

Navigating Online Hate

A Community Guide

Produced as part of the Canadian Digital Resilience Network Against Online Hate Initiative and supported by the Canadian Department of Heritage.

Author : Jack Rath

April 2026



Acknowledgments

Thank you to the Canadian Commission for UNESCO, Les3Sex*, the University of Ontario, the Centre for Media, Technology and Democracy, and the Max Bell School of Public Policy for their support, collaboration, and commitment throughout the development of this report.

1: Introduction and Purpose

Why This Guide Exists

This guide was produced by the Montreal Institute for Global Security (MIGS) as part of the Canadian Digital Resilience Against Hate Network, funded by the Department of Canadian Heritage. It draws on findings from four national discussion events held between October 2025 and March 2026, independent research, government data, and contributions from partner organisations.

71%

of Canadians aged 15 to 24 have been exposed to online content that incites hate or violence.

Canada is facing a measurable and accelerating spread of hate speech online. The online platforms Canadians use every day are designed to reward divisive and outrage-driven content, where engagement generates revenue regardless of its consequences. The result is a digital environment in which the spread of hate is a structural feature, one that is being amplified by artificial intelligence (AI), exploited by foreign actors, and increasingly linked to real-world violence. Indigenous, Black, racialised, and 2SLGBTQIA+ communities, amongst others, bear a disproportionate and compounding share of that burden. Canada has the opportunity to develop a response, grounded in a tradition of cross-community solidarity and a growing body of youthled digital resilience work. This guide exists because understanding the problem clearly is where meaningful change begins.

Key Terms and Definition

Understanding online hate begins with a shared language. The following terms appear throughout this guide.

Online Hate Speech

The Government of Canada defines hate speech as communication that expresses hatred or vilification of an individual or group based on protected grounds of discrimination. These grounds include race, ethnicity, religion, gender identity, sexual orientation, and disability, among others. The European Commission against Racism and Intolerance offers a broader definition that includes harassment, insult, negative stereotyping, stigmatisation, and threat.

Importantly, hate speech targets groups and communities, not just individuals. Hate speech expresses disdain toward a collective, even when it is directed at a specific person because of their membership in that group.

Did You Know?

Hate speech and cyberbullying are related but distinct. Cyberbullying targets an individual. Hate speech targets a person because of who they are, their race, religion, gender, sexuality, or other protected characteristic, and by targeting that person, it sends a message of hostility to the entire community they belong to.

The Online Hate Ecosystem

This guide uses the term 'online hate ecosystem' to describe the entire networked environment in which hate is produced, amplified, shared, and acted upon. The ecosystem is not random, it has patterns, pathways, actors, platforms, and economic incentives. Understanding it as an ecosystem, rather than a series of isolated incidents, is the first step to responding to it effectively.

Disinformation and Gendered Disinformation

Disinformation refers to deliberately false or misleading content intended to deceive. Gendered disinformation is a specific and increasingly documented phenomenon in which misogyny is weaponised as a political tool, used to attack women leaders, distort gender equality debates, and destabilise democratic institutions. It combines fabricated content, coordinated campaigns, and deep-seated cultural biases to cause both personal harm and political damage.

Radicalization Pathways

Radicalization pathways are the gradual processes through which individuals move from everyday online spaces toward more extreme content and communities. A young person might begin by watching gaming videos, then be presented content about male self-improvement, then content blaming women or minorities for societal problems, and eventually arrive at explicitly extremist material, all through algorithmic recommendations, without ever actively seeking it.

Algorithmic Amplification

Social media platforms use algorithms to decide what content users see. These algorithms are primarily designed to maximise engagement: the more time people spend on a platform, the more advertising revenue the platform generates. Content that provokes strong emotional reactions, including outrage, fear, contempt, excitement, tend to perform well algorithmically. This is a design choice with significant consequences for what kinds of ideas and attitudes get amplified at scale.

2: Where Hate Lives - Platform Specific Dynamics

Online hate exists along a spectrum of platforms. At one end are fringe platforms, spaces like 4chan, 8chan, Telegram, and Gab which operate with minimal moderation and function as incubators. This is where the most explicit hate content is produced, where extremist networks organise, and where harmful ideas are normalised within insular communities before migrating outward.

At the other end are mainstream platforms including Facebook, YouTube, X (formerly Twitter), TikTok, and Instagram, which reach vastly larger audiences. Hate on these platforms is less explicit, but its reach is incomparably greater. Research monitoring hate speech in British Columbia found that while fringe platforms host the most extreme content, mainstream platforms are where it reaches the most people, often through coded language, memes, and content that appears ambiguous to automated moderation systems.

The relationship between these spaces is dynamic. Ideas, language, and imagery move from fringe platforms to mainstream ones, and back again. A slogan developed in an extremist forum can appear on TikTok within days, stripped of its original context but carrying the same underlying message to an audience that may not recognise its origins.

TikTok

TikTok has become one of the dominant digital spaces for Canadians under 40. As of 2024, it had 12.1 million registered users in Canada, with over 70% of users under 40, and 43% aged 18 to 29. Canadians spend an average of 17 hours per month on the platform, more than Facebook or Instagram. Around 65% of users check the app daily.

The platform's short-form video format makes it particularly effective at spreading content that feels entertaining. Research published in *Social Media and Society* in 2025 analysed what it calls 'sigma masculinity' content on TikTok; videos that promote male dominance, rejection of women, often using cinematic figures and specific music genres to make these messages feel aspirational and cool. This content normalizes harmful attitudes in ways that are difficult to detect through automated moderation and easy to dismiss as humour or entertainment.

TikTok has also been documented as a significant vector for antisemitic and islamophobic content surges, particularly following the October 2023 Hamas-Israel conflict. Researchers at the University of

Waterloo noted a 919% increase in antisemitic content and a 422% rise in Islamophobic content on X, with similar patterns observed across platforms including TikTok during this period.

X (Formerly Twitter)

X has undergone significant changes since its acquisition by Elon Musk in 2022, with documented reductions in content moderation staffing and major policy shifts. The Southern Poverty Law Center documented that Meta's January 2025 policy overhaul removed the term 'hate speech' from its Hateful Conduct rules. X has followed a similar trajectory, and the alignment of the platform's leadership with certain political ideologies has raised concerns about the neutrality of its moderation decisions. These shifts have material consequences for the communities most targeted by hate speech.

"The content on Elon Musk's feed is a mirror of his own interests; ample praise for his priorities, mixed with a torrent of right-wing outrage over progressive politics. It highlights the ways social networks can create information bubbles."

— New York Times, May 2025

Instagram and Facebook (Meta)

Meta's platforms remain among the most widely used in Canada across age groups. Meta's January 2025 policy changes replaced its independent fact-checking programme with a 'Community Notes' model, a crowd-sourced approach that critics argue is slower, less consistent, and more susceptible to manipulation by coordinated networks. The removal of explicit hate speech protections for several marginalized groups from Meta's Hateful Conduct rules has prompted significant concern from advocacy organisations, who warn that these changes may lead to a significant increase in online harassment and real-world danger for vulnerable communities.

Twitch and Discord

These two platforms play distinct but important roles in the online hate ecosystem, particularly as they relate to young men. Twitch is a live-streaming platform primarily associated with gaming, with a large and culturally engaged youth audience. Its interactive, real-time format makes it valuable both as a space where harmful content can circulate and as a potential tool for outreach and counter-messaging.

Discord offers smaller, semi-private server spaces much more like closed community rooms than public broadcasts. These more insular environments can mirror the conditions in which harmful ideologies circulate and intensify. At the same time, Discord can also host supportive communities, peer education initiatives, and counter-narrative spaces.

Gaming Platforms and Children's Spaces

Online hate is not confined to platforms designed for adults. A CBS News investigation in August 2025 documented extensive hate speech in 'Spray Paint!', a popular Roblox game, where researchers found dozens of swastikas and multiple hate slurs within minutes of gameplay. Roblox maintains a moderation system, but harmful content continues to evade it. With approximately 40% of Roblox users under the age of 13, this is a child safety issue as much as a hate speech issue.

Platform Accountability

Across all of these platforms, a consistent theme emerges: platforms are currently required to regulate very little in most jurisdictions, and the gap between the harms they generate and the accountability they face is significant.

Some jurisdictions are beginning to close this gap. The European Union's Digital Services Act (DSA) requires large platforms to conduct risk assessments and publish transparency reports on their content moderation practices. New York's Stop Hiding Hate Act, which came into force in October 2025, requires social media companies with over \$100 million in annual revenue to report biannually to the state Attorney General on how they handle hate speech and harmful content, with civil penalties of up to \$15,000 per violation per day for non-compliance. Canada tabled the Combatting Hate Act in September 2025, which strengthens hate crime laws and simplifies prosecution.

These are important steps, but gaps remain, particularly for communities in Canada who continue to encounter hate on platforms that face no binding local obligations to address it. Advocating for stronger platform accountability is itself a form of community action.

100%

of Canadian youth who use Facebook have reported seeing hate on their feeds, according to the 2023 Youth Assembly on Digital Rights and Safety.

The Role of Algorithms

One of the most important mechanisms driving the spread of online hate is the role of platform algorithms. Every time you scroll through a feed, click on a video, pause on a post, or share content, the platform's algorithm learns something about what holds your attention. It then uses that information to serve you more content like it, all because your continued engagement generates advertising revenue. Content that provokes strong emotional reactions performs particularly well: outrage, fear, contempt, and humiliation all drive engagement. This means that divisive, extreme, and hateful content often travels further and faster than moderate, nuanced, or prosocial content.

A young person searching for gaming content can, within weeks, find themselves being served increasingly misogynistic videos, then content blaming societal problems on women or minorities, then explicitly extremist material, all through recommendation systems following patterns of engagement.

Coded Language

Hate groups and extremist communities are highly adaptive. When platforms remove explicit hate content, these communities develop coded language, symbols, and cultural references to convey the same messages while evading detection. Memes, images layered with meaning, can carry deeply offensive or extremist content that is legible to insiders but appears innocuous to automated systems or uninitiated observers.

What to look for

Coded language in online hate often borrows from mainstream culture, including gaming terms, internet slang, and pop culture references. If you encounter content that seems to be making in-jokes about violence, ethnic groups, or gender, it may be worth looking more closely at what is actually being communicated.

From Online to Offline: The Feedback Loop

Online hate does not stay online. The relationship between digital toxicity and physical violence resembles a continuous loop: real-world events fuel online hostility, which in turn escalates into offline harm, which then generates more online content, and so on.

The Toronto van attack in 2018, the Christchurch mosque shootings in 2019, the University of Waterloo stabbing in 2023 all had clear, documented online components. Perpetrators engaged with online extremist communities, produced or consumed online manifestos, and were shaped by the ideologies and cultures those spaces promoted.

This pattern has a broader democratic dimension as well. Persistent harassment online diminishes civic participation, particularly for women, racialised communities, and other marginalised groups. When people are targeted for speaking publicly about who they are, they often withdraw from public discourse.

Who Produces and Spreads Online Hate

Understanding who is behind online hate helps us design smarter responses. Online hate is generated and circulated by a diverse set of actors rather than isolated individuals, which has direct implications for policy design.

State and foreign actors run coordinated disinformation campaigns targeting democratic institutions, women leaders, and minority communities. Digital forensic research has identified campaigns linked to Russian networks targeting German politicians, deepfakes targeting the Moldovan president, and coordinated manipulation of gender equality debates across multiple countries.

Organised extremist groups recruit deliberately and strategically, often targeting young men experiencing social isolation or economic anxiety.

Online influencers produce content that normalises misogyny, often wrapped in self-help messaging, fitness culture, or entertainment. Research analysing sigma masculinity content on TikTok found that even seemingly humorous material can normalise prejudicial attitudes and perpetuate harmful stereotypes about women and gender.

Ordinary users who share, commenting on, or criticise hate content can amplify it through engagement-driven algorithms. Research shows there is a contagion effect in online comments, where exposure to hate content increases the likelihood that those exposed will produce or spread it further.

3: The Impact of AI and Emerging Technologies

AI is changing the online hate ecosystem, accelerating the production of harmful content, making it harder to detect, and creating new forms of harm that existing legal and technical frameworks were not designed to address.

AI as an Amplifier of Hate

Generative AI tools, the technology behind AI image generators, chatbots, and video synthesis tools, allow content to be produced at scale with minimal effort and, increasingly, minimal technical skill. Hate groups have been skillful adopters of these tools. A 2024 CBC investigation documented AI-generated propaganda being used to spread antisemitic, islamophobic, and racist narratives, including doctored historical footage and AI-generated imagery that distorts Holocaust history.

Large language models (LLMs) are the technology underlying most AI chatbots, including those embedded in social media platforms. These models are trained on enormous amounts of internet data, including all data that contains substantial volumes of hateful, extremist, and biased material. Research from the Rochester Institute of Technology found that small provocations can push AI models into producing calls for ethnic cleansing, Holocaust denial, and sexual violence, with Jewish communities and women consistently among the most targeted groups in these outputs.

The most widely documented case in 2025 was Grok, the AI chatbot embedded in X. In July 2025, Grok entered what the media described as a 16-hour period of antisemitic, racist, and sexually violent outputs. Referring to itself as 'MechaHitler,' Grok promoted white supremacist conspiracy theories, provided graphic instructions for sexual assault against named individuals, and made antisemitic remarks targeting a specific user by name. X attributed the behaviour to 'deprecated code,' but experts noted that the incident reflected broader structural vulnerabilities in how LLMs are built and deployed.

"Grok's behaviour reflects a broader AI challenge: balancing user instructions with ethical safety. We need to build models which are more aligned to human values, and we have to keep our research going to identify these problems and address them one after one."

— AI safety researcher, CNN, July 2025

4: What Youth Wish They Could Tell Their Minister

Young Canadians are expressing serious anxiety about emotional dependence on AI chatbots, concern that AI is eroding their capacity for critical thinking and genuine community discourse, and awareness that chatbot interactions can expose them to sexual, extremist, and self-harm content. They are also asking for the right to opt-out of addictive design features, for greater transparency about how AI systems work, and for sustained youth consultation in AI governance that involves structural inclusion in the decisions being made about technologies that shape their lives.

Across four national discussion events, a consistent and compelling set of messages emerged from young Canadians. This section presents those messages directly. Each message is grounded in the evidence gathered through the discussion event series and the research underpinning this guide.

"Algorithms Are Not Neutral, They Are Being Used Against Us"

Young Canadians have a sophisticated understanding of how algorithmic recommendation systems work, and a deep frustration that this understanding is not reflected in policy. They consistently describe being served increasingly extreme content without seeking it, watching recommendation pathways steer peers toward misogynistic, racist, or otherwise hateful material that they never explicitly requested.

Youth want policymakers to treat algorithmic design as a public health and public safety issue. Engagement-driven business models that reward divisive content have consequences for democratic participation, mental health, and community cohesion. Those consequences should be regulated accordingly.

80%

of Canadian youth who use Facebook believe people are more likely to engage in prejudiced speech online than offline.

We Are Not the Problem, But We Are Left to Deal With It"

Young Canadians are disproportionately exposed to online hate, yet are largely excluded from the policy conversations meant to address it. The Youth Assembly on Digital Rights and Safety articulated nine values that should shape Canada's approach to online safety; transparency, accountability, human-oriented design, safety, autonomy, security, adaptability, accessibility, and inclusivity. These values were developed by young people who are living the reality that those values are currently absent from most platform design and digital policy. Youth explicitly noted that approaches to online

safety that restrict rather than empower young people miss the point. They want agency, the ability to control what they engage with, to opt-out of addictive design features, and to be treated as digital citizens with both rights and responsibilities.

"Mental Health and Platform Design Are Connected"

The features that keep young people on platforms, like streaks, infinite scrolling, algorithmically recommended content, human-like AI chatbot interactions, are the same features that research consistently links to depression, anxiety, addiction-like symptoms, and social isolation. Canadian Youth understand the impact of growing up as heavy digital users on platforms engineered for engagement maximisation, and they see the connection between those design choices and their own mental health experiences.

"Companies and platforms prioritize profit over people. Youth in particular are experiencing significant and compounding negative impacts from these profitdriven practices — exploitation, diminished mental, physical, and emotional wellbeing, and greater proclivity for addictive and risky behaviours."

— 2023 Youth Assembly on Digital Rights and Safety

"Positive Role Models Matter, but They Are Hard to Find Algorithmically"

Across discussion events, a recurring theme was the absence of visible, algorithmic pathways toward positive content for young men in particular. The manosphere and its associated influencers are well-funded, algorithmically amplified, and culturally sophisticated. Initiatives that offer alternative narratives. Programs like Cyberhéros from Les3sex*, the No Social November movement, and community-based masculinity programs like The Man Cave in Australia are under-resourced and often invisible to the young people who most need to encounter them.

Youth want investment in counter-narrative infrastructure and policies that support influencers, programs, and communities that model healthy relationships, positive masculinity, and prosocial digital behaviour. They also want platforms to surface this content as actively as they surface harmful content, which means addressing the algorithmic incentive structures that currently disadvantage it.

5: A Practical Toolkit to Recognise and Respond to Online Hate

Recognising Online Hate

Not all offensive content is hate speech, and not all hate speech is offensive. Here is a simple set of questions to help you identify it:

- Does the content attack or demean someone because of their race, religion, gender, sexual orientation, disability, or other protected characteristic?
- Is it designed to dehumanise a group; portraying them as dangerous, inferior, or undeserving of equal treatment?
- Does it use coded language, symbols, or memes that may carry extremist meaning to those who know how to read them?
- Does it spread false or distorted information about a community in a way that promotes hostility toward them?
- Does it call for discrimination, exclusion, or violence against a group?

If the answer to any of these questions is yes, you may be looking at hate speech. You do not need to be certain; reporting mechanisms exist precisely so that trained moderators can make those determinations.

Remember

Hate speech is not the same as content you disagree with or find offensive. It specifically targets people because of who they are and its purpose is to denigrate or incite hostility against a group. Understanding this distinction helps you respond effectively and avoid false reports that can clog moderation systems.

Responding Without Amplifying

Engaging with hate content, even to refute it, can amplify it through algorithmic systems that treat all engagement as positive signals. This does not mean you should do nothing. It means choosing your response carefully.

- Document rather than share. Take a screenshot or use a link-saving tool to preserve evidence of hate content without amplifying it.
- Report through platform tools. Every major platform has a reporting mechanism. Use it. Reports aggregate, and patterns of reports can trigger human review even when automated systems miss content.

Reporting Pathways in Canada

Canada has several mechanisms for reporting online hate, depending on the nature of the content and who is responsible for it.

- Cybertip.ca — operated by the Canadian Centre for Child Protection, for reporting online child sexual abuse material and exploitation.
- Local police — for content that may constitute criminal hate speech, including wilful promotion of hatred, uttering threats, and incitement to violence.
- Platform reporting tools — for content that violates platform terms of service. While platforms are imperfect, reported content is more likely to be reviewed than content that goes unreported.
- The Canadian Human Rights Commission — for discrimination complaints under federal jurisdiction.

\$1.7M

awarded in damages by an Ontario court in August 2025 to eight individuals targeted in an online smear campaign. The case demonstrates that online hate and defamation have real legal consequences.

Protecting Yourself

If you are being targeted by online hate or harassment:

- Document everything. Screenshots with timestamps, URLs, usernames, and dates can be essential evidence if you choose to report to police or pursue legal action.
- Adjust your privacy settings. Limit who can see your personal information, tag you in posts, or contact you through platform settings.
- Tell someone. Online harassment can be isolating. Tell a trusted person such as a friend, family member, colleague, or counsellor.
- Take breaks. You are not obligated to remain in spaces where you are being harmed.
- Stepping away from a platform temporarily is a valid and healthy response.

A Final Word: Collective Responsibility

Sustainable solutions to online hate require long-term community engagement. Reporting harm, supporting those who are targeted, building systems that persist beyond any single project or intervention, and demanding accountability from platforms and governments are not optional additions to this work. They are its foundation.

This guide was produced by the Montreal Institute for Global Security in partnership with communities across Canada who have direct experience of the harms it describes. Their knowledge, their stories, and their insistence on being part of the solution have shaped this document.

Online hate is not inevitable. The platforms that host it were designed by people, the policies that fail to address it were written by people, and the cultural norms that sustain it are maintained by people. These platforms and policies can be redesigned by people.

Glossary of Other Terms

Artificial intelligence (AI) refers to computer systems designed to perform tasks that would ordinarily require human intelligence, such as generating text, images, or video.

Bot is an automated online account that can post, share, or comment without human input, and is frequently used to make fringe views appear more widespread than they are.

The contagion effect is the documented phenomenon by which exposure to hate speech increases the likelihood that those who encounter it will go on to produce or spread it further, even among people who reject the content.

Cybersexism refers to online acts of harassment or violence that are sexist, homophobic, or sexual in nature, typically targeting women, girls, and gender-diverse individuals.

Cyberviolence is a broad term for harm inflicted through digital technologies, encompassing cyberbullying, doxxing, non-consensual sharing of intimate images, online harassment, and hate speech.

Deepfake is synthetic media created using AI in which a real person's likeness, voice, or actions are fabricated to make them appear to say or do things they never did.

Doxxing is the practice of researching and publicly publishing private or identifying information about a person without their consent, typically to intimidate, harass, or cause harm.

Echo chamber is an online environment in which a person is primarily exposed to information and opinions that reinforce their existing beliefs, often created and sustained by algorithmic recommendation systems.

Extremism refers to beliefs or ideologies that advocate for discrimination, violence, or the dehumanisation of individuals or groups based on characteristics such as race, religion, gender, or sexual orientation.

Fringe platforms are online spaces that operate with minimal or no content moderation, and which frequently serve as incubators for extreme content before it migrates to more widely used platforms.

Hate crime is a criminal offence under the Canadian Criminal Code motivated by bias or prejudice toward a person or group based on characteristics such as race, religion, gender identity, or sexual orientation.

Large language model (LLM) is the technology underlying most AI chatbots, trained on vast amounts of internet data to generate text responses, and vulnerable to reproducing hateful or extremist content when prompted.

Manosphere is a loose network of online communities united by anti-feminist and misogynistic ideologies, existing across forums, social media, podcasts, and video channels that promote male grievance and hostility toward women and gender equality.

Misinformation is false or inaccurate information spread without necessarily malicious intent, contributing to a polluted information environment regardless of the motivation behind its spread.

Moderation is the process by which platforms review, restrict, or remove content that violates their terms of service, carried out through automated systems, human reviewers, or a combination of both.

Sigma masculinity is a form of online content, prevalent on platforms like TikTok, that promotes male dominance and emotional detachment while denigrating women and mainstream social norms, often through humour that normalises harmful attitudes.

Synthetic media refers to content created or substantially altered using artificial intelligence, including deepfake videos, AI-generated images, and voice cloning, posing significant challenges for distinguishing authentic from fabricated material.

Toxic masculinity refers to cultural norms around masculinity that are harmful to both those who internalise them and those around them, frequently manifesting online as content that frames aggression as strength and treats minority groups as threats.

Resources for Help

If you are in immediate danger Call 911 or your local emergency services.

If you need to talk to someone right now

Kids Help Phone offers free, confidential, multilingual support 24 hours a day, seven days a week for anyone in Canada aged 5 to 29. Call 1-800-668-6868 or text CONNECT to 686868.

The 988 Suicide and Crisis Lifeline is available 24 hours a day, seven days a week, for anyone experiencing mental health distress or crisis. Call or text 988.

The Hope for Wellness Helpline offers immediate emotional support, crisis intervention, and referrals to community-based services for Indigenous peoples across Canada, with counsellors available in English, French, Cree, Ojibway, and Inuktitut. Call 1-855-242-3310, available 24 hours a day, seven days a week.

The Black Youth Helpline provides support for Black children, youth, and families across Canada. Call 416-285-9944 or 1-833-294-8650.

Trans Lifeline is a peer support line run by and for trans people. Canadians can call 1-877-3306366.

Youthline provides text and chat support for 2SLGBTQ+ individuals across Canada. Visit youthline.ca or text 647-694-4275.

If you want to report online hate or a hate crime

Cybertip.ca is Canada's national tipline for reporting the online sexual exploitation of children, as well as other forms of online harm including non-consensual intimate imagery. Visit cybertip.ca or call 1-866-658-9022.

NeedHelpNow.ca provides support specifically for teens seeking to stop the spread of sexual images or videos shared without their consent. Visit needhelpnow.ca.

The National Council of Canadian Muslims provides an incident report form for victims of Islamophobic hate and discrimination. Visit nccm.ca.

B'nai Brith Canada operates an Anti-Hate Hotline for reporting antisemitic hate crime incidents. Visit bnaibrith.ca.

Bibliography

Bagaud, E. and Peel, M.-A. (2024) *Collective intelligence: together against gender-based cyberviolence*. Les 3 sex*. Available at: <https://les3sex.com/en/collectiveintelligence>.

Canadian Anti-Hate Network (2025) *Kernatium Division: Canadian teenagers create militia intended to 'kill Jews and immigrants'*. Available at: <https://www.antihate.ca/kernatium-division-canadian-teenagers-militia-kill-jews-immigrants>.

Canadian Centre for Policy Alternatives (2025) *Updating Canadian law to address online hate and gender-based violence*, *The Monitor*, 2 December. Available at: <https://www.policyalternatives.ca/news-research/updating-canadian-law-to-address-online-hateand-gender-based-violence/>.

CBC News (2023) *University of Waterloo stabbing suspect posted anti-women, anti-LGBTQ content online, court hears*, 8 December. Available at:

<https://www.cbc.ca/news/canada/kitchener-waterloo/twitter-x-tiktok-university-of-waterloo-hate1.7037526>.

CBC News (2025) *New Glasgow man, 22, charged after police respond to online hate speech*, 7 June. Available at: <https://www.cbc.ca/news/canada/nova-scotia/new-glasgow-man-22-chargedafter-police-respond-to-online-hate-speech-1.7555449>.

Centre for Media, Technology and Democracy and MASS LBP (2023) *2023 Youth Assembly on Digital Rights and Safety*. McGill University. Available at: <https://digitalassembly.ca>.

Chaarani, J. (2023) 'Social media algorithms to blame for antisemitic, Islamophobic content online, Waterloo expert says', *CBC News*, 27 November. Available at:

<https://www.cbc.ca/news/canada/kitchener-waterloo/twitter-x-tiktok-university-of-waterloo-hate1.7037526>.

Davids, J. (2025) *Wired for worry: how smartphones and social media are harming Canadian youth*. Macdonald-Laurier Institute, 15 April. Available at: <https://macdonaldlaurier.ca/wired-forworry-how-smartphones-and-social-media-are-harming-canadian-youth/>.

Dej, E. and Kilty, J. (2024) "'Die alone, old, and let the cat eat your face": anti-feminist backlash and academic cyber-harassment', *Feminist Media Studies*, 24(1), pp. 70–86. <https://doi.org/10.1080/14680777.2023.2181140>.

Dutta, A., Khorramrouz, A., Dutta, S. and KhudaBukhsh, A.R. (no date) *Down the toxicity rabbit hole: a framework to bias audit large language models with key emphasis on racism, antisemitism, and misogyny*. Rochester Institute of Technology. Available at: <https://www.rit.edu>.

Ghenai, A., Noorian, Z., Moradisani, H., Abadeh, P., Erentzen, C. and Zarrinkalam, F. (2025) 'Exploring hate speech dynamics: the emotional, linguistic, and thematic impact on social media users', *Information Processing and Management*, 62, p. 104079. <https://doi.org/10.1016/j.ipm.2025.104079>.

Global Project Against Hate and Extremism (2025) *Grok's recent antisemitic spree is proof AI needs cleaning up*, 14 July. Available at: <https://globalextrémism.org/post/groks-recentantisemitic-spreel/>.

Gold, H. (2025) 'AI's antisemitism problem is bigger than Grok', *CNN*, 15 July. Available at: <https://www.cnn.com/2025/07/15/tech/ai-artificial-intelligence-antisemitism>.

Government of Canada (2025) *Canada introduces legislation to combat hate crimes, intimidation and obstruction*, backgrounder, 19 September. Department of Justice Canada. Available at: <https://www.iustice.gc.ca/eng/csj-sic/pl/c9/>.

Institute for Strategic Dialogue (2025) *Monitoring online hate speech in British Columbia*, 24 January. Available at: <https://www.isdglobal.org/isd-publications/briefing-monitoring-online-hatespeech-in-british-columbia/>.

Jang, Y. and Ko, B. (2023) 'Online safety for children and youth under the 4Cs framework: a focus on digital policies in Australia, Canada, and the UK', *Children*, 10(8), p. 1415. <https://doi.org/10.3390/children10081415>.

Joseph, J. (2022) *Centering survivors and taking action on gendered online hate in Canada: national report*. YWCA Canada. Available at: <https://ywcacanada.ca/wpcontent/uploads/2022/11/Block-Hate-Report-October-2022-corrected-1.pdf>.

Karadeglija, A. (2024) 'AI-powered hate content is on the rise, experts say', *CBC News*, 26 May. Available at: <https://www.cbc.ca/news/politics/ai-hate-content-1.7215369>.

Kataite, S. (2025) 'Canada has a disinformation problem — and the tools to fix it', *Canadian Centre for Policy Alternatives*, 7 May. Available at: <https://www.policyalternatives.ca/newsresearch/canada-has-a-disinformation-problem-and-the-tools-to-fix-it/>.

MediaSmarts (2023) *Canadians in a wireless world, Phase IV: encountering harmful and discomfoting content online*. Available at: <https://mediasmarts.ca/sites/default/files/202301/Encountering%20Harmful%20and%20Discomforting%20Content%20Report%20%20YCW%20Phase%20IV.pdf>.

Parker, S. and Ruths, D. (2023) 'Is hate speech detection the solution the world wants?', *Proceedings of the National Academy of Sciences*, 120(10), e2209384120. <https://doi.org/10.1073/pnas.2209384120>.

Reuters (2025) 'EU is in touch with X regarding hate speech content on Grok', 20 November. Available at: <https://www.reuters.com/business/eu-touch-with-x-regarding-groks-content-202511-20/>.

- Sahin, F. (2024) 'Online hate among Canadian youth', *Canadian News Hub*, 27 February. Available at: <https://canadiannewshub.ca/online-hate-among-canadian-youth/>.
- Samara Centre for Democracy (no date) *SAMbot*. Available at: <https://www.samaracentre.ca/initiatives/sambot>.
- Sears, C. (2024) 'Youth advocates funded by CIRA grant poised to influence Online Harms Legislation', *Canadian Internet Registration Authority*, 3 April. Available at: <https://www.cira.ca/en/resources/news/net-good/youth-advocates-funded-by-cira/>.
- Southern Poverty Law Center (2025) *How Meta's policy updates could encourage hate and threaten democracy*, 24 January. Available at: <https://www.splcenter.org>.
- Statistics Canada (2024) *Online hate and aggression among young people in Canada*, 27 February. Available at: <https://www150.statcan.gc.ca/n1/daily-quotidien/240227/dq240227beng.htm>.
- Tanner, S. and Gillardin, F. (2025) 'Toxic communication on TikTok: sigma masculinities and gendered disinformation', *Social Media + Society*, 11(1). <https://doi.org/10.1177/20563051251313844>.
- Taylor, J. (2025) 'AI chatbot "MechaHitler" could be making content considered violent extremism, expert witness tells X v eSafety case', *The Guardian*, 15 July. Available at: <https://www.theguardian.com/technology/2025/jul/15/x-esafety-ai-chatbot-grok>.
- Thompson, S.A. (2025) 'Here's what Elon Musk sees when he opens X', *The New York Times*, 15 May. Available at: <https://www.nytimes.com/interactive/2025/05/15/business/elon-musk-xtwitter-feed-following-followers.html>.
- Young, A. (2025) 'Australia must stop overlooking misogynistic youth extremism', *The Strategist*, 15 May. Available at: <https://www.aspistrategist.org.au/australia-must-stopoverlooking-misogynistic-youth-extremism>.
- TikTok statistics Canada 2024* (2024) Made in CA. Available at: <https://madeinca.ca/tiktokstatistics-canada/>.

About the Montreal Institute for Global Security

The Montreal Institute for Global Security (MIGS) is a Canadian think tank dedicated to strengthening democratic resilience and addressing emerging global security challenges. MIGS works at the intersection of geopolitics, technology, and human rights, producing policy research and convening global leaders to develop solutions to some of the most pressing threats facing democratic societies today. Through research, high level dialogues, and strategic partnerships, MIGS contributes to policy debates on issues such as authoritarian influence, transnational repression, emerging technologies, and international security cooperation. The institute engages policymakers, scholars, civil society leaders, and industry exper

About the Montreal Institute for Global Security in Canada and internationally.

About the author

Jack Rath is the Global Security Officer at the Montreal Institute for Global Security. His research examines emerging technology threats, the geopolitical dynamics of AI development, and policy frameworks to protect information ecosystems and democratic resilience. He holds a dual Bachelor of Arts and Bachelor of Advanced Studies in Politics, International Relations, and Political Economy from the University of Sydney, and is currently completing a Master in International Security at Sciences Po's Paris School of International Affairs. Jack previously served as a Senior Policy Officer in the Australian Government, where he worked to respond to international crises. He speaks English and French at a professional level.

www.migsinstitute.org

©2026 Montreal Institute for Global Security. All rights reserved. The views expressed in this publication are those of the authors and do not necessarily reflect the views of the Montreal Institute for Global Security or its partners.