities are immoral or less competent) that have been traditionally employed to subordinate, denigrate, or justify violence toward a group.[9] Hate speech thus draws on and reinforces "systemic discrimination" against a group.[10]

All of these categories of harmful communication affect Canadians. A 2016 Angus Reid Survey found that the harassment rate experienced by all social media users was 31% and that rates were higher among young people (44% among those aged 18-34), visible minorities (38%) and LGBTQ people (58%). Men and women reported online harassment at similar rates, but women were twice as likely to report being stalked or sexually harassed and that social media harassment had an impact on their offline lives (28% vs. 19%).[11] The survey also revealed self-censorship: 61% of social media users said they have decided not to share some content to avoid unwelcome responses.

Challenges of identifying harmful online communication

Identifying harmful communication online, even the narrower category of hate speech, is challenging for both human evaluators and content moderation algorithms.

First, hate speech is dynamic, contextual and often ambiguous:

- Actors who promote hate use new terms, euphemisms and memes to avoid detection. They may also veil their communication as satire or humour to evade responsibility.

- Hate speech can include "positive" messages. Groups promoting hate ideologies, such as white supremacist and incel movements, both denigrate the targeted group (e.g., visible ethnic minorities, women or non-binary individuals) and emphasize the special value of their own group (e.g., white, male).

- The same term can be used in different contexts or by different actors and mean different things. For example, terms associated with hate speech may be used in counter-speech by anti-hate activists, or may be used by a members of a community (e.g., African Americans) in ways that would be considered hateful if used by non-members.

[H]ate speech is dynamic, contextual, and often ambiguous

- These challenges are exacerbated when processes to find hate speech are not sufficiently inclusive. Most obviously, US-based platforms have long struggled to address hate speech in languages other than English, something that Mark Zuckerberg himself admitted in a statement to a US Senate Committee in 2018.[12]

Second, these difficulties are magnified in attempts to identify hate speech "at scale" in databases or on social media platforms. Researchers often employ "supervised machine learning" approaches, which use training data labeled by human beings to develop and test complex algorithms to identify hate speech.[13,14] These algorithms can include components such as lists of terms, specific orders of words and sentiment

analysis (e.g., terms and text patterns indicating emotions like anger or revulsion). While these approaches are increasingly sophisticated, they may inherit or magnify limitations of human interpreters. For example, algorithmic systems have been shown to inappropriately flag African American discourse as hateful since it uses terms that would be considered derogatory if used by non-African American speakers.[15]

Third, researchers tend to examine the text of individual messages on single platforms, isolated from its context. This focus on individual messages is attractive to social media companies – it is easier to set and enforce rules that equate hate speech with discrete messages, which are more easily detectable by algorithms.[16] However, any harm may only be detectable by assessing broader contexts, such as:

- *Relationships* between users over time. An innocuous message can be part of a threatening pattern of harassment or hate. Social media platforms' focus on messages in isolation has reduced their capacity to recognize harassment over time, such as cyber-stalking.

- *Forums and platforms* may support patterns of communication and user experiences that are threatening, discriminatory or risky for members of certain groups.

- *Public discourse*: Looking at communication across social media platforms and other media and public spaces, researchers can identify tropes, narratives and other features that may have discriminatory impacts on people with different identities or experiences. For instance, anti-Muslim and anti-immigrant content is relatively common on social media, websites and some news media.[17,18,19]

Fourth, hate speech can be targeted or disseminated in different ways and in different online spaces:

- *Victim-targeted*: Actors directly target individuals or members of a group with harmful messages, including those sent via email, messaging apps and messaging functions that overlay social media or gaming platforms (e.g., WhatsApp, Facebook Messenger, Direct Messaging on Twitter, LinkedIn messages, Xbox).

- *Broadcast*: Actors post content to public media spaces (e.g., Twitter, Reddit, public Facebook pages and groups) which may be seen by targets, bystanders or the actor's sympathizers. This content may affect the targeted individual or group; may promote harmful narratives, images or opinions in broader populations; or may attract or mobilize supporters of a hateful ideology.

- *In-group limited*: Content may be disseminated among users likely to sympathize with a hateful ideology. Such content is less likely to be flagged to platforms or authorities, particularly when communicated via private, closed, or dark spaces (e.g., private Facebook groups; WhatsApp or Telegram groups; or restricted online chatrooms and forums).

Actors promoting discriminatory or hateful ideologies frequently leverage multiple platforms, using both private and public spaces to coordinate activities and attack individuals or groups. For example, a recent study of right-wing extremism in Canada found over 6,000 extremist channels, pages, groups and accounts across Facebook, Twitter, YouTube, 4Chan and smaller sites such as Gab, which promoted content that was anti-immigrant, anti-Muslim, anti-Semitic, or misogynist and often was anti-government.[20] Social media platforms provide "avenues for a broad spectrum of right-wing extremists to mobilise by recruiting new members, broadcasting disinformation and propaganda, harassing opponents and co-ordinating activity including publicity stunts, protests and acts of violence."[21]

Policy responses should recognize and seek to address the relationships between online and offline harms

Policies limited to a single platform will likely be ineffective, especially because different platforms have different audiences, community standards and capacities for content moderation, including whether messages are encrypted or visible to platform managers and whether moderation is primarily done by users, staff or algorithmic systems.

Finally, online communication rarely causes harm entirely on its own, but instead does so through its relationship to offline messages, behaviors and experiences. For instance, political candidates interpret flirtatious messages differently if they have been the target of stalking and race-related messages differently if they face systemic discrimination offline. Policy responses should recognize and seek to address the relationships between online and offline harms.

Case study: Online abuse of candidates and elected officials in Canada

Social media channels are critical to electioneering and constituency engagement in Canada, but candidates and elected officials face abuse and threat at all levels of government. Globally, online abuse is recognized as a form of political violence and one that appears to be particularly damaging to women and members of marginalized groups.[22,23] The initiative to address "intimidation in public life" in the United Kingdom, led by an independent advisory body to the government, found MPs experienced "persistent, vile and shocking abuse," and concluded that "widespread use of social media platforms is the most significant factor driving the behaviour we are seeing."[24]

While there is much anecdotal evidence of online abuse of candidates and elected officials in Canada, there are few systematic studies,[25] and none which combine large-scale data analysis with in-depth interviews with candidates themselves. We therefore conducted a study of the 2019 federal election. We analyzed a set of approximately one million tweets directed at candidates between mid-August and October 31.[26] (We were only able to directly study Twitter messages, as Facebook does not enable researchers to collect comments on candidates' posts or pages.) We also interviewed over 30 candidates or their communication staff.[27] Here we highlight some preliminary findings:

What kinds of negative messaging do candidates encounter and how much do they receive?

Interviewees helped us create a three-tier assessment of negative messaging:

- *Low negativity* messages are not simply policy criticisms or salty language, but are dismissive or disrespectful toward the candidate or other individuals, e.g. "Oh how do I hate thee, @XXX" and "@XXX Maybe you can help reduce GHG's by not releasing so much hot air."

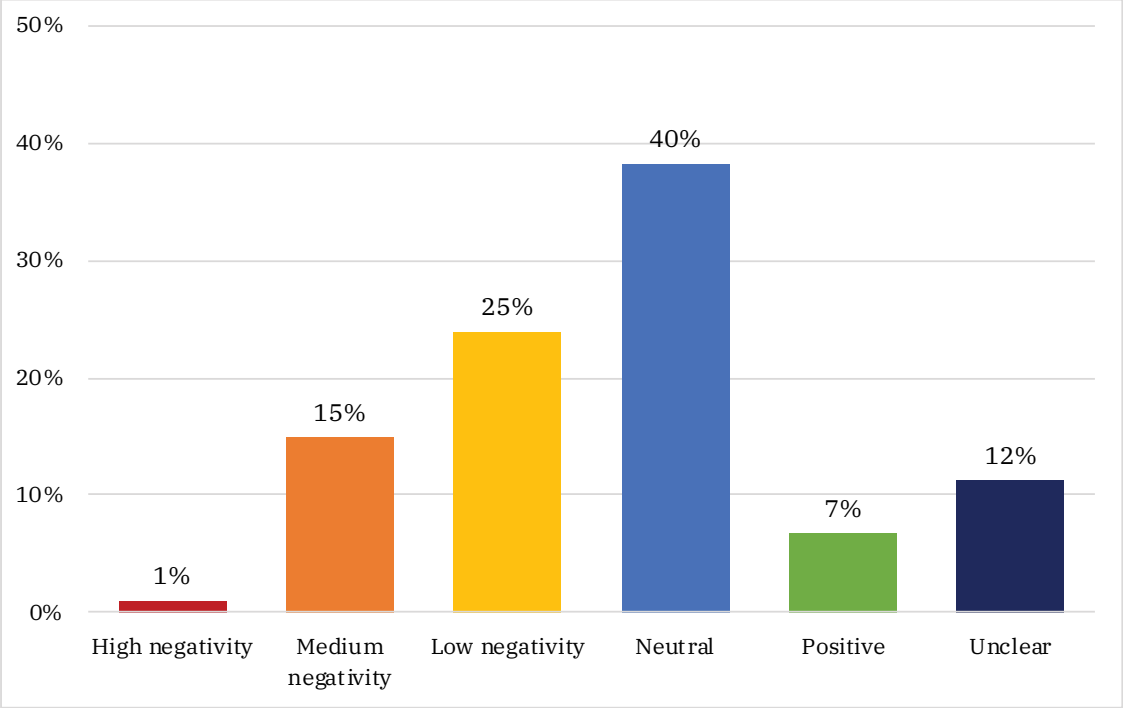
- *Medium negativity* messages are offensive, insulting, or advance negative stereotypes of social groups, e.g. "@XXX So the carbon tax will save the world. Infuckingcredible. You are the stupidest person to walk the planet" and "@XXX pander from the panzy."

- *High negativity* messages include hateful language, threats, or unsubstantiated accusations of moral or criminal wrongdoing directed at certain social groups, e.g. "@XXX Don't forget to take your antidepressants pills bitch" or "@XXX Eat a dick you Canadian fag. If it wasn't for us you'd be nothing. Die."

Our team read and categorized over 3,300 tweets, using those categories as well as "neutral," "positive," or "unclear" (meaning it was impossible to determine the tweet's sentiment, e.g. it only contained URLs). We used this data to train and test a machine-learning model. We then used this model to classify all tweets in our corpus of approximately one million messages.[28] The results we present here are preliminary.[29] We will fine-tune our model, further analyze its outputs and publish peer-reviewed findings.

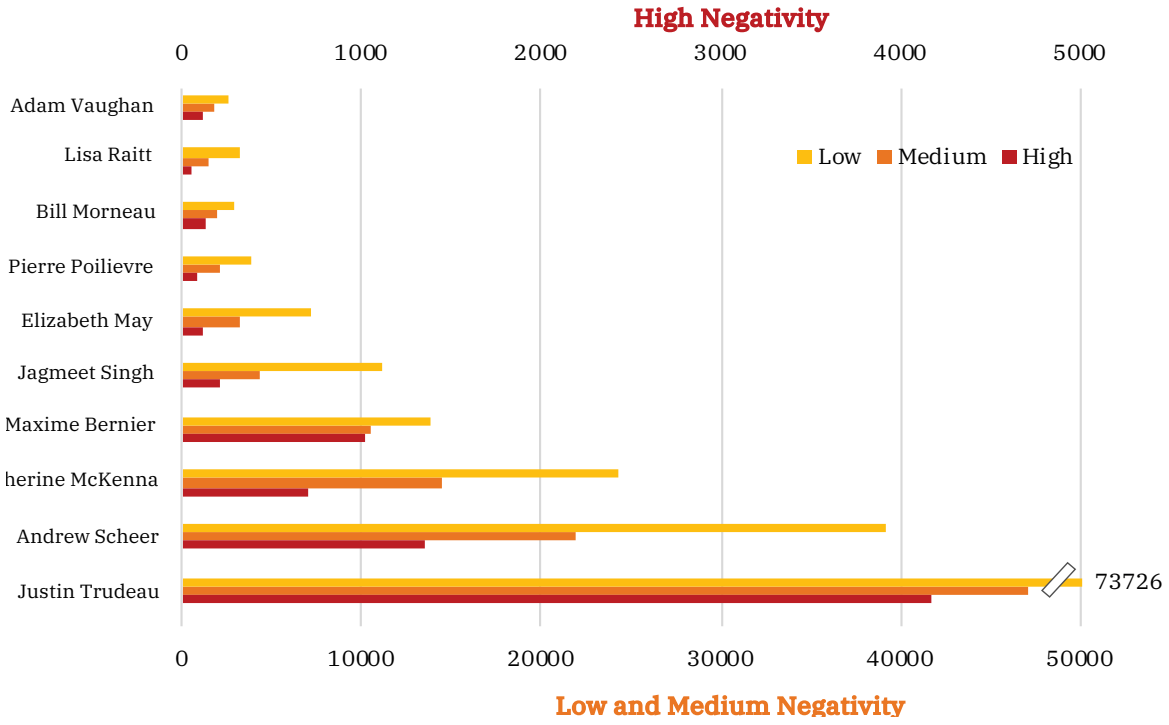
As Table 1 shows, about 40% of all tweets at candidates are *uncivil* (including low, medium and high negativity messages); about 16% *abusive* or *intolerant* (medium and high negativity) and about 1% explicitly hateful, threatening or potentially defamatory (high negativity). While 1% is a low proportion, it suggests that approximately 10,000 such messages exist among the 1 million tweets directed at candidates.

Figure 1: Overall prevalence of negative tweets



The volume of negative messages faced by candidates varies enormously. Table 2 shows the ten candidates who received the most negative tweets. They include party leaders (the Bloc Québécois leader did not receive significant numbers of English tweets), as well as Liberal cabinet ministers and prominent Conservative candidates. To clarify the extreme skewing in volumes of negative messages, compare our estimates of negative tweets for Justin Trudeau, #1 in this list (125,031 negative tweets), Adam Vaughan at #10 (4,445 negative tweets), and the #100 candidate 242 negative tweets.

Figure 2: Candidates who received the most negative tweets



How were candidates affected by negative messaging?

Candidates told us that high negativity messages affected their sense of wellbeing and security for themselves and their staff and could require time-consuming engagements with police or civil servants to address security concerns. Medium and low negativity messages, when encountered at high volume, could be demoralizing for candidates, campaign staff and volunteers.

Candidates and their campaign teams develop strategies, often in an ad hoc and reactive way, to manage abuse and harassment. They do so to defend themselves and their image, but also to protect the online space they cultivate to engage the public. Many expressed frustration that negative messaging directed at them or at social groups was undermining possibilities for productive political debate, exacerbating partisanship and limiting opportunities to engage with constituents.

The relationship between gender or visible minority status and negative messaging is complex. Recent research suggests there is no simple correlation between the volume of negative tweets and Canadian politicians' perceived gender or race.[30, 31] Nor did our study.[32] However, interviewees explained that online messages had a greater impact when they reinforced or echoed experiences of risk and discrimination. Or, as one MP observed, in a context of recent episodes of misogynist and racist violence in Canada, threats and insults on Twitter had her "looking over my shoulder." She explained: "You don't know when the person behind an account may turn violent. You don't know if this faceless avatar on Twitter is someone who may one day be triggered and come after me."

Candidates desire greater accountability for users who post abusive, threatening, or hateful language. This unaccountability is not simply the result of anonymity, since accounts apparently linked to real identities on Twitter and Facebook are responsible for significant amounts of high negativity messages. One interviewee observed:

> On Facebook you have a guy who lives in southwestern Ontario and his profile picture is of him and his daughter on the first day of Grade 1, and he is using hateful, misogynistic, violent language … We could block that account, which might cause us other problems, but there is no way to call that individual out and have them take responsibility.

Many candidates expressed frustration with the limited scope of actions they could take when facing highly negative messages or high volumes of abusive content. Criminal charges are very rare (and arguably rarely appropriate) and civil actions are extremely slow and costly. Some candidates echoed the Law Commission of Ontario: "there is currently no practical legal remedy available to many Ontarians victimized by online defamation." [33] Furthermore, candidates said that some of

the most problematic language may come from people with mental health issues, making interventions other than legal actions or platform bans desirable.

Many candidates claimed that social media companies, too, should be exposed to greater accountability for the volumes of abusive or hateful speech that they allow. Interviewees across parties also voiced the need for more transparency and procedural accountability for material that platforms remove. Candidates expressed their concerns that private online groups, or viral content on encrypted messaging apps, were facilitating the coordination and spread of hateful material, such that they only saw the tip of a toxic iceberg. Interviews conducted with women in North America and Europe indicate that what we see in Canada echoes a broader international pattern of structural problems.[34]

Policy responses to harmful online communication

The preceding sections suggest several challenges for addressing harmful online communication:

- Communication may be harmful at the level of a message, an exchange over time, or a pattern of communication on one platform or across media spaces.
- Some forms of harmful communication are relatively easy to identify, like a clear threat of violence. However, many are difficult to interpret for both humans and algorithms.
- Different social media platforms and online spaces facilitate different forms of harmful communication and present different challenges to identifying and addressing it.
- The harm of online communication is often related to offline experiences and behaviors, both at the individual level and through patterns of systemic discrimination.
- There are numerous obstacles to accountability for harmful communication. For instance, users may employ fake identities or engage in groups on dark online spaces. Internet intermediaries do not provide enough information to identify harmful patterns of content or enforce their content moderation rules.

A multi-dimensional policy framework will be needed to address the multi-dimensional problem of harmful online communication. The framework should be responsive to the diverse experiences and vulnerabilities of Canadians and fit within a constitutional framework that protects equality and multi-culturalism as well as freedom of expression.[35] We will examine several policies implemented outside of Canada to identify elements that may or may not be appropriate here. Before doing so, we will first address the argument that no new policies are necessary.

Shouldn't we just enforce existing laws?

Some commentators have argued it is hazardous to regulate any speech that is not already illegal, other than through better enforcement of existing laws.

In Canadian criminal and civil law, many forms of communication are illegal, from uttering threats to defamatory libel to hate propaganda. These laws set a high standard of proof regarding the intentional actions of particular individuals. We agree that only severe and clearly-defined forms of communication should be subject to legal sanction, whether criminal or civil. We do not support holding platforms responsible for all activities of their users. However, we argue that existing laws are insufficient for four reasons.

First, the categories of illegal communication and harmful-but-legal communication are not and should not be static. Laws develop in particular social, political and legal contexts that may have sidelined the views and experiences of marginalized groups. Former BC Supreme Court justice Lynn Smith observes that policies to address online hate and misogyny need to take account of "the weight Canadian courts give to equality, including gender equality, when they are required to strike a balance between equality and other Charter rights or values, such as the right to freedom of expression." [35] We need ongoing and inclusive debates about forms of communication that should be treated as objectionable, harmful, or impermissible.

First, the categories of illegal communication and harmful-but-legal communication are not and should not be static

Second, research by us and others suggests that people's political participation and wellbeing are undermined by content that is and should be legal, due to the volumes of content they encounter and the relationship between online communication and offline contexts. We need to encourage the development

and testing of interventions by platforms, users, civil society groups and other stakeholders to address these complex causes of harm.

Third, harmful communication is not necessarily best addressed by removing it or sanctioning its propagators. Research with diverse groups suggests that "a one-size-fits-all approach to online harassment may fail to support some users while privileging others." [36] Rather than focusing on the removal of problematic content or users, responses to harmful speech should include reparative measures such as apologies, mediation, or attempts to educate or rehabilitate offenders.

Fourth, it is not clear that platforms will act on harmful communication in an inclusive and effective manner without government intervention. Platform companies have strong business incentives to hide or disregard harms they contribute to and dominant platforms can do so with limited concerns that people will abandon their service. Moreover, some competitors to dominant platforms may facilitate harmful communication (as in the case of Gab). Sectoral regulation can both encourage standards for recalcitrant dominant platforms and avoid a race-to-the-bottom by some smaller platforms.

Learning from regulatory efforts outside Canada

Policy-makers are considering myriad solutions to addressing harmful online communication, including anti-trust and tax policies. However, those enacted so far have focused on addressing content on existing platforms. We have chosen four prominent examples of such interventions to demonstrate the range of policy options and to assess their efficacy.

1. European Union's Code of Conduct for Countering Illegal Hate Speech Online

In 2015-16, the European Commission cooperated with major companies like Facebook and Microsoft as well as civil society organizations to create a code of conduct around online hate speech. Signatories committed themselves to quickly review and act on notifications about content that may be hate speech. They also committed to vaguer provisions regarding public education and transparency.

The European Commission has claimed successes for the Code based upon the number of signatories and the increasing number of deletions within 24 hours.[37]

Both the development and impact of the Code have been criticized. First, the Code was not developed in a transparent, participatory and consultative manner. EU member states had no proper opportunity to shape the final text through their comments; free speech organizations do not seem to have been included in consultations; several digital rights organizations walked out of discussions because of concerns about a lack of transparency in the negotiations. Second, it did not include guarantees of due process to address concerns that undue deletions would limit freedom of expression. Third, removals are a problematic measure of success. The reports on the Code shows that removals have consistently increased over time, but is this an indication of success in fighting hate speech, in increasing production of hate speech, or that the metric itself is incentivizing over-deletion? We also do not know what types of hate speech were removed (e.g., anti-Semitic content, anti-Black content etc.). Fourth, there is no access for independent researchers to assess the reports or underlying data, or access to data on the broader patterns and impacts of removals.[38] Finally, while the Code offers rhetorical support for civil society actions against hate, it doesn't provide concrete assistance.

2. Germany's NetzDG: Online enforcement of pre-existing speech laws

The German parliament passed the Netzwerkdurchsetzungsgesetz (Network Enforcement Law, NetzDG for short) in 2017. While often called a hate speech law, NetzDG actually enforces 22 statutes of pre-existing German speech law online. NetzDG attracted global attention as the first major law to fine American-based social media companies for not adhering to national statutes. The law required companies with over two million unique users in Germany to act upon user complaints on pieces of content within 24 hours or face fines of up to 50 million Euros per post. Platforms that received over 100 complaints also had to produce a semi-annual transparency report – this included Instagram as well as Facebook. In summer 2020, the German government is updating the law to mandate that the companies supply information to the Federal Criminal Police Office about posts deleted for illegal content.

Because Germany was the first major country to create a law specifically addressing social media platforms, its approach has served as a lodestar for many other countries. Canada's Liberal Party suggested a similar policy in its election platform.[39] France's National Assembly in May 2020 adopted "Lutte contre la haine sur internet", but the French Constitutional Council struck down significant portions of that law in June. Authoritarian countries such as Russia have also claimed inspiration from NetzDG to mandate removal of politically-inconvenient content under vague and overly broad categories.[40] While many authoritarian countries were already pursuing such censorship measures, it is important to consider how democratic countries can set a higher standard of democratic accountability that is harder to co-opt for authoritarian purposes.

It is important to consider how democratic countries can set a higher standard of democratic accountability that is harder to co-opt for authoritarian purposes

NetzDG has raised many concerns around freedom of expression, including that sanctions would incentivize companies to over-delete content.[41] It is hard to assess the veracity of these claims without further transparency on what is deleted. Paradoxically, complaints made under NetzDG are removed under platforms' terms of service. There are only a few posts that are allowed under platforms' terms of service, but not under NetzDG, which are then deleted because they violate NetzDG. We might then think of NetzDG as a "terms of service enforcement law," pushing companies to decide on complaints more quickly.

The NetzDG case provides additional lessons. First, companies can act far more swiftly than their initial rhetoric suggests. Facebook quickly hired several thousand more content moderators in Germany. Second, companies can produce detailed country-level reports if required. Still, these reports require monitoring for accuracy. The only fine issued under NetzDG was to Facebook in July 2019 for underreporting hate speech.[42] For all the controversy over NetzDG, most civil society groups agree that transparency reports are helpful, though they can be improved. Third, it is very hard to measure if deleting individual pieces of speech has affected the quality of political discourse, online activities of hate groups, or extremist beliefs. This is in part because NetzDG came into force around the time that Facebook dramatically reduced researchers' access to data. Fourth, subsequent analysis of NetzDG seems to have changed few minds on the law's efficacy. This could possibly have been addressed through greater inclusion in designing the law and more support for researchers and civil society to test its impact

Figure 3: Policy Responses

	EU Code of Conduct on Hate Speech	**NetzDG**	**UK Online Harms**	**Facebook Oversight Board**
Regulatory type	Mostly self-regulatory	Law	State-enforced (by independent agency)	Self-regulatory
Inclusive rule-making	Civil society engagement, though some groups walked out of consultations	No	Development responded to open comment process	Many consultations with researchers and civil society
Sanctions for non-compliance	Threat of statutory regulation	Fines	Warnings, notices, fines, possibly ISP blocking, senior manager liability	None
Transparency	Regular monitoring exercises by civil society organizations and public bodies	Semi-annual reports required for platforms receiving over 100 complaints	Reporting	Decisions made public
Appeal	No	Through companies first	Mechanism for companies	Acts as an appeals body
Unit Targeted	Individual posts	Individual posts	Online platform policies	Individual posts
Measure of success	Number of removed posts/complaints	Number of removed posts or complaints	Company compliance	Unlcear

3. United Kingdom's "online harms" approach

The UK government introduced a novel approach to regulation with its 2019 Online Harms White Paper. Among other things, the framework would place a "duty of care" on companies, requiring them to show that they have "appropriate systems and processes in place to react to concerns over harmful content and improve the safety of their users - from effective complaint mechanisms to transparent decision-making over actions taken in response to reports of harm."[43] This will likely entail creating codes of practice around various types of

harms. The framework will likely be overseen by Ofcom, an existing regulatory agency for communication.

New legislation may require companies to remove illegal content, such as terrorist or child sexual exploitation content, in a timely fashion. However, unlike NetzDG or the EU Code, the online harms approach is not primarily focused on removals of individual pieces of content. It will also compel companies to make clear and to enforce the boundaries of acceptable behaviour and content on their sites. Higher standards will be required to protect children. Companies will also be required to provide appropriate and swift redress mechanisms for users to report harmful content or to appeal a takedown. The regulator will be able to intervene with warnings, fines and potentially even liability for senior managers if companies do not live up to their promises.

Critics worry that the distinctions between illegal content and legal but harmful content, the notion of harms and the regulatory instruments themselves are too loosely defined. This has raised concerns that the approach could undermine rights to free expression, a concern the UK government has agreed to address, though the exact procedures are still in development. More broadly, the development of the online harms approach has several positive aspects. First, it acknowledges that regulation may be needed to address forms of harmful communication that are not illegal and has prompted a robust debate about how to do so without threatening legal rights to free expression. This includes suggestions about creating a more differentiated understanding of responsibility for and responses to online harms.[44] Second, the approach has emphasized improvements in transparency by platforms, both to users and to independent researchers.[45] Third, the government has so far pursued extensive and public consultations in developing the process.

4. Facebook Oversight Board

Over the past few years, Facebook has developed an Oversight Board to guide content moderation. The Oversight Board is operated by a trust that is independent from Facebook, although the first board members were vetted and chosen by the company. The Board will review specific pieces of content from around the world and decide if that content should be removed or not. Their judgments will initially apply to Facebook and Instagram content, not material on WhatsApp. Facebook has stated that it will abide by those decisions, provided they do not conflict with applicable laws, and that they will serve as precedents. The first set of cases for review were announced on December 1, 2020. [46]

Much remains unknown about the Board's future processes and impacts. Can a board of 40 part-time members make decisions on cases from around the world in ways that are sensitive to context? How will these decisions actually shape Facebook policy across its various platforms? Will independent researchers be able to verify the effects of these decisions and policy changes on users? Even if the Board is successful on its own terms, a single platform approach may not address harmful communication in the digital media ecosystem. Hate propagators could leverage other platforms to push content across countries, languages and platforms, leading researchers to predict that "policing within a single platform (such as Facebook) can make matters worse, and will eventually generate global 'dark pools' in which online hate will flourish."[47]

The Oversight Board may introduce greater accountability and due process for content moderation decisions on Facebook. However, its occasional decisions may not significantly affect content on Facebook's platforms in Canada, let alone content on other platforms.

Recommendations

To address harmful communication as a multi-dimensional problem, responses should have the following aims:

- Develop policies transparently and through extensive and inclusive consultation;
- Increase accountability and disincentives for those who intentionally cause harm;
- Increase incentives for platforms to responsibly remove illegal communication and reduce the impact of legal but harmful communication, and do so in ways that foreground a human rights framework, including but not limited to freedom of expression;
- Focus on downstream consequences for individuals and for disadvantaged groups; and
- Provide clear consideration of what success might look like and how it could be measured and verified by independent researchers, including through robust transparency measures.

To achieve these aims, we suggest thinking about recommendations in three areas:

1. Address the systemic processes rather than individual pieces of content.

Policies should not be limited to acting on discrete cases of problematic speech, for several reasons:

- It is very difficult to identify communication that will cause harm without being attentive to context, including the identities of propagators and targets of communication;
- There are concerning examples of how deletion policies lead to suppression of speech from marginalized and racialized groups;
- Doing so may not help us to understand and address systemic harms (e.g., insulting or "medium negativity" content that is disproportionately targeted at some groups or that reinforces offline harms); and
- It over-emphasizes takedowns by platforms, when it may be more useful to give people more agency over what they can avoid or address themselves.

In other work, we have suggested the institution of a social media council.[48,49] This council would not just set a code of conduct, but also foster the creation of a public and inclusive process to do so. It could therefore help develop principles and best practices for effective, context-aware content moderation for the social media sector. Amongst other things, this approach would likely require having more content moderators who understand and are trained in Canadian law and context.

Whether it be a social media council or other solution, it is important to consider two additional issues. First, the process should be designed to avoid capture by companies, political factions and government agencies.[50] Second, initiatives have to consider how to accommodate small and medium enterprises, whether through sliding scales or other mechanisms, to avoid regulation that unintentionally locks in the big players.

2. Support people alongside working on content.

While much policy attention is rightly devoted to the online space, policies should also aim to support people who can help address harmful communication:

- Civil society groups and technologists that are developing innovative responses to harmful communication – support need not be limited to resources, but could also include access to data and incentives for platform companies to work collaboratively;

- Providers of social and psychological assistance to help affected individuals address the online and offline impacts of harmful communication; and

- Support better labour conditions for the content moderators employed by platform companies, who suffer serious consequences through work that aims to minimize harms to others.[51]

An important reframing of the issue of harmful communication is to take a reparative justice approach

An important reframing of the issue of harmful communication is to take a reparative justice approach. Content moderation processes are usually opaque to users, and in turn disguise platforms' power and reduce possibilities for recognition and satisfaction for targeted individuals. A reparative approach would add additional interventions to the current focus on removing content and occasionally punishing "perpetrators." Processes like mediation and targeted education could emphasize instead the repair of individual wellbeing and social relationships.[52]

3. Measure efficacy as a core principle.

Developing and testing policies to address harmful communication will require greater transparency. For companies, this must go beyond treating content removals as a measure of success, but also include greater transparency about how harmful communication is disseminated or targeted, how it is identified and how content moderation policies are enforced. This and other information can enable independent researchers to better assess the downstream effects of potentially harmful communication and platform policies. Currently,

companies have incentives for ignorance, because it is harder to be accountable for harms they don't measure and because they face legitimate obstacles to sharing information about user activities.[53] One proposed solution is to create a transparency regulator to facilitate tiered levels of access for government agencies, independent researchers and the public.[54] Transparency mechanisms are also needed to ensure that relevant stakeholders know how government agencies or self-regulatory bodies are developing and enforcing policies.

Ultimately, these transparency measures should help determine the positive and negative consequences of efforts to address harms from internet use and to help promote opportunities for more fair, equitable and democratic online communication.

Conclusion

Harmful communication online is a moving target: it evolves swiftly, may look different to different groups on different platforms and requires significant research to conceptualize and investigate. Here, we suggest that policymaking should move beyond a focus on individual pieces of content. While individual pieces of content matter, we risk missing the wood for the trees. Instead, we suggest approaches that focus on *processes*, *people* and *public accountability*, rather than particular posts. For instance, to help address the abuse and incivility faced by Canadian public figures, we would suggest: 1) the creation of a social media council or similar public forums to develop Canada-specific principles for reducing harmful communication; 2) greater support and more tools for candidates to manage incivility and occasional threats; 3) greater efforts to understand, educate and, if needed, hold to account individuals who promote hate in public and private spaces; and 4) greater opportunities for independent researchers to access the data needed to evaluate whether these efforts work.

The sad truth is that harmful communication will never disappear entirely. But we can implement policies that dramatically reduce its impact both online and offline.

Endnotes

1 Connolly, A. (2019). "It needs to stop": McKenna slams political vitriol after office defaced with vulgar slur, Global News, 24 October.

2 Bellemare, A. (2019). How a misleading YouTube video is stoking fears about Shariah law before the election, CBC News, 22 July.

3 Davey, J., Hart, M. and Guerin, C. (2020). An Online Environmental Scan of Right-wing Extremism in Canada. London, UK: Institute for Strategic Dialogue.

4 Davey, Hart, and Guerin, "An Online Environmental Scan of Right-Wing Extremism in Canada."

5 Thorpe, R. (2019). Homegrown hate: Inside a neo-Nazi group attempting to gain a foothold in Winnipeg and across the country, Winnipeg Free Press, 16 August.

6 Rossini, P. (2020). Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk, Communication Research.

7 For legal definitions in Canada, see Gill, L. (2020). Background Report, Canadian Commission on Democratic Expression.

8 Community Standards. Facebook.com.

9 For more, see the "Hallmarks of Hate" identified by the Canadian Human Rights Tribunal (2006) Warman v. Kouba.

10 Gelber, K. (2019). Differentiating hate speech: a systemic discrimination approach, Critical Review of International Social and Political Philosophy.

11 Angus Reid Institute. (2016). Trolls and tribulations: One-in-four Canadians say they're being harassed on social media, 21 October.

12 Alba, D. (2018). Why Facebook Will Never Fully Solve Its Problems with AI, BuzzFeed News, 11 April.

13 MacAvaney, S. et al. (2019). Hate speech detection: Challenges and solutions, PLoS One, 14(8).

14 Siegel, A. Online Hate Speech, in Tucker, J. and Persily, N. (eds.) Social Media and Democracy: The State of the Field. Cambridge University Press.

15 Davidson, T., Bhattacharya, D. and Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets, arXiv preprint.

16 As Ananny aptly observes: "Categories are crucial parts of infrastructures because they define people and data into predictable units that can be aggregated, combined, and analyzed." Ananny, M. (2020) Public Interest and Media Infrastructures: Regulating the Technology Companies that Make 'Pictures in Our Heads.' Canadian Commission on Democratic Expression.

17 Farokhi, Z. (2020). Rise of Islamophobic Emotional Rhetoric During the 2019 Canadian Federal Election, in Dubois, E. and Owen, T. (eds.) Understanding the Digital Ecosystem: Findings from the 2019 Federal Election. Ottawa, ON: Digital Ecosystem Research Challenge, pp. 38–40.

18 Lamoureux, M. (2019). YouTube Pulls Canadian Anti-Islam Vlogger Following Huge Defamation Loss, Vice, 14 May.

19 National NewsMedia Council (2018). 2018-59: Hunter vs Toronto Sun, 11 December.

20 Davey, Hart, and Guerin, "An Online Environmental Scan of Right-Wing Extremism in Canada."

21 Ibid, p. 4.

22 Amnesty International (2018). Troll Patrol Findings: Using Crowdsourcing, Data Science & Machine Learning to Measure Violence and Abuse against Women on Twitter. London, UK: Amnesty International.

23 Sobieraj, S. (2020). Credible Threat: Attacks Against Women Online and the Future of Democracy. Oxford University Press.

24 Committee on Standards in Public Life (2017). Intimidation in Public Life: A Review by the Committee on Standards in Public Life. United Kingdom Parliament, p. 7.

25 Though see: Rheault, L., Rayment, E. and Musulan, A. (2019). Politicians in the line of fire: Incivility and the treatment of women on social media, Research & Politics, 6(1), pp. 1–7.

26 The 1,021,803 tweets we collected were directed at candidates who were declared by their party and had a public Twitter account by August 31. We collected English language tweets using Twitter's Streaming API. For full results, see Tenove, C., and H. Tworek. (2020) Trolled on the Campaign Trail: Online Incivility and Abuse in Canadian Politics. Vancouver: Centre for the Study of Democratic Institutions, University of British Columbia.

27 We interviewed Conservative, Green, Liberal and NDP candidates and their staff. Over 50% are women and one-third are racialized. We also interviewed elected officials at other levels of government who have experienced significant online abuse, such as former premier Kathleen Wynne.

28 Special thanks to our colleague Trevor Deley for conducting this quantitative analysis, and to Grace Lore for helping with interviews of election candidates.

29 To evaluate the accuracy of our machine learning model, we tested how reliably it assigns the same label to a tweet (e.g. "high negativity") as did our team of human coders. This is calculated as an F-score, with 1.0 being 100% reliable. Our F-score is 0.73 when our model seeks to distinguish all six categories.

30 Rheault et al ("Politicians in the Line of Fire") found that women politicians in general may not receive higher proportions of incivility, but high-profile women –cabinet ministers and premiers – did.

31 Gruzd, A. et al (2020). Toxic Interactions and Political Engagement on Twitter, in Dubois, E. and Owen, T. (eds.) Understanding the Digital Ecosystem: Findings from the 2019 Federal Election. Ottawa, ON: Digital Ecosystem Research Challenge, pp. 30–33.

32 The previously referenced studies, and ours, did not determine whether people with intersecting identities associated with marginalization are particularly targeted for incivility, such as non-heterosexual or racialized women. Research outside the Canadian context has found that to be the case. See Amnesty International's "Troll Patrol Findings".

33 Law Commission of Ontario (2020). Defamation Law in the Internet Age: Final Report. Toronto, ON: Law Commission of Ontario, p. 15.

34 Sobieraj, Credible Threat: Attacks Against Women Online and the Future of Democracy.

35 Smith, Q.C., T. H. L. (2020). Marlee Kline Lecture: Real Hate in a Virtual World: The Law and Cyber Misogyny. Allard School of Law, University of British Columbia, 28 January.

36 Schoenebeck, S., Haimson, O. L. and Nakamura, L. (2020). Drawing from justice theories to support targets of online harassment, New Media & Society, p. 15.

37 European External Action Service (2019). Countering illegal hate speech online – EU Code of Conduct ensures swift response, European Commission.

38 Bukovská, B. (2019). The European Commission's Code of Conduct for Countering Illegal Hate Speech Online: An analysis of freedom of expression implications. Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression.

39 Liberal Party of Canada (2019). Forward: A Real Plan for the Middle Class.

40 Mchangama, J. and Fiss, J. (2019). Analysis: The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship. Copenhagen: Justitia.

41 Tworek, H. J. S. and Leerssen, P. (2019). An Analysis of Germany's NetzDG Law. Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression.

42 Delcker, J. (2019). Germany fines Facebook €2M for violating hate speech law, POLITICO, 2 July.

43 Department for Digital, Culture, Media & Sport and Home Office (2020). Online Harms White Paper - Initial consultation response. United Kingdom Parliament, p.7.

44 Tambini, D. (2019). The differentiated duty of care: a response to the Online Harms White Paper, Journal of Media Law, 11(1), pp. 28–40.

45 Department for Digital, Culture, Media & Sport and Home Office, "Online Harms White Paper - Initial Consultation Response," 10–11.

46 Oversight Board (2020). Announcing the Oversight Board's First Cases and Appointment of Trustees. Oversight Board.

47 Johnson, N. F. et al. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology, Nature, 573(7773), pp. 261–65.

48 Tenove, C., Tworek, H. and McKelvey, F. (2018). Poisoning Democracy: How Canada Can Address Harmful Speech Online. Ottawa, ON: Public Policy Forum.

49 Tworek, H. et al (2020). Dispute Resolution and Content Moderation: Fair, Accountable, Independent, Transparent, and Effective. Amsterdam: Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression.

50 Tenove, C. (2020). Protecting Democracy from Disinformation: Normative Threats and Policy Responses, The International Journal of Press/Politics.

51 Roberts, S. T. (2019). Behind the Screen: Content Moderation in the Shadows of Social Media. New Haven: Yale University Press.

52 Schoenebeck, Haimson, and Nakamura, "Drawing from Justice Theories to Support Targets of Online Harassment.".

53 Tworek, H. J. S. (2019). Social Media Platforms and the Upside of Ignorance, in Big Data, Platform Governance, Internet Governance. Centre for International Governance Innovation.

54 MacCarthy, M. (2020). Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry. Amsterdam: Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression.

PUBLIC POLICY FORUM
FORUM DES POLITIQUES PUBLIQUES
PPF

Centre for MEDIA, TECHNOLOGY and DEMOCRACY

McGill

MAX BELL SCHOOL of PUBLIC POLICY