
# POLICY PAPER: OPPORTUNITIES TO BETTER COMBAT THE SPREAD OF ONLINE HATE IN CANADA

## Digital Peace Project

September 2023

Université Concordia University
MONTREAL INSTITUTE FOR GENOCIDE AND HUMAN RIGHTS STUDIES
Government of Canada / Gouvernement du Canada
Canada

# Table of Contents

## Introduction

The rise of hate speech poses a significant challenge to our collective digital landscape, requiring concerted efforts to address these issues while safeguarding freedom of speech. In response to this pressing concern, the Montreal Institute for Genocide and Human Rights Studies (MIGS) at Concordia University launched the Digital Peace Project. Supported by funding from the Department of Canadian Heritage, this initiative has played a role contributing to national efforts in mitigating online hate by engaging civil society actors—particularly those from ethnic, cultural, religious and visible minority communities—and empowering marginalized groups by amplifying their voices while seeking to increase our shared capacity to confront racism, discrimination and prejudice prevalent in the online sphere.

Building upon MIGS's decade-long commitment to combating offline and online hate, including projects such as the Digital Mass Atrocity Prevention Lab and the Canadian Task Force to Combat Online Antisemitism, the Digital Peace Project aligns with the U.N. Strategy and Plan of Action on Hate Speech. Its core mission has been to employ a multi-faceted approach to better understand and combat hate speech and online harms by encompassing a range of activities culminating in this final policy paper.

This policy paper presents research findings, and proposes policy implications and recommendations generated through the Digital Peace Project. By shedding light on the challenges posed by online hate and elucidating potential solutions, this policy paper seeks to serve as a valuable resource for policymakers, researchers, journalists and advocates committed to fostering a safer and more equitable digital ecosystem.

## Definitions

**2SLGBTQI+:** A variant acronym (Two spirit, Lesbian, Gay, Bisexual, Transgender, Queer, Intersex, Plus). This terminology is used in the policy paper to broadly describe any gender or sexual identity that falls outside a cis-gender and heterosexual framework (ie. someone who does not identify with gender assigned at birth and/or does not identify as being sexually attracted to the opposite sex).
**Algorithms:** the basis of computer programming and are used to solve problems ranging from simple sorting and searching to complex tasks such as artificial intelligence and machine learning
**Algorithmic Bias:** refers to the systemic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others. It's a prevalent concern today, with artificial intelligence (AI) and machine learning (ML) applications increasingly permeating every aspect of our lives.
**Digital Literacy:** the confident and critical use of digital technologies for information, communication and basic problem-solving. This includes using computers to retrieve, assess, store, produce, present and exchange information, and to communicate and participate in collaborative networks via the Internet.
**Hate Speech:** Any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor. This definition is found in the UN Strategy and Plan of Action on Hate Speech.
**Online Harms:** Online harms describes digital content that is harmful or broadly offensive in nature, but may not necessarily meet the recognized definition of Hate Speech. In scope, online harms may include hate speech, in addition to terrorist propaganda, violent content, child sexual exploitation and the non-consensual sharing of intimate images.

## Methodology

The Digital Peace Project focused on seven main buckets of work to help build awareness about online hate, contribute to national efforts to mitigate and moderate harmful content, bring together communities who often feel under-represented in these discussions and to give community partners a proactive role in solution finding exercises.

- CONSULTATIONS: First, MIGS conducted six consultations on project design. In closed-door sessions, staff and members of civil society were asked to contribute to the project design.
- PUBLIC SURVEY: After the consultations were completed, MIGS worked with RIWI to organize a public perception survey. Over 4,000 Canadians aged 16+ were asked about their perceptions on online hate and content moderation, including what people groups they believed were most often the targets of online hate and incitement to violence. The survey was conducted in two waves. Findings from the survey were shared in an online event and fed into the design of the virtual roundtables.
- VIRTUAL ROUNDTABLES: Five virtual roundtables were organized. Each roundtable focused on a particular people group that is frequently targeted by online hate: Women, 2SLGBTQI+ individuals (held in French), Indigenous Peoples, Religious Groups and Minorities and Racial and Ethnic Minorities. Community representatives, tech representatives and members of targeted communities participated on the panels. Each roundtable was live streamed onto Facebook, YouTube, Twitter and LinkedIn to reach as many Canadians as possible.
- FOCUS GROUPS: Actors from civil society, the tech sector and policy makers came together in two closed-door sessions to discuss the policy implications of the public perception survey to help guide the recommendations in the policy paper. Nine attendees were present at the first focus group. Eight attendees were present at the second focus group.
- PODCASTS: MIGS organized a podcast series featuring interviews with experts on online hate, community members and youth engagement professionals.
- SOCIAL MEDIA MONITORING BOT: In partnership with Areto Labs, MIGS launched a social media monitoring bot which is looking at the narratives expressed towards Canadian politicians, particular those who identify as women or other marginalized groups to learn from the speech that is directed at them as public facing figures of their identity groups.
- POLICY PAPER: The writing of this paper was guided by the above activities. It is our objective that this paper be useful in informing the decisions of Canadian policymakers.

The next sections will expand upon the findings of the social media monitoring bot, the public survey, virtual roundtables and the focus group sessions.

## Mapping Canadians' Experiences with Online Hate

### a. Public Perception Survey

A survey of 4,002 Canadians found that there is a tendency to conflate hate speech with comments that may be personally offensive and other online harms[1]. This may indicate a disconnect between legal definitions and the popular experience, or it may indicate that people living in Canada weigh Online Harms as holding comparable severity to Hate Speech. Among civil society actors, there could be an opportunity for additional research on perceptions of online hate speech among Canadians, or for the implementation of educational programs addressing Hate Speech and Online Harms in the contemporary digital landscape.

When asked what the most common reasons people experience hate speech online, Canadians answered: race/ethnicity, religion, sexual orientation and gender identity. These responses were generally shared across surveyed demographics with some distinctions[2].

Disaggregated trends suggest that racialized people feel less comfortable with freedom of expression online, and feel less safe sharing their opinions. We might infer that groups that are more likely to experience hate speech are less likely to support unfettered freedom of speech on digital platforms.

Racialized respondents also reported witnessing a greater frequency of hate speech online when compared to non-racialized respondents. This may suggest an opportunity for platform owners to address algorithmic bias to foster a less hostile experience for racialized users.

Some marginalized and vulnerable groups also tend to be opposed to any significant increase in content moderation. At first this seems counterintuitive, but it may be related to a disbelief in the effectiveness of content moderation resulting from the lived experience of algorithmic bias on digital platforms, like the over-censorship of 2SLGBTQI+ content or Black Lives Matter content. This could present a significant

[1] From December 2022 through February 2023, the Montreal Institute for Genocide and Human Rights Studies (MIGS) paired with RIWI, a global trend-tracking and prediction technology firm, to conduct a public perception survey of Canadian thoughts on online hate and content moderation. RIWI employs Random Domain Intercept Technology (RDIT), which is a survey technology patented by RIWI, designed to minimize bias among random samples of demographically dispersed online respondents. Data is anonymized and for the purpose of this policy paper, insights have been provided based on both its aggregate and disaggregated forms. The survey reached 27,294 unique respondents across Canada, including 4,002 complete responses. Expanded findings of the survey with graphs can be found here.

[2] The survey was run across Canada in both English and French. 26% of respondents were between the ages of 16 and 24, 41% were between the ages of 25 and 44 and 33% were over the age of 45. 60% of respondents identified as male, 30% identified as female and 10% identified as "Other". The top three reported ethnicities among the respondents were white (50%), "Other" (13%) and East Asian (8%). Indigenous respondents made up 4% of the sample group. The top three reported religious faiths amongst respondents were "No Religion" (38%), Christianity (25%) and "Other" (14%).

challenge to content moderators seeking to protect more vulnerable groups of users while avoiding the false-positives or over-moderation of content.

It was clear from the survey that Canadians' prefer humans to be the primary moderators of hate speech, instead of AI tools. For oversight, Canadians indicated that government, social media companies and additional regulators should oversee online harms moderation through a more collaborative model. However, it was fairly evident that Canadians did not want the government to take on the primary role of moderation. The lower trust in government moderation compared to social media self-moderation could suggest a fear of over-regulation or a general lack of trust in government institutions post-pandemic among respondents. Cooperation through a co-design model that involves groups most exposed to online harms could ensure community informed decision making.

## a. Social Media Monitoring of Canadian Politicians and Journalists

Trends in online hate can be tracked on social media, and are often most blatant when used against people in the limelight that might visibly identify as a member of a marginalized community.

During a five month period spanning from January 27, 2023, to June 30, 2023, Areto Labs conducted a comprehensive analysis of the occurrences of hateful or abusive speech within the realm of the X (formerly Twitter) platform.

The social media bot tracked 369,758 comments in English and French on X (formerly Twitter), and monitored for instances of hate against racialized individuals, members of the 2SLGBTQI+ community, Indigenous, Métis, and Inuit populations, women, as well as individuals affiliated with different religious identities.[3]

Of particular concern was the elevated prevalence of such derogatory discourse when directed towards women and trans individuals.

Among the 43 accounts subjected to the highest influx of hate-fueled discourse throughout the measurement period, a pattern emerged: these recipients were exclusively women. Notably absent from this subset were cisgender men, indicating a gender-specific dimension to the distribution of abusive comments. This finding underscored the significance of recognizing the gendered nature of hate speech within the context of online platforms, urging the need for targeted interventions to counteract such patterns.

[3] Areto Labs uses 9 types of abuse in their content flagging. These include: Insult, Homophobia, Transphobia, Ableism, Physical Threat, Racism, Sexism, Spam, and Gender Microaggression.

## Identifying the Challenges to Addressing Online Hate Speech and Providing Policy Focused Solutions

Several key themes emerged from our research with experts and those with lived experience dealing with online hate: Scale of online attacks, Identity, the impact and nature of the harms, personal nature of the attacks, impact of democracies and conspiracy theories.

### a. Scale of Online Attacks

Over the past decade, the boundaries between physical spaces and online platforms have become increasingly blurred as more people spend time online, leading to a rise in digital-first experiences. However, this shift has also brought about online harms that have real-world consequences, including the genocide in Myanmar, violence against women journalists globally,and extremist attacks in New Zealand and the US Capitol. . Marginalized groups are particularly at risk to these types of harms and experience hate speech or online harms at a disproportionate frequency and scale.

Social media gives unprecedented access to others and an unfortunate side effect of that is the spread of dehumanizing speech and incitement to violence. The algorithms on platforms have the ability to amplify hateful narratives on an astronomical level that is so different from the landscape we were in several decades ago. This is in stark contrast to the initial desire many people had for using social media, which was to connect with others around the world. However, the algorithm pushes people to not only find increasingly radical content, but also links them with others that are ingesting it, creating separate groups of people with more extreme views. The business model that the algorithms support erode the potential for connection and coming together that social media provides and instead build more silos. One speaker reflected that 30 years ago people would be passing out anti semitic or islamophobic pamphlets on street corners with minimal reach. But now, one social media post can be seen by thousands or more all over the world.

One study by the Center for Countering Digital Hate found that 700 hateful social media posts were viewed 7.3 million times. Additionally, 84% of reported antisemitic social media posts on Facebook, Twitter, TikTok, Instagram and Youtube did not generate responses from the platform.

### b. Identity

Identity factors play a significant role when considering hate speech and online harms, including religion, ethnicity, nationality, race, sexuality, and gender. Harmful actors develop narratives that target individuals or groups based on these identity factors, thereby intertwining hate speech with identity politics. Severe forms of hate speech targeting minoritized groups have the potential to escalate into violence, atrocities and even genocide. Minoritized communities are targeted with persistent narratives

of "us versus them" that are often aimed at exploiting or manufacturing grievances to dehumanize specific groups.

The Center for the Prevention of Radicalization Leading to Violence published a research report in 2021 measuring the extent of hate motivated behaviour in Quebec. One of their findings indicated that being a part of a religious minority increases your risk of being subjected to hate motivated acts by over two times. Intersectional identity is also a key factor. Muslim women, for instance, are often the most visible and therefore the most targeted.

In the context of Indigenous Canadians, the impact of digitally enabled social platforms and community hubs has been nuanced. While the internet has allowed for new ways to share Indigenous stories in normative society and facilitated the reclamation of spaces through social channels online, it has also led to a concerning influx of digitally enabled hate speech and online harms specifically targeting Indigenous individuals. This issue became so significant that CBC had to temporarily close comments on its Indigenous stories due to a disproportionate amount of harmful user-generated content. As a result, Indigenous peoples face cultural censorship, hindering participation and representation of Indigenous peoples in these spaces.

It was also interesting to note that in Canada, over a ten year period 21% of those accused of hate crimes were between the ages of 12 and 17 and 87% of that cohort were boys. This is a phenomenon that impacts across our society, and digital literacy and the ability to educate youth about the harms of these hateful narratives is so important.

Overall, addressing the root causes of identity-based online hate requires understanding the historical and systemic factors that underpin it, as well as a willingness to take collective responsibility while creating more inclusive and equitable spaces online through collaborative and relational regulatory approaches.

## c. Role of Algorithms

Currently, the biggest platforms operate by incentivizing engagement, or in other words, time spent on the platform. Built into the business model of the main social media platforms are algorithms that actively push radical content. Indeed, studies have shown that negative messages, messages of hate and disinformation spread far quicker because this type of content keeps users longer on the platform. This drives users to stay on the platform longer, causing the disproportionate spread of more radical and harmful content over fact-based and more measured content. Consequently, the publication of harmful online content, including attacks against women leaders in politics and journalism, marginalized groups and others, becomes a source of revenue for social media companies.

The algorithmic bias common to social platforms contributes to online harms faced by racialized and otherwise marginalized groups. By this we are referring to systematic errors within platforms that

arbitrarily favor one group over another. For instance, queer and racialized Instagram users are often arbitrarily flagged as inappropriate and banned, and YouTube channels hosting certain religious or racialized content can be demonetized. These biases create a greater cognitive load for these users, which leads to highly draining experiences of vulnerability online. Members of marginalized groups must constantly navigate questions like "how do we interact with our communities?" and "how are we allowed to express ourselves?"

Regulations therefore have to go beyond user level moderation and terms-of-service reforms and instead target the systemic roots of engagement-bias and the business model and development of the social media platforms and search engines.

## d. Impact and Nature of the Harms

The boundaries between physical spaces and online platforms are blurred as Canadians are spending more time online with an increasing frequency of digital-first experiences. As a consequence digital harms are also real-world harms, and there are many ways this can manifest:

- Incitement to violence and dehumanizing speech;
- Bullying, insults, mockery, propagation of rumours;
- Hacking;
- Outing, which is the deliberate or accidental sharing of another person's sexual orientation or gender identity without their explicit consent;
- Doxxing, which is the intentional revelation of a person's private information online without their consent, often with malicious intent. This includes the sharing of phone numbers, home addresses, identification numbers, etc.;
- Sextortion, which is the the act of threatening to share nude or explicit images;
- Algorithmic bias.

## e. Personal Nature of Online Hate Speech

Building off the impact and nature of online harms, the personal nature of attacks came up in the panel on online hate against women in particular. Several women in the panel indicated that they expected to be attacked based on their opinions or work, but did not expect to be attacked for their appearance. More so than in other panels, the speakers identified that when there is incitement of violence against women online, it often extends to their families and loved ones. These instances of hatred targeting children, spouses and mothers can act as a significant barrier to prevent many women from participating in leadership or public facing roles. This is really amplified for women from racialized communities who may experience hate from multiple different angles.

## f. Impact on Democracies

Hate speech poses an enormous threat to modern democracy due to its divisive nature and potential for inciting violence. Social media has added a new dimension to the threat of hate speech due to its vast reach and the speed in which content and information can travel. This has become a particularly imposing issue during national elections as foreign state actors and well funded interest groups are capable of manipulating public narratives at an incredible pace, however, the capacity for civic harm is not limited to elections.

Across all panels, the notion of the silencing or “chilling” effect of online hate and incitement to violence came up. The chilling and silencing effect is even more severe for those who may identify as more than one of the targeted groups. For example women who are also part of a religious, ethnic or racial minority may feel even more pressured to "stay silent" on online discussions and political discourse. This silencing effect is detrimental to democratic societies and presents the appearance that women and girls are absent from digital commons, and are therefore not a priority in their planning or design.

Jon Penney outlines the chilling effect as “an act of compliance with or conforming to social norms in a given context. Chilling effects arrive out of contexts of ambiguity -—such as ambiguity in the law or a circumstance where a person is aware they may be monitored by the government. If a person wishes to say or do some particular thing, but face ambiguity as to whether their conduct is legal or may attract scrutiny if they are being monitored, they face uncertainty about how to act.And in such moments of uncertainty, behavioral social science tells that people tend to act the way they believe others would act in the same circumstance, that is, they follow the norm.” This can also happen if someone feels like they are being surveilled and monitored by other users on social media which has resulted in some form of hate speech, harassment, disinformation or other negative outcome.

The Samara Center for Democracy’s SAMbot tracked toxic tweets received by incumbent candidates and party leaders during the 2021 Federal election. It found that “Female candidates were more likely to receive toxicity than male candidates. Approximately 21% of all tweets directed at female candidates were likely toxic while 18% of tweets directed at male candidates were likely toxic.”

#ShePersisted, an organization working to end gendered disinformation and abuse against women in public life, explains the problem succinctly “These types of attacks do not only represent a threat to the women they target. Weaponized by malign foreign and domestic actors, these attacks threaten democratic institutions and have important ramifications for global peace and security and the broader human rights system.” A weak democracy is susceptible to foreign subterfuge where conspiracy theories and harmful tropes can gain stronger footing in the absence of diverse perspectives and lead to the manipulation of political discourse.

## g. Anonymity

Anonymity on the internet bolsters people's confidence to say things they would not say offline. However, anonymity can also allow women and others who feel targeted by online hate to participate in

these spaces without fearing for their safety. More thought is needed on how to balance the benefits of the drawbacks of the capacity to remain anonymous online.

## h. Conspiracy Theories

One unique theme that came out of the panel on religious minorities is the impact of conspiracy theories. One panelist noted the persistence of antisemitic conspiracy theories related to the myth of a global Jewish conspiracy to control the media, the economy, government or other institutions, including the "great replacement theory" which is also anti-Muslim, amongst others. QANON also spread a host of antisemitic theories that gained a lot of traction during the COVID-19 pandemic. Radicalized groups are able to promote some ideologies present in conspiracy theories that can turn rhetorical hatred into active violence. When these conspiracy theories make their way into mainstream political discourse it results in the rollback of rights for multiple communities, which we have seen results in an increase in hate incidents.

## Solutions:

### Regulation of big tech

Content moderators are needed in local languages (e.g. of Myanmar). Additionally, many of the harms caused by gender disinformation is a result of these platforms existing without any sort of oversight or regulatory framework. Currently, the biggest platforms operate by incentivizing engagement, or in other words, time spent on the platform. It doesn't matter if the engagement is positive, negative, harmful or abusive. As long as users are logged-in as much as possible. So, regulation has to go beyond user level moderation and terms-of-service reforms and instead target the systemic roots of engagement-bias.

Canada is the only G7 country without comprehensive intermediary liability laws in place. The federal government should pass legislation that would clarify when platforms should or should not be held liable for harms caused by content on their platforms. Those liability protections should then be incorporated in the Canada-US-Mexico Agreement (CUSMA) as well as other trade agreements. Additionally, laws need to have some "teeth" and impose a penalty for the platform's failure to uphold its obligations. Some helpful resources include UNESCO's Guidelines for Regulating Digital Platforms and Europe's Digital Services Act, which also maps out a co-regulatory model.

The Santa Clara Principles On Transparency and Accountability in Content Moderation are another resource on regulating internet platforms. The first version of the Principles outlined minimum standards that internet platforms must meet to provide adequate transparency and accountability about their efforts to moderate user-generated content or accounts that violate their rules and was endorsed by Apple, Facebook, Google, and Twitter. The second version adds "standards directed at government and state actors to beef up due process and expand guidelines for reporting on and notifying users about takedowns." Tech companies should consider endorsing this second version of the Principles.

### Accessible monitoring and reporting:

There is a need for reporting and monitoring mechanisms that are accessible, useful and reliable.

### Getting back to the basics of data protection and privacy

Focusing on data protection and privacy is a tangible way to take important steps to combat online harms like hate speech. Data feeds into biased algorithms and perpetuates harm.

### Reassess our focus on innovation at all costs

A human rights lens needs to be applied when it comes to the adoption of fast evolving technology. Our focus on speedy innovation as an essential component to our socioeconomic future gives tech companies an enormous amount of lobbying power. We need to understand that regulation does not have to be seen as undermining innovation and is not contrary to our economic interests.

### Building strong partnerships

There also needs to be more work towards developing strong partnerships between government, for profit companies and civil society actors. Government legislation is critical in defining the benchmark for cyber hatred and online harms while giving platform owners and civil society actors tools to respond. However, more work is needed to establish mechanisms for civil society and private actors to participate in lateral development of best practices that benefit everyone. Additionally, allyship of affected groups to counter hate is important. A particular challenge with monitoring social media platforms is that algorithms are not always capable of differentiating between hate speech and speech that is meant to counter hate speech. When it comes to content moderation, we need to engage social media companies to make sure that they adopt a nuanced approach that is developed with the help of targeted communities.

### Building and utilizing an international network

The UN focal point on hate speech could play a central role in developing and maintaining a non-hierarchical international network of different actors involved in this work. Given that there is not a clear consensus internationally on how to deal with online hate, tech companies are trying to respond to multiple different takes by governments. There are many logistical hurdles, such as the evolution of technology, anti-regulation lobbies becoming more powerful, and bad actors becoming more sophisticated. Creating such a network between national and civil society actors is a critical step to countering hate speech and online harms.

### Empowering users

Equip users with tools to respond to cyber hatred (i.e. counterspeech, knowing how to use tools like blocking and muting and understanding that there is an algorithm that profiles and feeds you certain types of content). Citizens assemblies are also powerful tools to hear from people on their perspectives on the digital rights space.

### Role of the media

Mainstream media filters down mis- and dis- information found online. The media is also partially responsible for public education, and needs to be responsibly reporting on freedom of speech topics and not playing an obfuscating role. So, there needs to be consideration towards profit and incentification for those organizations, and support in place to protect women media professionals and their families from both offline and online expressions of hatred and violence.

### Education - Beyond Digital Literacy:

Digital literacy skills are extremely important. One of the common reasons behind increased visibility of harmful content online is that platform users have the perception that they can get away with it. Users need to know both how to use the internet and how to be on it (know how to identify mis- and dis-information and how/if to engage with hateful rhetoric). They also need to understand that the internet is an individualized experience. Education is also required to understand and address the offline root causes of online hate and what it means to live in a diverse society. It is also important to equip these users with tools to respond to cyber hatred and equip users with the tools to act in situations where

they are experiencing racism or other forms of discrimination on digital platforms. Additionally, we need to ensure elected officials have the knowledge and vocabulary to understand the industry to ensure proper regulations and monitoring mechanisms can function. Civil society organizations can be great partners for educational needs, however, they also need funding.

### Beyond technology:

We need to work together to tackle the root causes of intolerance and hatred. The faithful rights framework within the UN system was specifically developed to look at how peer-to-peer learning and engagement can try and bridge understanding and collaboration across different religious or belief communities. We need to engage diverse stakeholders, including parliamentarians, religious or belief leaders, educators and social media influencers to try and move towards having systemic and societal change that effectively tackles online hate. At the local level, there should be accessible mechanisms for reporting and monitoring of online hate against villages or belief communities.

## Conclusion

Online hate speech is a key issue in Canada that will not disappear without strong legislation and enforcement mechanisms. Key challenges when considering policy recommendations include: the sheer scale of content online and the wide reaching effects of online attacks, the role of identity as a influencing factor for both perpetrator and victim groups, the impact of algorithms on the content we see and the propagation of hate speech, the impact and often personal nature of online hate speech, the role of anonymity as both a tool to keep yourself safe but also bolster users that feel empowered to spread hate speech, conspiracy theories and the impacts on democracies globally.

Solutions include strong regulation of big tech with usable enforcement mechanisms, accessible monitoring and reporting mechanisms, a renewed focus on data protections and privacy, reassessing our focus on innovation at the expense of human rights, empowering users and the media to combat the spread of hate speech, building strong partnerships across sectors, borders, and between targeted groups and education tools that go beyond digital literacy and help users understand how platforms work and what their rights are.