Private Messaging Public Harms

Disinformation and Online Harms on Private Messaging Platforms in Canada

May 2021

Sam Andrey | Alexander Rand | M.J. Masoodi | Stephanie Tran

RYERSON LEADERSHIP LAB>>>

ROGERS cybersecure catalyst

How to Cite this Report
Andrey, S., Rand, A., Masoodi, M.J. & Tran, S. (2021, May). *Private Messaging, Public Harms*.
Retrieved from https://www.cybersecurepolicy.ca/privatemessaging

350 Victoria St, Toronto, ON M5B 2K3
ISBN: 978-1-77417-024-3

Contributors
Sam Andrey, Director of Policy & Research, Ryerson Leadership Lab
Karim Bardeesy, Executive Director, Ryerson Leadership Lab
Sumit Bhatia, Director of Communications and Knowledge Mobilization, Rogers Cybersecure Catalyst
Jahanmeer Bhatti, Research Assistance, Ryerson Leadership Lab
Zaynab Choudhry, Design Lead
Charles Finlay, Executive Director, Rogers Cybersecure Catalyst
Braelyn Guppy, Marketing and Communications Lead, Ryerson Leadership Lab
Mohammed (Joe) Masoodi, Policy Analyst, Cybersecure Policy Exchange
Alexander Rand, Research and Policy Assistant, Cybersecure Policy Exchange
Stephanie Tran, Research and Policy Assistant, Cybersecure Policy Exchange
Yuan Stevens, Policy Lead, Cybersecure Policy Exchange

This project is sponsored by the Government of Canada through
the Privy Council Office's Democratic Institutions Secretariat.

@cyberpolicyx @cyberpolicyx Cybersecure Policy Exchange

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

Executive Summary

More than eight in ten people in Canada use online private messaging platforms, such as Messenger, WhatsApp and Snapchat; and over half are receiving messages about the news or current events at least weekly. This growing vector for news is coming under increased scrutiny, as evidence from jurisdictions around the world reveal private messaging apps' role in spreading disinformation and a broad range of already-illegal materials, including hate speech, inciting violence, cybercrime, sexual abuse and child sexual exploitation materials.

The spread of disinformation and other online harms poses risks for social cohesion, public safety and democracy; and, as a result, has raised calls for technical and regulatory changes. At the same time, concerns have also been raised regarding over-censorship of content and that any such changes may negatively impact freedoms and rights, particularly the right to free expression. Adding to the complexity, cybersecurity and privacy experts fear that police and security agencies may use online harms to justify the weakening of encryption technologies used by many messaging apps, allowing them access to message content, and thus posing significant implications for privacy rights and civil liberties.

To date, Canadian regulatory proposals has focused largely on social media content that remains publicly accessible. The purpose of this report is to explore **the role of private messaging in information ecosystems in Canada**, including Canadians' use of messaging apps; their exposure to disinformation and other online harms; and potential policy and technical approaches to mitigate potential harms.

A representative survey in March 2020 of 2,500 people in Canada found that:

- **46% report receiving private messages that they suspect are false** at least monthly;
- 39% report receiving private messages that they initially believe to be true, but later find out are false, at least monthly;
- **26% report receiving messages containing hate speech** at least monthly, with rates higher among people of colour;
- A majority have about the same level of trust in news that they receive through messaging apps as they do in news from websites, television or social media;
- Those who use private messages as a news source report seeing false information more frequently, with a majority of Telegram and WeChat users receiving false information at least weekly; and
- **Those who believe in COVID-19 conspiracy theories are more likely to regularly receive news through private messages** (77% more likely to get news through WhatsApp and 34% more likely to get news through Facebook Messenger).

Against this backdrop, messaging platforms and some governments around the world have taken some steps in an attempt to mitigate disinformation and other online harms, including:

- Labels and limits on message forwarding to create more friction for messages to go 'viral';
- Limits on group size to reduce message reach;
- Mechanisms to enable users to report harmful content to moderators; and
- Features to encourage users to verify information they receive.

We conclude with three recommendations for the Government of Canada to better understand and mitigate these complex challenges:

1. Invest in **research and innovation** specific to disinformation and other online harms on private messaging platforms in Canada;
2. **Require transparency** from large online platforms to better understand online harms through private messaging; and
3. Make investments in **policy-informed digital literacy efforts** that build resilience to disinformation through private messaging platforms.

Introduction

In 2014, Facebook acquired a messaging app that few in North America had heard of — WhatsApp. The $19-billion purchase for an app with an estimated $20 million in revenue, drew a stunned response from industry observers, prompting *Time Magazine* to ask, *"What is WhatsApp?"*[1] For Facebook, the purchase was a calculated move to reach WhatsApp's quickly growing global user base (estimated at 450 million active monthly users at the time) and its increasing share of online engagement time. Today, WhatsApp's active user base has more than tripled in size,[2] becoming the most popular messaging app in the world, with over two billion average monthly users in 180 countries.

Since then, several messaging applications have risen in popularity, with many becoming a staple of modern communications in different parts of the world, including Canada.[3] A variety of economic, social, political and technical factors help explain these trends. The ability to send and receive text messages, photos, videos and calls to people across the world using the internet, rather than a telecommunications network, is thought to have driven messaging apps' popularity, particularly in countries with underdeveloped or expensive cellular networks. Some states have banned certain platforms, leading to the rise of one platform over another. Several messaging apps have also proven popular among different diasporas as a convenient way to communicate with loved ones back home. For instance, in Canada, 84% of newcomers use WhatsApp daily, while 60% of recent immigrants from China use WeChat daily.[4]

The shift is also part of a strategy by major platforms to address falling engagement on open social networks.[5] As Facebook's Mark Zuckerberg stated in 2019:

> "Today we already see that private messaging, ephemeral stories, and small groups are by far the fastest growing areas of online communication. There are a number of reasons for this. Many people prefer the intimacy of communicating one-on-one or with just a few friends. People are more cautious of having a permanent record of what they've shared… In a few years, I expect future versions of Messenger and WhatsApp to become the main ways people communicate on the Facebook network."[6]

Moreover, COVID-19 and its accompanying restrictions on in-person engagement have accelerated growth in messaging app communications, with total messaging on WhatsApp and Messenger reportedly growing by over 50% in the first month of the pandemic.[7]

The widespread growth of messaging apps has led to increasing research interests around the world in understanding the varied impacts of the technologies, including — and perhaps most controversially — their role in facilitating the spread of disinformation. Indeed, there is increasing concern among experts, researchers and policymakers regarding the rise of online disinformation, cybercrime, and other hateful and illegal content, and their toxic effects on social cohesion, public safety and democracy. To date, there has been no significant research on the issue of online disinformation or harms specifically on private messaging platforms in the Canadian context.

The purpose of this report is to outline the **realities and policy challenges of mitigating disinformation and other harmful online materials shared through private messaging applications and services in Canada.** Mixed methods were used to explore these issues, as well as potential solutions.

We use the terms "messaging applications" or "messaging platforms" to mean internet-enabled media designed to offer two-way and multi-way communication between a finite number of users determined by the sender.[8] These platforms are sometimes also referred to as "closed" messaging platforms, to differentiate from the more open format of platforms like Facebook, Twitter or YouTube, and the much older protocol of e-mail. They are also sometimes called direct messaging applications. We have chosen to use the term "private" messaging, as we believe it most accurately shapes the object of debate for policymakers and users. Out of scope for this report are social media interactions that blur the line between 'public' and 'private' such as Facebook groups and Instagram 'close friend' stories, among others.

It is important to stress that this is a highly complex policy challenge. There is an inherent tension at play between mitigating online harms and competing democratic values of free expression and privacy. Therefore, there is no perfectly calibrated solution to address disinformation on private messaging apps. Content moderation of digital communications can easily infringe on free expression to an undesirable degree; while a lack of content moderation can risk the further spread of harmful content, potentially posing real-world consequences. The fact that many of the messaging apps and services offer end-to-end encryption to preserve the privacy of messages makes this tension even more challenging.

This report attempts to wrestle with, and be sensitive to, these tensions and explore Canadian views on this complex challenge, while still recommending a path forward to better understand and tackle the challenge in Canada.

Methodology

This report used a mixed-methods approach to explore the policy challenges of disinformation and other online harms on private messaging platforms in Canada. A **jurisdictional and literature scan** was conducted in order to review the extent of the challenges faced in Canada and other jurisdictions around the world, as well as the types of policies that other governments have considered or adopted. Unstructured **interviews** to better understand the jurisdictional landscape were also conducted with experts:

- **Aviv Ovadya,** Founder of the Thoughtful Technology Project;
- **Craig Silverman,** Media Editor for BuzzFeed News; and
- **Dr. Claire Wardle**, Co-Founder and Leader of First Draft News at the Shorenstein Center on Media, Politics and Public Policy at Harvard University.

We also conducted a **national representative survey** to better understand the use of private messaging apps in Canada. Abacus Data administered an online survey to 2,500 residents of Canada age 16 and older between March 17 and 22, 2021. A random sample of research study panelists were voluntarily invited to complete the survey, which included response quotas to ensure the results were representative by language, age range, gender and region. The survey sought to: 1) Collect up-to-date detailed data on the use of private messaging apps in Canada, including demography, frequency of use, levels of trust toward messages received, and the typologies of disinformation users are most exposed to; and 2) Analyze the relationship between messaging app use and online harms (see questionnaire in Appendix 3). The margin of error for a comparable probability-based sample of the same size would be ±2 percentage points, 19 times out of 20. The data were weighted according to the latest Canadian census to ensure that the sample matched Canada's population by age range, gender, educational attainment and region. Totals may not add up to 100% due to rounding.

Finally, an invitational **stakeholder workshop** was held on March 12, 2021. Participants were selected to represent academia, industry, policymakers, journalism and civil society, including community groups representing ethnocultural groups (see full list in Appendix 1). This participatory workshop used Chatham House Rule to allow the authors to elicit rich and in-depth information by creating a respectful space of interaction for participants to share and discuss the issues and concerns relevant to the study. Workshop facilitators provided opening remarks as a synopsis to the current public policy challenge; and pre-determined questions followed suit, allowing participants to engage and discuss (see Appendix 2). Facilitators took workshop notes, which informed the drafting of this report.

A convergent design was used to validate and compare the findings from the various data sources.[9] This involved assessing the sources of data, separately analyzing the data, and reviewing results through side-by-side comparison. Such an approach enables validation between the various sources.[10]

It is important to note that the varied perspectives of our participants, including experts and stakeholders, greatly informed this report; however, the statements and recommendations are solely those of the authors.

Though not an exhaustive list, policymakers increasingly describe the range of illegal and

What Are Online Harms?

Though not an exhaustive list, policymakers increasingly describe the range of illegal and harmful conduct online to include:[11]

Disinformation: The deliberate spread of false narratives[12] and information that undermine political processes,[13] national security[14] and public health.[15]

Hate and Harassment: Online hate speech against identifiable groups and harassment can lead to real-world harm, particularly for minority groups.[16]

Incitement to Violence: Online media have been used to mobilize and incite violent events in many countries,[17] including the Québec City mosque shooting, the Toronto van attack and the storming of Rideau Hall.

Sexual Violence: People can use online applications to spread sexual abuse material, including child exploitation and non-consensual sharing of intimate images.

Fraud and Cybercrime: Criminals use online platforms to deceive and impersonate others, including to gain access to personal information and bank accounts.

Background

Online Disinformation and Private Messaging Apps

The role of social media and Cambridge Analytica in the 2016 U.S. election and the Brexit referendum sparked many Western governments' concerns over online disinformation and its impacts on democracies. Subsequent research on social media and disinformation soon followed and frequently made reference to the polarizing political discourse fostered by online platforms, as well as individual behaviour modification made possible by algorithmic profiling in ways that undermined voter autonomy, free elections and even democracy itself.[18] More recently, research has demonstrated that public fear surrounding COVID-19 — coupled with demands for rapid answers — has led to **even more false, misleading or misinterpreted information online**, often with a politically motivated goal to sway public opinion.[19]

In a 2019 survey, eight in ten Canadian residents indicated that an increase in deliberately false information was a problem affecting Canadian society.[20] About half reported seeing **deliberately false or misleading news** or political information online at least weekly, and those who use social media platforms for their news were significantly more likely to report encountering false and hateful content. While Canadians have relatively high levels of trust in news media compared to international peers, a 2019 survey conducted by the Canadian Journalism Foundation showed that social media is the most-used source of news for Canadians, while also being the least-trusted source.[21] This trend was accentuated among millennials, who received news via social media more often than the general population, while also revealing low levels of trust toward those sources compared the general population. Similarly, a 2020 survey indicated that, while nearly all Canadians said they had experienced COVID-19 misinformation online, only about 20% verified the false claims they encountered, and half said they **shared false information they found online before discovering that it had, in fact, been false.**[22] Recent research from the Institute for Research on Public Policy found a relationship between susceptibility to false news about COVID-19 and reduced propensity to abide by public health recommendations,[23] with further research pointing out a connection between online information and vaccine hesitancy in Canada.[24] In addition, a 2021 poll commissioned by the Canadian Race Relations Foundation found that nearly 80% of Canadians were concerned about the **spread of hate online**, with large majorities also specifically concerned about right-wing extremism and growing political polarization.[25]

The current online ecosystem has been described as an **"infodemic"** — a chaotic information environment consisting of an abundance of inaccurate information that not only undermines the effectiveness of public health measures in the current global health pandemic context, but also leads to real-life violence,[26] discrimination,[27] chilling effects[28] and confusion. At a macro level, many experts have pointed to the long-term consequences of disinformation, including its role in the erosion of democratic values and principles, human rights and social cohesion.[29]

Although much of the focus of research and journalism on disinformation has been on social media platforms where content shared remains generally open and accessible to the wider community to view, comment on and share, there has been relatively less scrutiny on the role of private messaging apps in the propagation of disinformation. Such platforms — particularly WhatsApp — have received significantly more political and academic attention and investigation in other parts of the world. These platforms are often widely used in the developing world, and have been subject to criticism for their role in spreading disinformation and influencing elections, including in Brazil, India and Indonesia.[30]

While it is unclear whether those running the online platforms themselves have a complete picture of the substantial amount of harmful content shared on their messaging apps, it is clear that the small percentage of communications that are harmful are sufficient to produce real-world harms. One challenge associated with combating harmful content in this space is that these small kernels of harmful content may be nested within an overpowering abundance of innocuous chatter. In other words, for all the socially useful possibilities enabled by online platforms, there nonetheless remain **significant risks to social cohesion and the integrity or quality of information that is shared.**

In North America, messaging apps have often evaded policy discussions up until recently, with the U.S. Capitol riots playing a particularly catalyzing role. Soon after the riots at the U.S. Capitol, news reports began to reveal that many of those who stormed the Capitol were using messaging apps to communicate and coordinate. The 'closed' design of these messaging applications allows for smaller group-level discussions between users and are often encrypted, making it not only more difficult to monitor the spread of harmful material, but also a challenge for researchers to study using the methods often employed to track disinformation on open platforms, like Twitter or Facebook. Proposals and actions to date from the platforms have ranged from adding friction, such as group size and message forwarding limits, to mechanisms that enable users to verify information more easily.

As social media continues to create harm in the real world, people have called for governments to play a larger role in regulating it. These calls have come from governments,[31] academics[32] and civil society at large.[33] In some cases, even the social media companies themselves have echoed these calls.[34] As part of its multi-pronged approach in regulating the big tech giants, the Canadian government announced its intention in late 2019 to introduce new regulations to regulate the timely removal of illegal content on open online platforms.[35]

Adding to the complexities are concerns about over-censorship and implications for freedom of expression.[36] In addition, cybersecurity and privacy experts fear that police and security agencies may use events like the U.S. Capitol riots to justify arguments for weakening encryption technologies used by messaging apps, enabling them access to otherwise unavailable content — and thus posing significant risks to privacy rights and civil liberties. Although it is not yet clear if Canada's proposed legislation will attempt to regulate content moderation on messaging apps and services, the government has expressed its views on the dangers of disinformation, with the President of the Queen's Privy Council suggesting that the government "would be open" to considering legislation that makes it an offence to knowingly spread such materials.[37]

Thus, questions remain as to how messaging apps in Canada could be impacted by such regulatory developments, as well as what would be the short- and long-term impacts of any policies that bring private messaging apps into their fold. Chief among these concerns is striking the appropriate balance between free expression and the prevention of online harm, as in the case with social media content regulation in Canada and abroad.

These policy questions exist, as do so many policy questions about the regulation of communications technologies generally, in a context in which the technologies themselves often evolve and are adopted at a rate faster than public policy's current ability to adapt. The technologies are powered by data sets and network effects that (apart from internet protocols themselves) are often under private domain; or controlled by, or disproportionately enjoyed by, a handful of large platform technology companies. These data sets and benefits of network effects are not available in the same way to governments, regulators or members of the public generally.

What is End-to-End Encryption?

The issue of disinformation on private messaging apps is tied to the debate on encryption policy more generally, because several of the most popular messaging platforms offer end-to-end-encryption by default for message content. This prevents anyone other than the sender and recipients of a message from viewing the content of that message. For example, under end-to-end encryption, not even WhatsApp can read messages sent using its platform. Instead, the platform can only collect metadata related to users and messages, such as where, when and how the app is used.[38] Other platforms, notably Signal, do not collect any such metadata besides those required for the app to perform its basic functions.[39] Other encrypted messaging apps, such as LINE and Viber, fall along a spectrum in terms of the metadata they collect.[40]

The core dynamic of the encryption debate is the tension between preserving privacy and mitigating online harms. End-to-end encryption is seen as an important technical bulwark against threats to individual rights, democracy and national security.[41] However, others criticize the constraining effect that encryption has on law enforcement and security agencies.[42] Perhaps, most notably, state security actors themselves tend to resent encryption as a hindrance to the investigations and intelligence gathering activities that they are legally authorized to carry out.

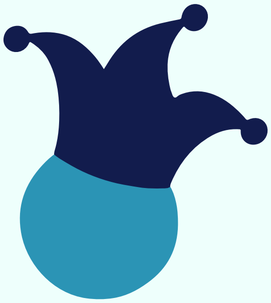
A detailed discussion of the debate around law enforcement's access to encrypted messages is beyond the scope of this report. However, the fact that many private messaging applications use end-to-end encryption means that the tension between privacy and harm mitigation plays a central role in discussions of regulating those platforms. This creates challenges for policymakers that do not exist when trying to regulate public, unencrypted platforms such as Facebook and Twitter.

What is Mis- and Disinformation?

Misinformation is information that is false and spread regardless of intention to deceive, including unintentionally. In contrast, **disinformation** is information that is false — and the person disseminating it *knows it is false*. In this instance, the dissemination of the false information is deliberate and intentional, and is often performed by malicious actors.[43] Many instances of disinformation are then spread further unintentionally as misinformation, including through private messages.

Malinformation, on the other hand, is authentic information used to inflict harm, such as leaks of private information or forms of harassment and hate speech.

Disinformation can come in many forms, and some of these forms are more harmful than others.[44] The harmfulness of a piece of disinformation can depend on whether it is intended to deceive people. Research by Dr. Claire Wardle has identified at least seven different types of mis- and disinformation, ranked based on their harmfulness and intent to deceive (least to most):[45]



Satire. When aspects of a true story are misrepresented for the purpose of humour. Satire is often easy to identify because it intentionally distorts reality in ways that seem exaggerated and ridiculous. For this reason, it can be less harmful than other types of disinformation. But even satire has been known to mislead people, with real-world consequences.[46]

False Connection. A piece of media uses an image or a headline that does not correspond to the actual content of that media. Often called "clickbait," a media outlet may use this tactic to try and draw people into clicking on their articles for economic benefits. Alternatively, they can be the product of poor journalism, with images or headlines carelessly attached to stories for which they are ill-suited.

Misleading Content. A piece of content includes only a few selected aspects of the story that it claims to represent. For example, a written story could use a quotation without providing adequate context, thereby distorting the intended meaning of the speaker. Similarly, a piece of content could include a cropped photo or video that shows only one aspect of an event, while concealing other salient aspects. This type of selective presentation can be done on purpose, with an intention to mislead. But it can also be done inadvertently, as the product of poor journalism.

False Context. A video or photo is shared with the claim that it shows a particular event, but in fact does not show that event. This is done initially with an intention to mislead, but can later be shared by users who have been fooled by the initial post. Even the official White House Twitter account following the U.S. Capitol uprising has shared videos accompanied by false context in an effort to advance its political messages.[47]

Imposter Content. A piece of content is shared by someone falsely claiming to be a member of a well-known organization. One recent example of this was the use of a fake BBC news logo in a story covering the Kenyan elections of 2017.[48] This story was shared widely on WhatsApp before being identified as fake. This type of manipulation is, of course, done with a clear intention to mislead.

Manipulated Content. A piece of media is digitally altered to depict events that did not actually occur. Perhaps the most well-known example of media manipulation technology is Photoshop, which has been widely used for many years to manipulate photos. Such digital tools to manipulate audio and video have become popular in recent years. While often created with a clear intention to mislead, manipulated content can also be used for satirical purposes.

Fabricated Content. A piece of media is created 'from scratch,' rather than simply being altered or recontextualized to mislead. This could be in text form, such as a news story that is pure fiction. However, recent advances in artificial intelligence have produced tools that can generate video and audio from scratch, without directly altering a particular piece of media. These generated video and audio clips are sometimes called *Deepfakes*.

Icon Illustrations inspired by First Draft "Fake news. It's complicated." by Claire Wardle

Private Messaging Harms

Global Understandings to Date

Election Interference

There have been numerous studies on the role of private messaging apps in propagating mis- and disinformation during elections.

Disinformation on private messaging apps played an important role in the Brazilian election of 2018.[49] In one study, researchers used a tip line established in partnership with 24 Brazilian media outlets, in order to document thousands of instances of misleading political information on WhatsApp, the country's most popular private messaging app. Using this large set of crowdsourced messages, the researchers were able to get a glimpse into the details of the mis- and disinformation ecosystem. These researchers found that the most common form of misleading information involved visual images, with video and text seeing only about half as many instances, respectively. These images tended to make use of what First Draft News (see above) refers to as *false context*; that is, they would often include photos of real political documents, news events or statements from public officials, but with a caption that suggested a false context. The primary false narrative being pushed by groups favouring Jair Bolsonaro for President claimed that the integrity of the election was under threat, pointing to opponents of Bolsenaro who were accused of making efforts to rig the outcome.

One study found that nearly 70% of viral audio-based misinformation included claims of election fraud. These messages spread virally on WhatsApp using what researchers described as "the tactics of mid-1990s chain emails."[50] For example, one message that saw wide diffusion read: "if you send this message to just 20 contacts in a minute, Brazil will unmask this criminal. DO NOT break this chain. The unwary must know the truth."

Researchers wrote that, if they could choose one piece of misinformation to represent the overall mis- and disinformation picture during the Brazilian election, it would be "a real picture of electronic ballot boxes, presented out of context, denouncing electoral fraud meant to harm then-candidate, now-President Jair Bolsonaro. That image would be coupled with a short text mixing real and false misdoings from the opponent's party, urging everyone to share it wildly."[51] The fact that many of these messages were widely distributed via private messaging platforms highlights an important ambiguity around the distinction between public and private political speech.

The Indian elections of 2019 also saw widespread use of misleading information on WhatsApp.[52] With over 400 million WhatsApp users in India, the platform was a key battleground for political parties during the election. For example, the governing party of Prime Minister Narendra Modi recruited up to one million "WhatsApp Volunteers" to create a significant number of interconnected WhatsApp groups for the dissemination of party information.[53] Mis- and disinformation in these groups ultimately stoked racial tensions between Hindu and Muslim communities, and promoted Hindu nationalism to energize the party's political base. For example, one such 'WhatsApp Volunteer' was quoted as saying that he used WhatsApp to communicate with

60 voters who had been assigned to him. He shared numerous stories about anti-Hindu violence perpetrated by Muslims, including some stories promoted by Modi's party — which have been debunked as misinformation — as well as fake polls suggesting that Modi's party was performing much more strongly than it was in reality. This example illustrates the spread of mis- and disinformation in a highly distributed fashion, making use of party volunteers to share misinformation with voters either via small networked WhatsApp groups or directly to individuals through private messages. While WhatsApp is the most-used private messaging application in India, other platforms have also been used to spread disinformation, such as ShareChat and Helo.[54]

Disinformation on WhatsApp also influenced the political landscape during the Indonesian election of 2019.[55] This influence included the use of hoax campaigns by politically motivated actors. For example, several unfounded news stories circulated on WhatsApp suggesting that both domestic and foreign state actors such as China engaged in election interference. This false story spread virally in the form of an audio message on WhatsApp. Similarly, a narrative claiming that domestic state police were interfering in the vote-counting process began circulating on WhatsApp. One study found these narratives had a polarizing effect in Indonesia, and undermined trust in its electoral institutions.[56]

Undermining Public Health

In the early days of the pandemic, the Government of India issued a statement warning citizens not to trust rumours spreading on messaging apps such as WhatsApp.[57] This followed a wave of unfounded suggestions for avoiding or curing the illness, as well as *false context* media purporting to show the devastating impact of the virus in other countries. Much of this information was shared on WhatsApp, owing to its role as India's most widely-used messaging application.

One recent study cites an increase in the use of private messaging apps in Turkey as a major challenge for countering COVID-19 disinformation in that country.[58] Turkey also saw the growth of narratives that were specific to the Turkish context. These narratives tended to relate the pandemic to internal social political divisions. Similarly, a study on disinformation in Malaysia argues that addressing false information about COVID-19 on WhatsApp in that country amounts to a "Herculean task".[59]

State Actors

Another challenge stemming from private messaging platforms is that state actors appear to be increasingly turning to encrypted messaging platforms to carry out international propaganda campaigns.[60] Russia's history with the private encrypted platform Telegram is emblematic of this emerging reality. The country initially sought to ban the platform within its borders in 2018 due in part to a failure to ensure the decryption of user data upon request by the state. However, this legal ban did little to block the actual use of the platform in Russia. Even while the platform was legally outlawed, the Russian government used it to spread information concerning the COVID-19 pandemic. In 2020, the Russian government unblocked the platform. In the interim, Russian state actors appeared to have made extensive use of Telegram to spread propaganda among Ukrainian users of the app in support of Russian strategic objectives in the region.[61] There have also been cases of Facebook pages pushing disinformation, driving users to Telegram channels.[62]

In China, the popular messaging platform WeChat is widely known to be a part of the Chinese government's mass surveillance network.[63] Not only are users' activities analyzed, tracked and shared with Chinese authorities upon request, but the app also censors what it deems politically sensitive topics. This includes data sent from WeChat users registered outside of China, which is also surveilled and used to further build censorship algorithms in China.[64]

Fear and Violence

False stories spreading on WhatsApp have enabled the activity of violent mobs in places such as India.[65] Specifically, false stories circulating on WhatsApp purported that a group of child kidnappers were making their way into various villages across the country, creating widespread fear, and leading to attacks and killings of people falsely accused. In Sri Lanka, the government temporarily banned WhatsApp and Viber to stop the spread of rumours following a wave of terrorist attacks and to curb mob violence.[66] WhatsApp was also identified as a source of disinformation contributing to the arson and vandalism attacks against cell phone towers in the United Kingdom. Specifically, conspiracy theories began to emerge, falsely tying the emergence of the COVID-19 pandemic to the implementation of 5G infrastructure, leading to the cell phone tower attacks.[67]

White Nationalism and Radicalization

The activities of White nationalist groups on public, unencrypted forums such as Facebook have been well-documented in the Canadian context.[68] But these groups also appear to have begun mobilizing on encrypted private messaging platforms, where their activities have received less attention from the research community. The messaging app Telegram gained popularity among White nationalists during former President Trump's term.[69] Several news reports reveal that the platform features large groups of anonymous users with names such as "Only White Lives Matter," who promote White nationalist narratives and violence. Studies suggest that these groups are highly networked, with over 20% of overall content being forwarded from another group; and it has been demonstrated elsewhere that non-violent content on social media can often act as a gateway to more violent and extreme views.[70]

Public figures who are known to be proponents of White nationalism, including far-right commentators Milo Yiannopoulos and Alex Jones, have attempted to increase their follower count on Telegram following their ban from social media platforms, including Facebook and Twitter, for inciting hate and violence.[71] More recently, following the decision of Twitter and Facebook to ban President Trump's accounts, many of the former President's more extreme followers have begun to move toward encrypted messaging platforms, such as Signal and Telegram.[72] The week that President Trump was banned from major platforms, Signal and Telegram were the number 1 and 2 most downloaded applications on Apple and Google's app stores.[73] However, at the same time, WhatsApp had introduced changes in its terms of agreement, resulting in a large number of users switching to alternate apps that they believed were more privacy-protecting, which may have also influenced the migration.[74]

Studies have also shown that the terrorist organization Daesh has made extensive use of private messaging applications to support its activities, including using Telegram to spread propaganda, recruit new members and encourage violence though methods such as the use of selective hashtags, as well as more explicit recruitment tools such as recruitment videos showing graphic imagery.[75] Some studies have evaluated the impact of platform interventions to disrupt terrorist messaging networks. For example, a 2019 "Action Day" based on referrals from Europol to Telegram of terrorist content resulted in a sustained reduction in the number of terrorist posts on Telegram and in migration to other platforms, such as Twitter, Rocket.Chat, TamTam, nandbox and Hoop Messenger.[76]

Sexual Abuse

Private messaging platforms have been used extensively to distribute sexual abuse material. For example, over 20 million instances of child sexual abuse material were identified and removed from Facebook in 2020, of which over 99% were detected through automated mechanisms rather than user reports.[77] The National Center for Missing and Exploited Children estimates that 70% of Facebook’s reports are from private messages on Messenger and Instagram.[78] However, Facebook Messenger and Instagram direct messages do not offer encrypted messaging by default. The use of encrypted messaging platforms to disseminate this harmful material poses challenges not faced when this material is distributed on public or non-encrypted platforms. For example, Facebook has acknowledged that, although it has seen success in using automated means to identify child sexual abuse material on its non-encrypted platforms, those methods will not be applicable to encrypted platforms. The company reports it has begun to develop new approaches to address the spread of this material on encrypted platforms, including identifying patterns of activity and scanning unencrypted information (such as profile and group information) for abusive content.[79]

Disinformation and Algorithms

Disinformation in private messaging presents different challenges than disinformation on more open platforms, such as Facebook and Twitter. For example, disinformation on open platforms is often spread through recommendation algorithms. A report published by the NGO Avaaz has pointed out that many of the top news sources promoted by Facebook often shared false health information and conspiracy theories, while receiving nearly four times more views than the top sites promoting reputable health information.[80] The ability of disinformation to spread more quickly because of social media algorithms is well-established, with numerous studies drawing connections between the tonal features of disinformation, the psychological propensity to share a particular story, and the tendency of algorithms to promote stories that are already widely shared.[81]

Disinformation on private messaging apps is not necessarily subject to this kind of algorithmic propagation. However, the ecosystem presents challenges of its own; for example, it may be the case that information received through a private messaging app is seen as more credible, or is more likely to be read, when it is received directly from a contact known to the person in real life. A research study conducted in 2018 by Kantar Media found that news content received on WhatsApp was more likely to be trusted than news found on Facebook.[82] This was largely due to the personal nature of the app, such as the close relationship with the sender and the directness of content dissemination. Because the intimacy of private messaging apps can contribute to perceptions of content credibility, a major concern about the spread of false information on private messaging apps is not necessarily virality, but the level of trust placed in the content that is shared.

Open platforms can also guide further traffic toward disinformation groups and materials on private messaging apps by promoting links such as YouTube channels and using hashtags on Twitter in a process sometimes referred to as **"multi-homing."** Human actors are identifying clever strategies and tactics to algorithmically amplify disinformation, including by exploiting the virality function of hashtags. YouTube curation algorithms can lead users down a disinformation 'rabbit hole,' where similar videos are sequentially recommended to watch.[83] Thus, a cross-flow of (dis)information between public platforms and private messaging apps exists. For these reasons, platform and regulatory action to address disinformation on social media platforms can act more broadly to reduce disinformation spread through private platforms as well.[84]

This crossflow of information between private and public platforms raises questions about the appropriate way to counter malicious actors on public platforms. For example, if a group spreading misinformation is pushed from public platforms, it may cause that group to focus their efforts on private messaging applications, where it is more difficult for researchers and authorities to track their activities, and where they may nevertheless be able to indirectly influence more public channels. For this reason, it is still a topic of debate among some observers whether removing malicious actors from public platforms is likely to produce a net-positive effect on the information environment.

Findings

Use and Experiences with Private Messaging in Canada

A national representative survey was conducted to provide an up-to-date account of Canadians' experience with disinformation and other online harms encountered through private messaging, and overall trust in information shared on such platforms.

Overall Use of Messaging Apps

Overall, **83%** of respondents reported using at least one private messaging app in the last year. Facebook's three messaging services — Messenger, WhatsApp and Instagram — remain the dominant private messaging platforms used by those in Canada. The proportion of Canadians that reported using Messenger, WhatsApp, Snapchat and WeChat in the last year were all consistent with past surveys in 2019 and 2020.[85]

Top Private Messaging Apps in Canada

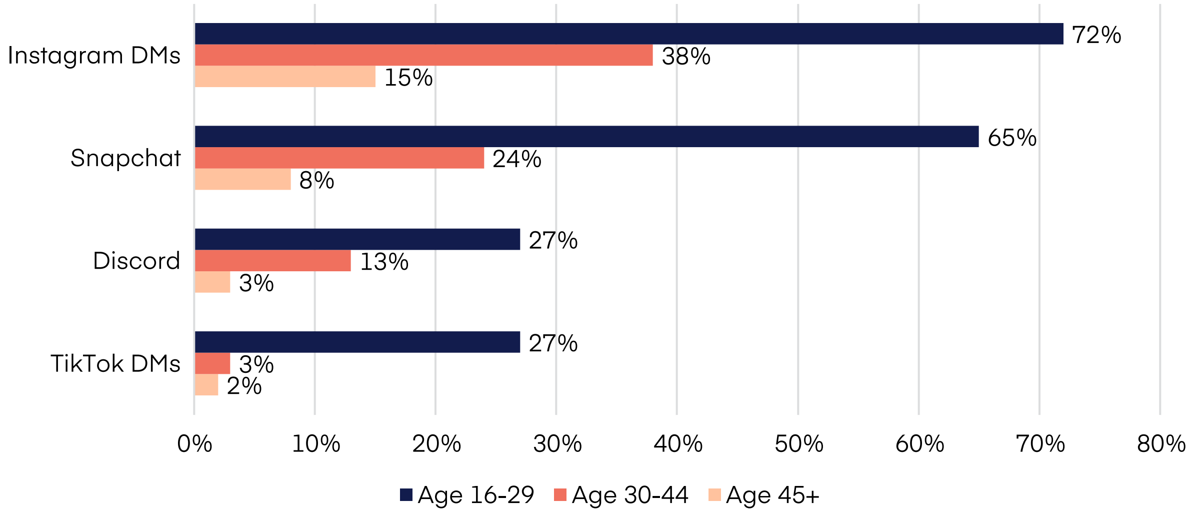
83% Used At Least One Messaging App

72%	Facebook Messenger
35%	WhatsApp
33%	Instagram Direct Message
24%	Snapchat
13%	Twitter Direct Message
10%	Discord
8%	TikTok Direct Message
8%	Telegram
6%	WeChat
5%	Signal
4%	Viber
4%	LINE
3%	QQ
3%	Weibo
2%	Clubhouse
1%	imo

n = 2,451

Four messaging apps had significantly different use patterns across age groups. Instagram, Snapchat, Discord and TikTok were all used by a minority of those aged 45 and older, while used by significant proportions of those aged 16-29.

Private Messaging App Use by Age in Canada



n = 2,451

WhatsApp, Telegram and WeChat each had significantly different use patterns in different **ethno-cultural communities** and among **newcomers** in Canada. Each of these apps are among the dominant messaging apps in South Asia, the Middle East and China, respectively.[86]

WhatsApp:

Overall: 35%

People of Colour: 68%

In Canada <10 years: 84%

Speak Language Other than English or French Most Often at Home: 72%

Middle Eastern: 89%

South Asian: 77%

Telegram:

Overall: 8%

In Canada <2 years: 31%

Middle Eastern: 24%

Latin American: 23%

Black: 20%

WeChat:

Overall: 6%

East Asian: 32%

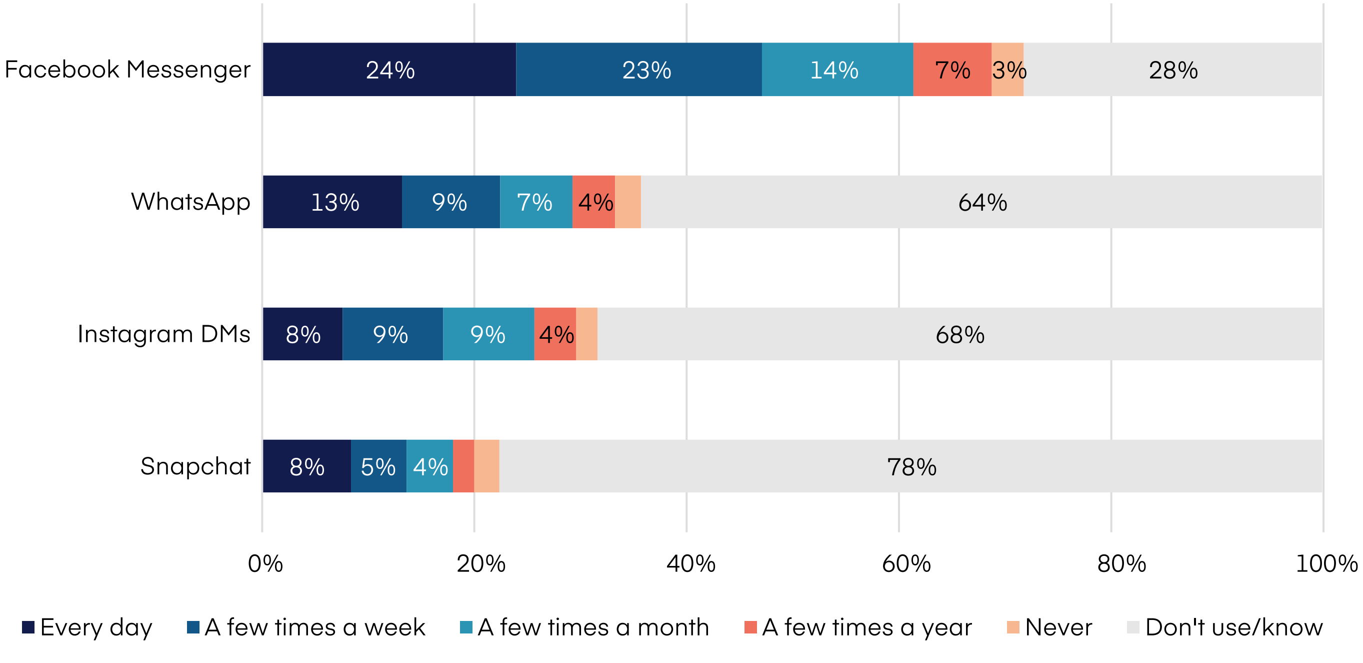
Speak Cantonese or Mandarin Most Often at Home: 48%

Using Messaging Apps for News

When asked which sources they use to stay up-to-date with the news or current events, **21%** of respondents said that they rely on private messages from friends, family or colleagues. This is an increase from 11% when the same question was asked in August 2019.

Overall, Messenger, WhatsApp, Instagram and Snapchat were the top sources of messages about news or current events. Nearly half of respondents report receiving messages about the news or current events at least weekly on Facebook Messenger, with 22% saying the same about WhatsApp and 17% for Instagram direct messages.

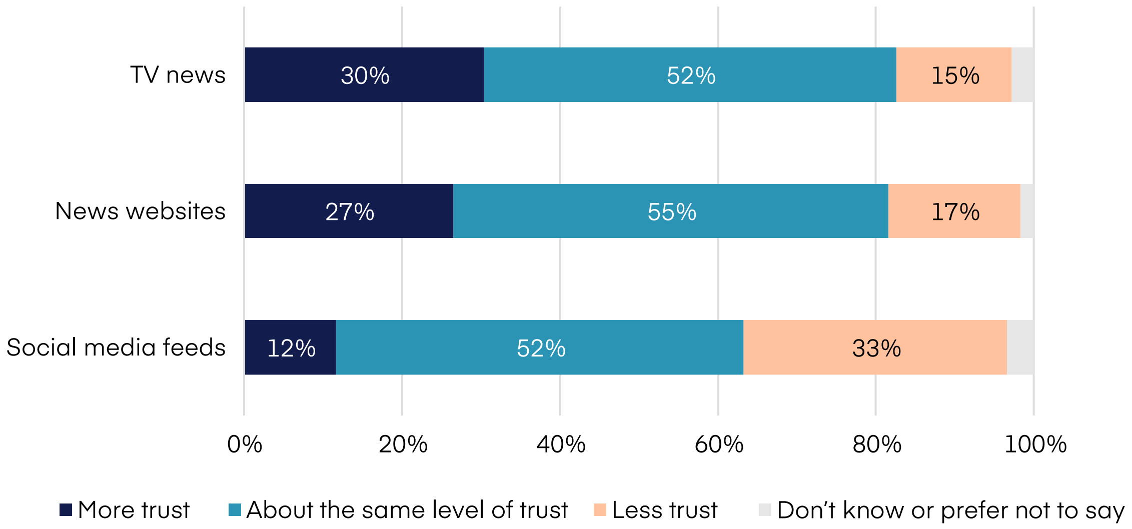
Frequency of Messages About the News or Current Events



n = 2,451

We asked survey respondents how much trust they have that messages they received about the news or current events were accurate and authentic compared to the other sources they use. In each case, a majority indicated that they had about the **same level of trust in information from messaging apps as TV news, news websites and social media feeds.** About one-third have less trust in social media; and a similar proportion have more in TV news and news websites.

Trust that News is Accurate and Authentic Compared to Messaging Apps



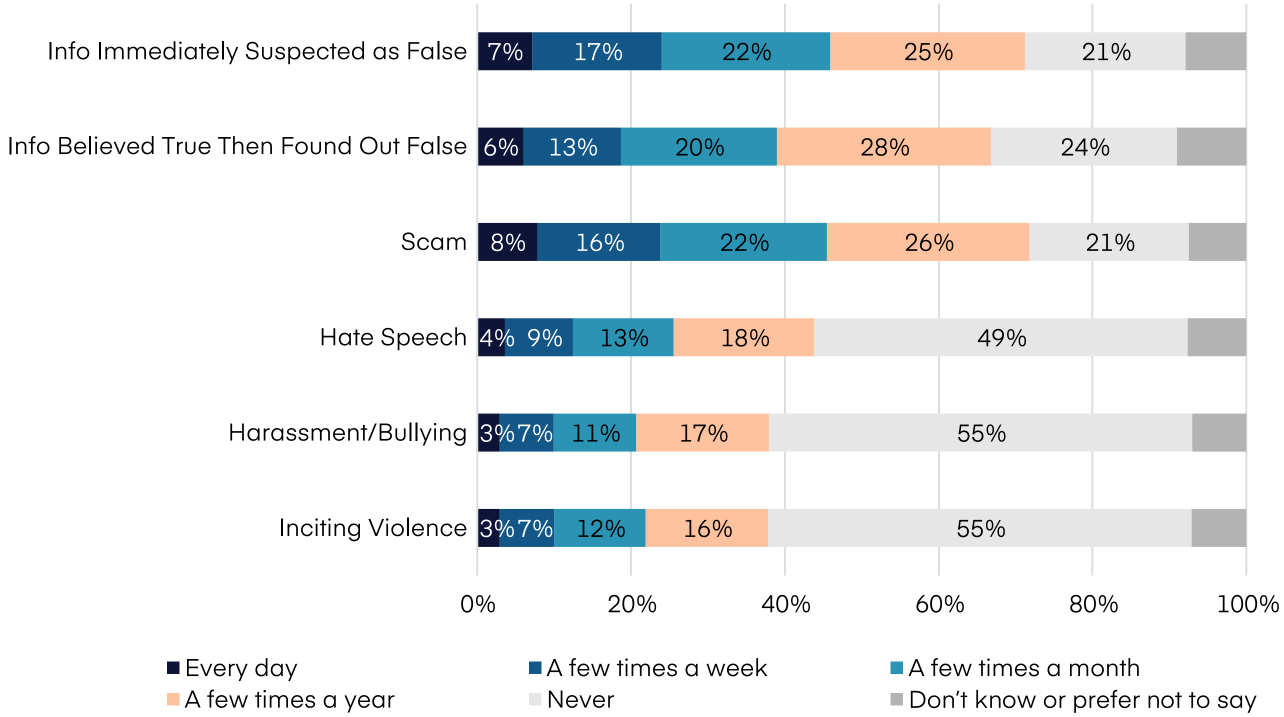
TV news n = 1,204;
News websites n = 975;
Social media n = 892

False Information through Messaging Apps

We asked respondents how frequently they encountered a range of online harms through private messaging apps. About half (46%) reported seeing **information that they immediately suspected was false** at least a few times a month; while 39% reported seeing information that they initially believed was true, but later found was at least partially false, with the same frequency. **Scam or phishing** messages were also reported as a relatively frequent occurrence, with 46% reporting receiving these messages at least a few times a month.

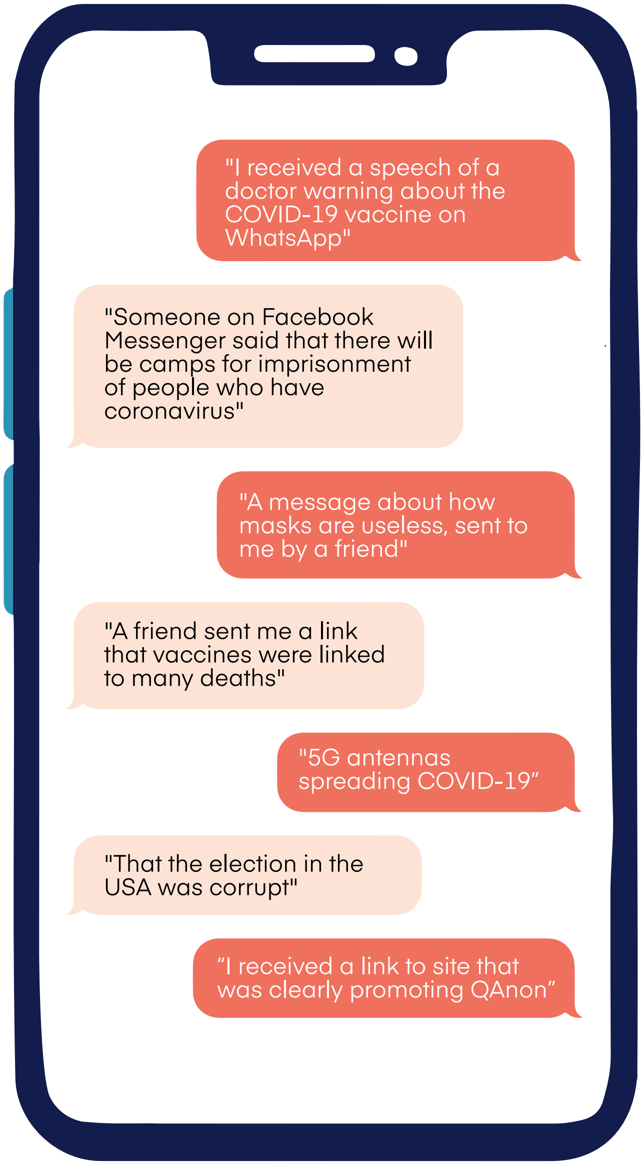
We also reviewed the relationship between respondents' frequency of receiving news through private messages and the frequency of reporting false information. Overall, 30% of respondents who said they relied on private messages as a news source reported seeing information that they immediately suspected to be false at least a few times per week, compared to 24% overall. Those who received news at least a few times per week through Telegram (51%), WeChat (51%), Instagram direct messages (33%), WhatsApp (31%), Snapchat (29%) and Facebook Messenger (28%) all reported higher levels of seeing false information at least a few times per week than the overall population.

Frequency of Online Harms through Private Messaging Apps



n = 2,044

Respondents were asked to give an example of false information that they had received on private messaging apps, which about 20% (n=509) answered. About 40% of those who provided answers mentioned messages related to COVID-19, and 10% related to the U.S. election. The other half of answers referenced various scams, phishing attempts and celebrity news. Of the 40% of answers related to COVID-19, 42% were about **misinformation related to vaccines** in particular.



"I received a speech of a doctor warning about the COVID-19 vaccine on WhatsApp"

"Someone on Facebook Messenger said that there will be camps for imprisonment of people who have coronavirus"

"A message about how masks are useless, sent to me by a friend"

"A friend sent me a link that vaccines were linked to many deaths"

"5G antennas spreading COVID-19"

"That the election in the USA was corrupt"

"I received a link to site that was clearly promoting QAnon"

Sample of survey responses

We also asked respondents how much truth they thought there was to a set of COVID-19 conspiracy theories or misinformation. Overall, 10% believed there was a great deal or some truth to at least three of the four statements (see Appendix 5).[87] Those respondents were much **more likely to receive news through private messaging:** 63% receive news through Messenger at least a few times a week, compared to 47% overall (34% more likely); and 39% receive news through WhatsApp at least a few times per week, compared to 22% overall (77% more likely). This echoes the findings from previous research that found a relationship between consuming news on social media platforms and the propensity to believe in COVID-19 conspiracy theories.[88]

Hate speech was reportedly received through private messaging apps by about one-quarter (26%) of respondents at least a few times a month. However, rates were much higher among Latin American (58%), Middle Eastern (44%), Southeast Asian (44%), Black (40%) and South Asian (32%) respondents. Harassment or bullying, and promotion or encouragement of violence, were reportedly received at least a few times a month by about 21% of respondents.

Global Approaches To Private Platform Regulation

Canada

There are several existing laws that could constrain efforts to regulate communications on private messaging apps. However, the relevance of certain laws is complicated by the fact that messaging platforms can be used for private conversations between two individuals, as well as wider broadcasts to large groups. Regardless of these challenges, there are a few laws that likely apply in the Canadian context, including, but not limited to:

- **International Covenant on Civil and Political Rights**: Canada has committed to uphold certain international standards of civil and political rights. Article 17 of the International Covenant on Civil and Political Rights provides protection against "arbitrary or unlawful interference" with one's "correspondence."

- **Canadian Charter of Rights and Freedoms:** Section 2 of the Charter protects the right to "freedom of thought, belief, opinion and expression, including freedom of the press and other media of communication." These freedoms can be curtailed "only to such reasonable limits prescribed by law as can be demonstrably justified in a free and democratic society."[89]

- **Criminal Code of Canada:** The Criminal Code makes it an indictable offence to communicate "statements, other than in private conversation, [that] wilfully promote hatred against any identifiable group."[90] There are defences, including if a person can establish that the statements communicated were true; or if they believed them to be true and the discussion was on a subject of public interest. Whether communication of hate speech in private messaging groups is "private conversation" is contextual.[91]

- **Competition Act:** Canada's *Competition Act* prohibits directly or indirectly sending electronic messages that are "false or misleading in a material respect" specifically for the purpose of promoting business interests or the supply or use of a product, including criminal responsibility if this is done knowingly or recklessly.

Several jurisdictions around the world have taken steps to regulate online platforms, as detailed in the following section. But it should first be noted that these steps have not been tailored to the specific challenges posed by private messaging apps. In short, to date, no jurisdiction has yet taken specific actions to address online harms perpetrated on private messaging apps in a way that takes a sensitive and realistic posture toward the challenges posed by encrypted messaging in particular. This creates a challenge and opportunity for Canada to build upon the actions of other jurisdictions and become a leader in this space.

European Union

In December 2020, the European Commission announced the *Digital Services Act* (DSA) and the *Digital Markets Act* (DMA). The DSA calls for more fairness, transparency and accountability for digital services' content moderation processes, including messaging services.[92]

"Hosting services" which includes private messaging apps would be required to have user-friendly mechanisms to electronically report content that users consider illegal, as well as provide notice to users if it removes or disables content, including the reasons for its decision and available redress possibilities. The law would also require annual reports outlining their content moderation activities, including the number of user reports by type of alleged illegal content, action taken and average time needed for taking action, as well as proactive measures taken as a result of the application and enforcement of their terms and conditions. When enabled by national laws, EU member states would also be able to order hosting services to remove illegal content; however, how this may apply to private messaging is not yet clear.

The DSA differentiates "hosting services" which include private messaging services from "online platforms" which exclude private messaging and groups. Specifically, "online platforms" only include services that make "information available, at the request of the recipient of the service who provided the information, to a potentially unlimited number of third parties" and may be subject to a number of additional requirements, such as risk management, data sharing and reporting of criminal offences.

In contrast, the DMA proposal is concerned with economic imbalances, unfair business practices and their negative consequences, such as weakened contestability of platform markets.[93] One proposed implication for messaging apps has been the potential to require interoperability between messaging apps. Such a requirement could necessitate the ability for messages to be sent between platforms; for example, allowing message sent from WhatsApp to be received through iMessage or Signal.[94]

France

In November 2018, the National Assembly approved a bill with the aim of preventing the spread of disinformation before a general election.[95] With this bill, legal orders may be issued to online platforms, requiring them to take "any proportional and necessary measure" to stop the "deliberate, artificial or automatic and massive" spread of false information in the three months prior to an election. Judges must determine within 48 hours whether disinformation has been distributed on a significant enough scale, and whether it has led to a disturbance of the peace, thereby compromising the results of an election.[96] The definition of online platform is broad and could capture private messaging platforms, as it includes online services that "bring together several parties with a view to the sale of a good, the supply of a service or the exchange or sharing of content, a good or a service." The same law requires platforms to put in place a visible and easily accessible mechanism for users to flag misinformation; and platforms are required to provide an annual report detailing the measures that they have taken to stop the dissemination of misinformation. The law has yet to be used in a French national election.

Germany

In June 2017, the German Federal Parliament adopted the *Network Enforcement Act* or the NetzDG, which came into effect nationwide in October 2017.[97] The law aims to more effectively reduce hate speech, criminally punishable disinformation and other harmful content on social media. Under the *Act*, social networks with at least two million members in Germany are subject to multiple obligations, which include taking down or blocking access to unlawful material within 24 hours. However, messaging applications are specifically exempt. The Canadian government has indicated that it may follow a similar approach, with the Prime Minister's mandate letter to the Minister of Heritage calling for "new regulations for social media platforms, starting with a requirement that all platforms remove illegal content, including hate speech, within 24 hours or face significant penalties."[98]

United Kingdom

In December 2020, the UK announced plans to extend the powers of the Office of Communications (Ofcom), the national communications regulator, allowing it to publish codes of practice for online platforms regarding illegal content, as well as issue fines up to 10% of companies' global revenue and impose criminal sanctions on senior platform executives for non-compliance.[99] The government's plan indicates that private messaging platforms will be in scope, as it will apply to services that "host user-generated content" and/or "facilitate public or private online interaction between service users."

The UK government's proposal states that it will use codes of practice to set out how companies can fulfil their duty of care, including measures that may be expected in the context of private communications. These measures could include safety-by-design features, such as preventing anonymous adults from contacting children. An open letter to Facebook from Australia, the UK and the U.S. in 2019 emphasized particular concern with a "single platform that would combine inaccessible messaging services with open profiles, providing unique routes for prospective offenders to identify and groom our children."[100] In March 2021, Facebook announced it was implementing changes to Instagram, preventing those over the age of 18 from sending messages to those under 18 if they don't follow them.[101]

The UK government's plan also alludes to the possibility that these new codes of practice may influence the technical design of private messaging applications. It elaborates in the following way:

> "Online services and products can be designed in a way that limits the ability of users to engage critically with online content. For example, a user journey that allows the user to forward messages to an endless number of people risks limiting the user's ability to critically assess content, and leaves them more vulnerable to engaging with misinformation and disinformation online... The safety by design framework will provide organisations with practical guidance on how to design safer online

> services and products that empower users. As part of this role, Ofcom will develop a greater understanding of how service design strengthens users' media literacy skills. This dual approach will empower adult users to keep themselves safe online and ensure companies consider the impact of their design choices on user safety."

Singapore

In October 2019, Singapore introduced the *Protection From Online Manipulation and Falsehood Act*.[102] The *Act* is aimed at preventing the electronic communication of 'false statements' in order to suppress support for and counteract the effects of false communications, while safeguarding online information systems from manipulation. Specifically, it imposes fines and jail time of up to five years for the spread of 'false statements' as determined by the government, as well as jail time of up to 10 years for the use of bots to spread false information. While senior ministers in the Singaporean government have acknowledged that private messaging platforms are an important vector of false information, the existing legislation does not address the specific challenge of encrypted messaging platforms in any clear and practical way.

Brazil

In 2020, Brazil sought to expand the powers of government and the obligations of tech companies by introducing a new legal framework in the form of a "Fake News Law," to tackle the sharing of false or deceptive content with the potential to cause individual or collective harm. The draft law proposes to mandate that all online platforms, including private messaging platforms, retain records of "broadcasted" messages forwarded by at least five users to more than 1,000 users, including the users who originated and forwarded the message, in the event that such records need to be disclosed to authorities.[103, 104] Aside from the text of this law, the congressional proceedings surrounding its framing were criticized as lacking in appropriate consultation and debate, owing in part to its introduction during the COVID-19 pandemic. The text itself has been heavily criticized for its weakening of end-to-end encryption and its vague criminal provisions.[105]

India

India's main data privacy law is the *2000 Information Technology Act*. However, in recent years there has been a flurry of activity by Indian lawmakers to replace it with the *Personal Data Protection Bill,* first introduced in 2019.[106] Since being introduced, the bill has seen numerous rounds of deliberation and consultation. Generally, it lays out compliance requirements for various types of personal data, expands individual privacy rights, creates a State regulator to protect data, and puts in place special rules for certain types of sensitive data. The bill has been criticized by some commentators as enabling the State to access the private data of citizens.[107] For example, while the bill does ostensibly provide some new protection rights for citizens, it also provides the government with sweeping powers to exempt its various agencies from respecting these rights under a broad set of circumstances. These exemptions are largely left to the discretion of the executive branch, with no significant oversight mechanisms. This has drawn criticism even from the original drafter of the bill, who said in an interview with *The Economic Times*: "The government can at any time access private data or government agency data on grounds of sovereignty or public order. This has dangerous implications."[108]

United States of America

In April 2018, the *Allow States and Victims to Fight Online Sex Trafficking Act* (FOSTA) was passed into law.[109] The law amended Section 230 of the *Communications Decency Act*. Section 230 allows internet companies to avoid liability for what they publish — or remove in good faith — on their platforms, including user-generated content. FOSTA carves out a new exception to Section 230, stating it does not apply to charges of sex trafficking, or to conduct that promotes or facilitates prostitution. The effect of this change has meant that most major platforms altered their terms and conditions to restrict the posting of sexual content; however, it has not had a significant observed impact on the content moderation of private messaging. Former President Donald Trump released an executive order in May 2020, asking regulators to redefine Section 230 more narrowly, while the current U.S. administration has suggested its replacement with new legislation; and, as it stands, the impacts of such changes on private messaging remain unknown.[110] At the time of writing, the unamended parts of Section 230 otherwise remain in effect.

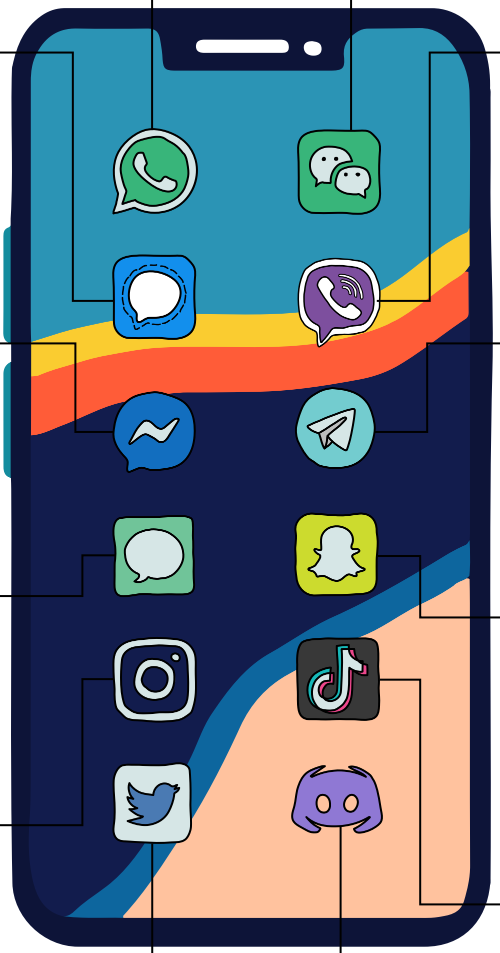
Harm Mitigation Measures

WhatsApp (Source)

% of Canadian Residents Used App in Last Year: 35%
Group Size Limit: 256
Forward Limit: Up to five times from original sender and then can only be shared in one group at a time
Can Users 'Broadcast'? To up to 256 contacts who've added the sender as a contact
End-to-end encrypted? Yes

WeChat (Source)

% of Canadian Residents Used App in Last Year: 6%
Group Size Limit: 500 (but only 100 without bank accounts attached)
Forward Limit: Up to five times from original sender then only once at a time
Can Users 'Broadcast'? Via Offical Accounts
End-to-end encrypted? No



Signal (Source 1) (Source 2)

% of Canadian Residents Used App in Last Year: 5%
Group Size Limit: 1,000
Forward Limit: No apparent limit
Can Users 'Broadcast'? Yes with no apparent limit.
End-to-end encrypted? Yes

Facebook Messenger (Source)

% of Canadian Residents Used App in Last Year: 72%
Group Size Limit: 250
Forward Limit: Only 5 people or groups at a time
Can Users 'Broadcast'? Only for approved pages to send non-promotional subscription messages
End-to-end encrypted? Only Secret Conversations, not by default

Apple iMessage (Source)

Group Size Limit: 32
Forward Limit: No apparent limit
Can Users 'Broadcast'? No
End-to-end encrypted? Yes

Instagram DMs

% of Canadian Residents Used DMs in Last Year: 33%
Group Size Limit: 32
Forward Limit: Can forward friends' stories to mutual friends, can't forward actual messages
Can Users 'Broadcast'? Yes with images/videos
End-to-end encrypted? No

Viber (Source)

% of Canadian Residents Used App in Last Year: 4%
Group Size Limit: 250
Forward Limit: No apparent limit
Can Users 'Broadcast'? Yes through Communities, but the Community 'creators' decide who gets to post
End-to-end encrypted? Yes

Telegram (Source)

% of Canadian Residents Used App in Last Year: 8%
Group Size Limit: 200,000
Forward Limit: 100
Can Users 'Broadcast'? Yes with no apparent limit
End-to-end encrypted? Only "secret chats" but not by default.

Snapchat

% of Canadian Residents Used App in Last Year: 24%
Group Size Limit: 63
Forward Limit: Can forward friends' stories to mutual friends, can't forward actual messages
Can Users 'Broadcast'? Yes with images/videos
End-to-end encrypted? Yes

TikTok DMs

% of Canadian Residents Used DMs in Last Year: 8%
Group Size Limit: No direct group chat option in TikTok
Forward Limit: Can't forward DMs
Can Users 'Broadcast'? No
End-to-end encrypted? No

Twitter DMs (Source 1) (Source 2)

% of Canadian Residents Used DMs in Last Year: 13%
Group Size Limit: 50
Forward Limit: Can't forward DMs
Can Users 'Broadcast'? No
End-to-end encrypted? No

Discord (Source)

% of Canadian Residents Used App in Last Year: 10%
Group Size Limit: 10
Forward Limit: Does not support forwarding
Can Users 'Broadcast'? Can program bots to send messages to all groups to which you have access, but no natively-implemented direct-messaging blast feature
End-to-end encrypted? No

User Content Policy & Harmful Message Reporting

WhatsApp

User Content Policy:
Prohibits content that is illegal, obscene, defamatory, threatening, intimidating, harassing, hateful, racially, or ethnically offensive, or instigates or encourages conduct that would be illegal, or otherwise inappropriate, including promoting violent crimes. Prohibits falsehoods, misrepresentations, misleading statements or "hateful" use of their app.

Harmful Message Reporting (In violation of terms of service or community guidelines): Yes

Accounts can be banned if "activity is in violation of our Terms of Service."

WeChat

User Content Policy:
Prohibits "fraudulent, false, misleading, or deceptive" content.

No content that "harms" or is "hateful"

Harmful Message Reporting (In violation of terms of service or community guidelines): Yes

Signal

User Content Policy:
Prohibits "illegal or impermissible communications such as bulk messaging, auto-messaging, and auto-dialing".

No explicit mention of 'false' or 'hateful' content.

Harmful Message Reporting (In violation of terms of service or community guidelines): Can block users, but no evident reporting function

Message requests feature was added to address spam

Viber

User Content Policy:
Prohibits "bullying", discrimination based on "ethnic or race origin" and "deceptive or fraudulent links".

Harmful Message Reporting (In violation of terms of service or community guidelines): Can report inappropriate content (spam, violence promotion, misinformation, etc.) in ads, communities, individiual posts and custom sticker packs

Facebook Messenger

User Content Policy:
(applies to Facebook and Messenger)

Prohibits sharing "unlawful, misleading, discriminatory or fraudulent" content, anything that "infringes or violates someone else's rights" or content that violates its Community Standards, which includes a detailed hate speech policy.

Harmful Message Reporting (In violation of terms of service or community guidelines): Yes

Telegram

User Content Policy:
Doesn't allow spam, content that promotes violence, or illegal pornography.

No explicit mention of 'false' or 'hateful' content.

Harmful Message Reporting (In violation of terms of service or community guidelines): In-app reporting of accounts sending spam or unwelcome private messages. Reported users then have limited accounts where they can only send messages to people saved as a contact.

"Repeated offences will result in longer periods of being blocked"

Apple iMessage

User Content Policy:
Prohibits "objectionable, offensive, unlawful, deceptive, inaccurate, or harmful content".

Harmful Message Reporting (In violation of terms of service or community guidelines): iMessages from non-contacts can be reported as junk or spam to Apple.

Snapchat

User Content Policy:
Prohibits threats, violence and harm, hate speech, harassment, deceptive practices and false information.

Harmful Message Reporting (In violation of terms of service or community guidelines): Yes

Instagram DMs

User Content Policy:
(applies to Instagram and DMs):

Prohibits "anything unlawful, misleading, or fraudulent."

Prohibits coordination of harm, promotion of violence or self-harm, hate speech, harassment and "misinformation that contributes to the risk of imminent violence or physical harm."

Harmful Message Reporting (In violation of terms of service or community guidelines): Yes

TikTok DMs

User Content Policy:
(applies to TikTok and DMs):
Prohibits hateful behavior, harassment, bullying, threats of violence and misinformation (including synthetic media)

Harmful Message Reporting (In violation of terms of service or community guidelines): Yes

Twitter DMs

User Content Policy:
(applies to Twitter and DMs):
Prohibits promotion of violence, child sexual exploitation, harassment, hateful conduct and self-harm.

Prohibits synthetic or manipulated media "that are likely to cause harm."

Prohibits and thoroughly defines abusive behavior.

Harmful Message Reporting (In violation of terms of service or community guidelines): Yes

Discord

User Content Policy:
Harrassement, violence, and hate speech are all prohibited. as well as other categories of online harmful content such as non-consensual sharing of sexual material.

Harmful Message Reporting (In violation of terms of service or community guidelines): Yes

Harm Mitigation Measures

As shown in the infographic above, platforms vary widely in terms of the degree of **measures adopted to counter disinformation and online harms.** They also vary significantly with respect to the level of detail in their terms of use, and the types of speech they prohibit.

Larger platforms, such as Facebook Messenger and Facebook-owned WhatsApp, have imposed measures like forwarding limits, often as a result of increased public scrutiny, and in an effort to pre-empt more stringent government action. Meanwhile, platforms such as Signal, which are operated by significantly smaller companies, face less public scrutiny. Small platforms may also find it advantageous to distinguish themselves from larger platforms in an effort to increase their market share by attracting users dissuaded by larger platforms' data-sharing practices.

It is possible that these competitive market dynamics and corporate strategic efforts to pre-empt state regulation play an important role in platform self-regulation. Policymakers should therefore be mindful of industry dynamics; and of how the spectre of government regulation may drive both greater moderation and censorship among larger platforms on the one hand, as well as user flight to smaller, less-stringent platforms on the other. However, a detailed study of these incentive structures is beyond the scope of this paper.

It is worth mentioning that platforms have increased self-regulation measures to remove harmful content, which appear to be at least somewhat effective. For example, in 2018, Facebook began periodically releasing *Community Standards Enforcement Reports*, in which they publish a breakdown of the harmful content proactively removed by the company during the period in question — though without providing separate information for each of their messaging products such as Messenger, Instagram direct message or WhatsApp.[111] These reports suggest increasing levels of content moderation. For example, in the spring of 2020, Facebook released one such report claiming to have removed nearly double the number of posts, including hate speech, compared to the previous quarter; and a falling proportion of hate speech detected by user reports. However, others have pointed to the uneven removal of certain harmful content; for instance, images of sexual violence may be flagged and deleted more frequently compared to hateful or harassing messages due to technical factors.[112] Further, such efforts have been criticized as lacking transparency, making it difficult for independent observers to get a clear sense of the ways in which Facebook is undertaking this content moderation; and obscuring the degree to which genuine progress is being made in this area, which we elaborate on in the following section.[113]

User Reporting Mechanisms

One approach to regulating private messaging applications is to require tools that enable users to report harmful content. These reports can be used as a basis for further action by the platform itself, such as banning users or closing messages for content that violates terms of service. Mandating user reporting and associated appeal mechanisms has been advanced in France, the UK and the EU. Nearly all private messaging platforms reviewed, with the exception of Signal, have user reporting mechanisms, many of which are easily accessible directly through the messaging interfaces.

For example, WhatsApp's user reporting feature requires that a copy of recently received messages be shared with WhatsApp.[114] WhatsApp also has a provision in its terms of service different from Facebook and Instagram that does not allow for "publishing falsehoods, misrepresentations, or misleading statements." In a white paper in 2019, WhatsApp indicated that it banned two million global accounts per month for misinformation, of which 95% were for "abnormal WhatsApp behaviour" detected automatically, while 5% were through user reports.[115]

Other platforms, like Telegram, allow certain members of a group to have 'administrative' powers, including deleting messages by other members of the group.[116] However, there is no guarantee that this power will be used to enforce desirable community standards from a public policy perspective.

There is limited public information available on the use and impact of these user reporting mechanisms, particularly at a country-specific level. Of the private messaging platforms, Snapchat provides the most detailed information on user reporting in Canada.[117] For example, in the six-month period from January to June 2020, a total of 247,864 content reports were received by Snapchat in Canada. Based on an approximate active user population of 10.4 million Canadians, this represents one report from 2.4% of users over a six-month period. Of the 247,864 content reports, 62,745 reports (25%) and 36,767 unique accounts (15%) were enforced. The majority of the reports (55%) were for sexually explicit content; 13% were impersonation; 11% were threatening violence or harm; 9% were harassment or bullying; and 3% were hate speech. There were also 1,041 account deletions for child sexual exploitation and abuse, of which 70% were proactively identified through content scanning for known child sexual abuse material.

Labels and Limits on Message Forwarding

Some messaging platforms have tried to slow the spread of disinformation and spam by placing limits on the number of times that a message can be forwarded between users. WhatsApp has introduced this feature, which involves slowing the rate at which a message can be shared if that message has already

Canadian User Reporting on Private Messaging Platforms

22% have reported someone for sending illegal, hateful or harassing content

26% have reported an account for being automated or sending spam

37% have found reporting to be very effective; 37% moderately effective; 22% not effective been shared five times.[118] This limit then only allows for the message to be shared once at a time, and also adds a warning label of "Forwarded many times" as a means of informing the recipient of the message.

Similarly, Facebook Messenger only allows a message to be forwarded to five users or groups at a time.[119] Other platforms, such as Signal and Viber, do not yet have an apparent limit on message forwarding.

WhatsApp has suggested that such measures have helped slow disinformation, claiming a 70% drop in the transmission of messages that had already been forwarded five times after the change was introduced. The company credits this, along with a similar measure introduced in 2018, with the 25% overall reduction in message forwarding on its platform over the past two years. However, researchers from the University of Minas Gerais and MIT argue that, although such measures are somewhat effective in slowing the spread of disinformation, they do not completely stop the dissemination of such content. Further, there remains potential for professionally orchestrated teams to exploit such measures through networked groups in ways to have disinformation still go viral.[120]

Limits on Group Messages

Platforms can potentially slow the spread of harmful content through private messaging on their platforms by placing a limit on the number of users allowed in a particular group. The most widely-used messaging apps like WhatsApp and Facebook Messenger have imposed group size limits of 256 and 250, respectively.[121] [122] Other platforms like Instagram and Snapchat have much lower limits of 32 and 63, respectively. However, other messaging apps that are increasingly growing in users such as Signal and Telegram, have much larger group limits, with Signal having a maximum limit of 1,000 people in group conversations[123] and Telegram's limit up to 200,000[124] at the time of writing. Signal has also expressed a desire to find technical solutions to make group sizes much larger.[125]

Compared to message forwarding, there has not been as much scrutiny or publicly available evaluation of the effect of group size limits on online harms. For example, WhatsApp increased its group size limit in 2016 from 100 to 256 with seemingly little scrutiny.

In our survey, **17%** of respondents in Canada reported being part of an **online group message with more than 50 people.** A 2019 study found that 34% of Canadian WhatsApp users were part of a group with people that they didn't know, while 8% were part of a group focused on discussing news with like-minded people. The same study found members of messaging groups focused on news in other countries were more likely to say that they trust the news they get from social media.[126]

Some advocates have also suggested changes to the design of group messages to add friction to the spread of viral content, such as: requiring users to opt-in to group messages; limits on the ability to broadcast messages (send identical messages to multiple contacts); and having message groups over a certain size be unencrypted to allow for content moderation.[127]

Enabling Information Verification

Platforms can try to make it easier for users to verify the reliability of information. There are many different ways of doing so that can be applied to text, audio, photo or video. In our survey, **43%** of respondents in Canada said they had **ever tried to fact-check a message** they had received about the news or politics using another source. Of those who had taken this step, 72% found it to be effective (ranked 7-9 on a scale from 1-9 least to most effective).

WhatsApp began testing a new feature in seven countries (Brazil, Italy, Ireland, Mexico, Spain, the UK and the U.S.), which is now available in Canada, that allows users to click on a magnifying glass icon next to a highly-forwarded message that directs users to a Google search of the message content. Such a measure could have the effect of encouraging fact-checking, though there has not yet been an independent evaluation of the new feature's effectiveness. The amount of time needed for fact-checking, compared to the speed with which viral disinformation can propagate through private messaging platforms, has also been raised as a concern with respect to the efficacy of these tools.

Aviv Ovadya, Founder of the Thoughtful Technology Project, has emphasized that these efforts to prompt users to verify information are not yet suitably adapted to the challenge of false information on private messaging. For example, the WhatsApp prompt to verify a piece of information on Google fails to capture the fact that the Google search engine is designed to provide results based primarily on relevance to the query's keywords, rather than the authoritative quality of the content. The design of an AI-driven search engine that emphasizes authoritativeness in its results could be a more suitable tool for verification purposes on private messaging. One current effort in this regard is MediaSmart's custom search engine, which allows people to search eight global fact-checking sites at once. However, as no tool has yet been designed for this purpose at scale including searching audiovisual media, there is a need for further innovation in this area.[128]

The instant messaging platform LINE, the most used messaging app in Taiwan, has implemented a novel method of empowering users to verify the news that they encounter by dialoguing with a chatbot. Users can send links or statements to the bot, named *LINE Fact Checker,* which automatically analyzes and verifies the content against a database of previously verified stories provided by a small group of third parties. This allows the platform to avoid having to read users' encrypted messages in order to provide in-app information verification services, while leveraging the pooled efforts of different fact-checking services. It also allows the platform to collect data on the types of queries the bot is exploring, providing insight into the information environment without identifying anonymous users.[129]

WhatsApp and the Poynter Institute launched a similar international chatbot in May 2020 specific to COVID-19 misinformation, while CBC developed a Facebook Messenger chat bot for users to learn about misinformation ahead of the 2019 federal election.

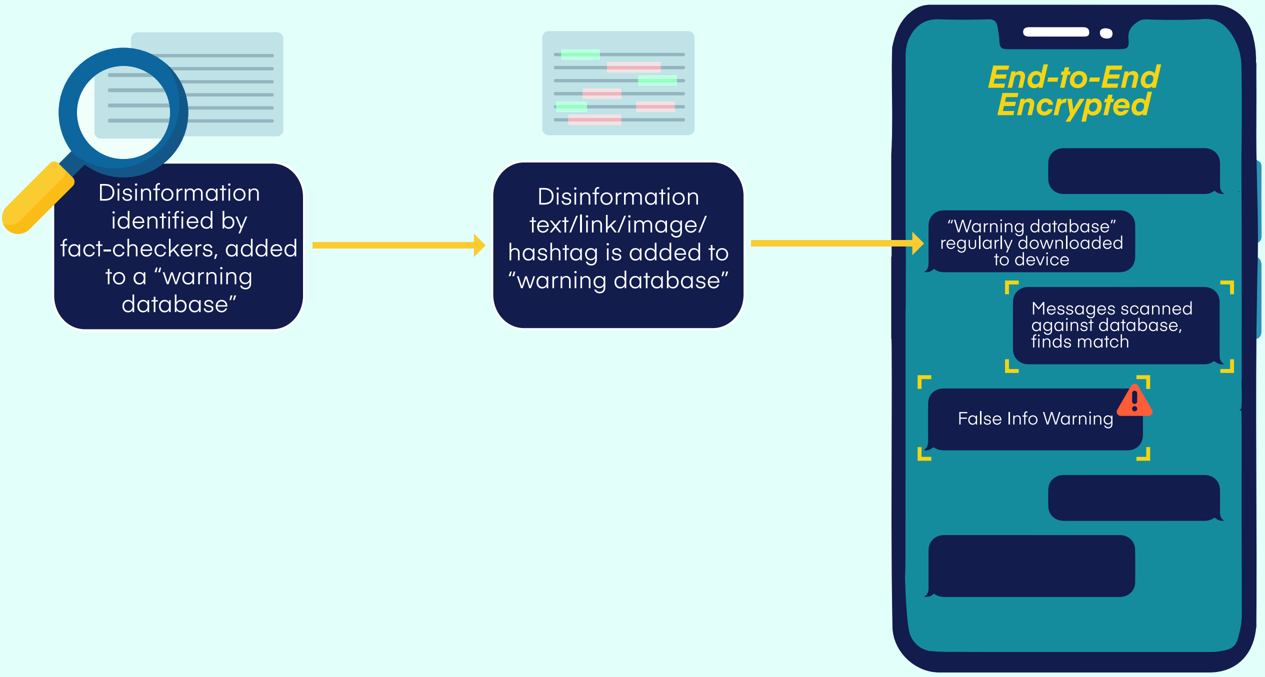
Client-Side Scanning

One approach to helping users identify disinformation that they encounter on open platforms is for third-party fact-checkers to monitor posts, and then label disinformation as false or misleading. Twitter and Facebook recently began using this method to label disputed or misleading information about the U.S. election and COVID-19,[130] and to direct users toward a curated page containing trusted sources.[131] Such measures have not yet been extended to direct or group messages such as Facebook Messenger, Instagram Direct Messages or Twitter Direct Messages, even though each of these messaging services is, by default, unencrypted.

This type of labelling becomes more challenging in the case of end-to-end encrypted messages, since even the platform operator itself is unable to access the content of messages. In this case, a proposed approach is increasingly gaining traction, known as *client-side scanning* or *client-side context*. Under this approach, messages received are scanned against a database of previously identified harmful content downloaded to the user's device. Content matched with the database could receive a warning label, or in some proposals even be disabled from being shared. Advocates of the approach argue that the content of the message remains private, preserving the end-to-end encryption architecture while still monitoring for harmful or illegal content. However, others have pointed out technical[132] and legal[133] challenges associated with this approach, with some arguing that it still breaks the 'essence of encryption'.

The approach also raises a wide range of social, ethical, technical and political questions. For instance, it is unclear whether the technology will slow down the app's or device's

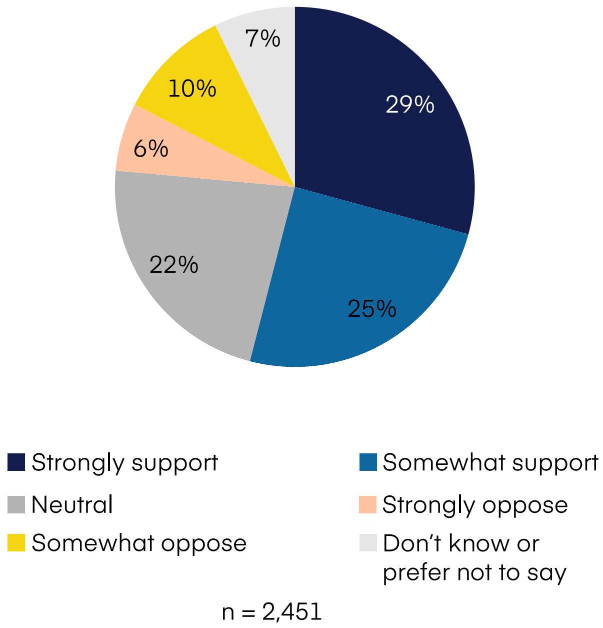
How Client-Side Scanning Might Work


Disinformation identified by fact-checkers, added to a "warning database"
Disinformation text/link/image/ hashtag is added to "warning database"
End-to-End Encrypted
"Warning database" regularly downloaded to device
Messages scanned against database, finds match
False Info Warning

performance, as regular updates to the 'harmful content database' are required. The approach raises further questions regarding who determines what is 'harmful' content. Such technologies have the potential to be used by autocratic regimes to suppress free expression and political dissent, and could be extended to include the censorship of particular types of speech. There are also issues of false positives and the misidentification of non-harmful content.

In our survey, we reminded respondents that social media platforms, like Facebook and Twitter, had started adding warning labels to some public posts that contain false information about news or current events, as determined by independent fact-checkers. We then asked about their level of support if the same approach of adding warning labels was applied to private messaging apps when false information is sent. Overall, **54% of respondents were supportive**, with 22% neutral and 16% opposed.

Support for Disinformation Warning Labels on Private Messages



Some platforms are also currently developing ways to automate the scanning of harmful online materials through artificial intelligence, that could be accompanied by warning labels or forwarding limits. For instance, since 2018, WhatsApp's Suspicious Links Detection feature has been automatically alerting users with a label if a link embedded in a message appears malicious by scanning for suspicious characters. Facebook, Instagram, TikTok and Twitter have all experimented with warning users before they publicly post harassing or inappropriate comments, though none has yet to use a similar approach in private messaging.[134] Facebook[135] and Google[136] have invested resources in developing tools to automatically detect misleading visual media using this approach, such as deepfakes. On an encrypted platform, it is possible that these automated methods could be implemented using client-side scanning. However, there are several challenges associated with this proposed method:

1. ***Technical Challenges:*** AI tools for identifying misleading media will only be effective if they can be trained against large volumes of existing misleading media, which would also need to be similar in character to what is currently being shared on the messaging platforms. If this fails to happen, the AI-based methods of verification may quickly become obsolete. As a recent example, one commentator demonstrated that a commercial manipulated video detection tool failed to identify recent deepfakes of actor Tom Cruise.[137] Similarly, implementation of AI-based methods on encrypted messaging platforms is subject to all of the technical challenges described previously.

2. ***Ethical Challenges:*** AI models can make choices that would appear inaccurate or otherwise unacceptable to a human decision-maker.[138] Algorithms trained to flag or delete certain types of content can inadvertently create harm. For example, an AI error led to the deletion by YouTube of hundreds of videos used by human rights groups to document war crimes in Syria.[139]

3. ***Free Expression Challenges:*** While misleading visual media is often used for online abuse or the spread disinformation, it is also used for more benign purposes, such as satire. This means that deepfake detection needs to be considered in the context of broader questions surrounding disinformation, such as what constitutes useful satire as opposed to harmful forgery, and the appropriate role of labelling.

4. ***Communications and Credibility:*** Once a piece of misleading media has been identified, the best way to communicate that information needs to be determined. In order to help dispel disinformation, detection strategies need to be accompanied by efforts to improve communications. Once detected, labelling something as mis- or disinformation needs to address the needs of different demographics, which could make detection more likely to be accepted as credible. Moreover, tagging some media as misleading may, paradoxically, cause some people to be more interested in consuming it.[140]

Legal Obligations and Codes of Conduct

Platforms are generally protected from accountability for the content they host due to laws such as Section 230 of the *Communications Decency Act* in the U.S., and the European Union's *E-Commerce Directive* in Europe. In Canada, some scholars have argued that the updated USMCA trade agreement extends similar protection to platforms in Canada.[141] But there is a growing movement toward requiring platforms to be legally responsible for content in order to help mitigate online harms. For example, as described in more depth earlier, Germany adopted its *Network Enforcement Act* in 2017, requiring large platforms to identify and remove illegal content in a timely manner. Similarly, the UK government is in the process of creating a 'duty of care' model, as described in a recent white paper; while the EU will require large platforms to identity and mitigate risks to fundamental rights.[142]

In Canada, government actions to combat disinformation include the *Christchurch Call,* a joint initiative including over 50 national governments and technology companies that have committed to preventing terrorist and violent extremist content on the internet.[143] In 2019, Canada also unveiled the *Digital Charter,* in which its second principle describes 'safety and security' as a requirement for Canadians online; while its eighth principle specifically describes the government's role in protecting against 'online threats and disinformation' while defending freedom of expression.[144] Finally, the Canada Declaration on Electoral Integrity Online also outlined commitments of signatory social media platforms to intensify efforts to combat disinformation, including removing fake accounts and inauthentic content, promoting transparency for Canadians about efforts to safeguard the internet ecosystem, and assisting users to better understand the sources of information that they see.[145]

In its 2021 report, the Canadian Commission on Democratic Expression provided six recommendations for how Canada can respond to online hate speech and other online harms.[146] One of the recommendations was the creation of a new legal standard, wherein Parliament enacts a Duty to Act Responsibly, in turn placing affirmative responsibilities upon platforms.[147] To oversee and enforce the Duty to Act Responsibly, the Commission recommends the creation of a new regulatory body that, granted with legislative authority, can produce and enforce a Code of Conduct by which platforms must abide, in order to meet their Duty to Act Responsibly.

The principal advantage of government-driven approaches is democratic and sovereign accountability. However, in terms of disadvantages, government initiatives can fail to stay ahead of the speed of technical change and capture technical nuance available to private firms, which can restrict the effectiveness and enforceability of a policy measure.[148] One possible example of this shortfall within a related policy area is the right to explanation provided under the EU's *General Data Protection Regulation*, which has been criticized as failing to capture technical realities associated with contemporary data-processing methods.[149]

Platform Transparency

Private messaging platforms can enable a better understanding of their role in online harms, and increase transparency by making available more information and data for civil society and researchers to review. Likewise, governments can ensure transparency and accountability by requiring and auditing this information, as is contemplated by the EU's *Digital Services Act*, which has proposed transparency reporting obligations for large platform service providers, including the publishing of annual reports detailing their content moderation activities.

Private messaging platforms could, for example, be required to regularly publish information regarding their content moderation activities. This could include:

- The amount of illegal and/or harmful content reported and enforced during a specified time period, in specified categories of reasons for enforcement;
- Time taken by the platform to review and enforce user reports, such as median time or proportion longer than specified time intervals;
- Use of automation for content moderation purposes, including information on outcomes, accuracy and implemented safeguards;
- The amount of illegal and/or harmful content removed by the platform during a specified time period, in specified categories of reasons for removal;
- The amount of user suspensions or bans, in specified categories of reasons for enforcement; and
- Measures and effectiveness of measures to slow the spread of disinformation, such as group sizes, message forwarding and message labels.

Several of the researchers consulted for this report noted that researching disinformation on private messaging apps can be methodologically and resource-intensive to undertake in an ethical manner.[150] More researchers and resources are needed that can apply various qualitative methods to provide a richer and in-depth understanding of the trends, behaviours, complexities and nuances exhibited on private messaging platforms in Canada. It was also noted that large platforms, such as Facebook, that operate several different types of services often do not break down their publicly available metrics in a way that differentiates private messaging from public posts. Greater platform transparency can also assist in providing a sense of scale to more qualitative efforts.

Understanding Root Causes of Disinformation and Promoting Digital Literacy

Another approach to curbing the harms of disinformation is to promote digital literacy. Digital literacy is a broad skillset, allowing users of digital technology to better assess the quality of the information that they come across online.[151] Research has shown that media literacy education can improve people's ability to identify misleading information.[152] Recognizing the growing harm of online mis- and disinformation, some governments[153] and education institutions[154] are already investing in media literacy programs, in some cases with the collaboration of large technology companies.[155] In 2019, the Government of Canada invested $7 million in digital, news and civic literacy programming as part of efforts to combat disinformation; and the 2019 federal Budget committed an additional $19.4 million over four years toward research and policy development on online disinformation in Canada.[156] Several media literacy organizations in Canada noted that they are in the process of developing new efforts specific to the risks of disinformation on private messaging.

Other ways of empowering users to verify the information that they receive could involve leveraging the capabilities of the global fact-checking community. For example, the platform WeVerify has organized a large, crowdsourced database of previously identified manipulated media, such as photos and videos, in which blockchain technology is used to identify known fakes. This is meant to encourage a global collaborative approach to detection, which they call *participatory verification*. The need for such accessibility is made apparent in a recent study on the state of technology in journalism, which showed that while almost three-quarters of journalists use social media to find stories, only 11% use media verification tools.[157] Governments could lend support and resources to encourage the growth of these grassroots verification resources.

While these efforts may help promote greater resilience to the threats, others have argued for a focus on the source of the current public policy issue at hand, pointing to the monopolization of big tech platforms and society’s reliance on these forces for technical solutions to problems that have emerged from their technologies — rather than on public interest technology solutions, or on the generation of alternative models of platform governance.

Future Challenges

This report has outlined the challenges of mitigating online harms on private messaging platforms, as distinguished from the challenges of regulating public online spaces. However, there are some digital communication platforms that do not fit neatly into these two categories. Just as mitigating the harms of private messaging platforms requires different strategies than would be used in public online spaces, it is possible that other kinds of platforms will pose challenges that undermine the strategies suited to private messaging. For example, there have recently been a proliferation of audio-based messaging, most notably through the new app Clubhouse. These apps capture the 'private gathering space' aspect of messaging platforms, but their lack of text-based communication may require changes of strategy or technological innovation to moderate effectively. Likewise, gaming-based gathering spaces, such as VRChat, can create new forms of online harms not possible on text-based messaging platforms. As such, it is not likely that policymakers will be able to simply adapt methods that worked for text-based platforms to address new harms occurring on audio- and video-based ones.

Moreover, the growing presence of decentralized messaging apps also requires careful consideration if governments proceed with the regulation of private messaging apps. This type of messaging app uses decentralized protocols, with many relying on blockchain technology.[158] There are several such messaging apps currently on the market, including Dust, which claims to keep no record of messages sent or received on servers.[159] This is a shift away from the centralized nature of traditional apps, where the data is hosted and managed on company servers. As there is no centralized control over data in decentralized apps, challenges may arise over how these spaces can be effectively moderated or regulated. For instance, a distributed, peer-to-peer infrastructure will make content removal difficult as the data will only be saved locally on the sender's and recipient's devices.[160]

Finally, some have suggested that innovative encryption schemes could enable a more active role for platforms in moderating encrypted messaging without compromising user privacy. Most notably, certain proposed forms of encryption could allow platforms or regulators to apply machine learning to the content of user messages while preserving the secrecy of any individual message.[161] While no fully homomorphic encryption scheme has yet been developed that is suitable for the commercial purposes of private messaging companies, this is an ongoing area of research among academic and industry experts in cryptography.

Conclusion

The regulation of private expression involves some of the most complex challenges that governments face today. The inherent challenges of maintaining the privacy and freedom of expression of Canadians while mitigating real harms cannot be overstated. As governments have struggled to combat disinformation, especially during the COVID-19 pandemic, human rights groups have raised concerns regarding the increasing use of disinformation laws by governments to curtail free expression, independent media outlets and political opponents.[162] Cambodia, for instance, has recently expanded its campaign against fake news by adding WhatsApp and Telegram to its list of government-monitored platforms,[163] a concerning development considering the country's documented use of 'fake news' charges to arrest government critics.[164] Against this backdrop, it is clear that this is a sensitive policy area that requires deliberate commitment to privacy rights if public buy-in is to be attained.

We should also acknowledge that technology-driven harms — and the efforts of platforms or governments to resolve them — exist in the content of rapid technological change.[165] Platform companies, groups of employees within these companies, platform influencers and would-be perpetrators of harm are also engaged in a constant struggle to out-innovate each other. This evolutionary dynamic results in a challenging and rapidly changing space for state actors, for whom the razor of technological innovation is largely unavailable at a competitive timescale, and who must therefore rely on the often blunt instrument of public policy to address these harms.

The 'duty of care' model being advanced by other jurisdictions, such as the UK and by the Canadian Commission on Democratic Expression, is a significant global trend for policymakers to consider as a way to advance democratic governance.[166] However, the limitations of this approach to date must also be duly acknowledged. For example, these approaches have tended to place theoretical duties on private platforms to mitigate harms, which may be difficult to meaningfully carry out or practicably enforce where there are inherent trade-offs between harms. This has been the case most notably with regulatory efforts to address harms on platforms using end-to-end encryption.

Considering these complex challenges and the underdeveloped, but growing, understanding of how harms are manifesting in Canada on private messaging platforms, this report recommends three next steps for the Government of Canada:

1. **Invest in research and innovation specific to online harms on private messaging platforms in Canada**

   Through our research, survey, interviews and workshop, we learned that experts and communities are increasingly concerned about the harm propagated through private messaging platforms, particularly in light of COVID-19. However, we also heard that there was often insufficient evidence available to guide mitigation efforts. The evidence from our own survey scratches the surface of the complexities of the Canadian context regarding online harmful content through private messaging, particularly given the unique information ecosystems of many diaspora communities that rely heavily on specific messaging apps. The Government of Canada can continue to play a supporting role in sponsoring dedicated research to gain a fuller understanding of the Canadian context.

2. **Require transparency from large online platforms to better understand online harms through private messaging**

   Canada should also consider following the lead of the EU in mandating greater transparency from online platforms with a significant number of users in Canada. Too much of what we currently understand about the core parts of our information ecosystem are from voluntary unaudited disclosures from the platforms, such as account and post takedown numbers. Specific to private messaging, much of the voluntary disclosure from large platforms is also not specific to their private messaging functions and lacks country-specific reporting. Compelling regular and independently audited information and data that can help guide and evaluate the effectiveness of harm mitigation efforts would be a positive use of the power of the state. This could include information specific to Canada on harmful content user reporting, enforcement and automated moderation, as well as measures to slow the spread of disinformation, such as group sizes, message forwarding and message labels.

3. **Make investments in policy-informed digital literacy efforts that build resilience to disinformation through private messaging platforms**

We heard and learned about the unique challenges of private messaging disinformation, given the intimate nature of the platforms and increased levels of trust in the sender. The Government of Canada and many others have invested significantly over the last two years in digital and civic literacy efforts. Increasing the understanding of what literacy efforts work, and with which populations, with a specific focus on how to build resilience to disinformation in the unique context of private messaging platforms, should be a priority. Public investments in public interest technology tools that advance digital and media literacy, such as multilingual rapid fact-checking and contextualization search engines, should also be considered.

The best media literacy approaches work hand-in-hand with approaches that incorporate responsible technology design and proactive public policy. Media literacy efforts are more likely to help reduce mis- and disinformation when it is easy for private messaging app users to report harmful content, or to check it against authoritative information; when private messaging platforms have positive obligations to build attributes into their products to limit harms; when private messaging platforms have transparency requirements about the possible extent of misinformation on their platforms; and when the media ecosystem is serving up factual journalistic content, from multiple trusted sources. These can all work together to build a resilient, vibrant and cohesive society and democracy in Canada.

About the Authors

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

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

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

Stephanie Tran is an experienced researcher with over five years of experience analyzing public policy and human rights issues related to digital technologies, with past experience working for the Citizen Lab, Amnesty International Canada, the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA) and more. She is a trained computer programmer, having earned a Diploma in Computer Programming from Seneca College. She also holds a dual degree Master of Public Policy (Digital, New Technology and Public Affairs Policy stream) from Sciences Po in Paris, and a Master of Global Affairs from the University of Toronto. She earned her BA degree from the University of Toronto specializing in Peace, Conflict and Justice.

Appendix

Appendix 1: Stakeholder Workshop Participants

The varied perspectives of the participants greatly informed this report; however, the statements and recommendations are solely those of the authors.

Aengus Bridgman	PhD Researcher, McGill University
Akaash Maharaj	CEO, Mosaic Institute
Aviv Ovadya	Founder and Researcher, Thoughtful Technology Project
Caio Machado	PhD Researcher, University of Oxford
Chris Beall	Policy Lead, Centre for International Governance Innovation (CIGI)
Chris Tenove	Postdoctoral Fellow, Political Science, University of British Columbia
Elizabeth Dubois	Associate Professor, University of Ottawa
Fabrício Benevenuto	Associate Professor, Universidade Federal de Minas Gerais (Brazil)
Farhaan Ladhani	CEO, Digital Public Square
Fenwick McKelvey	Associate Professor, Concordia University
Jennifer Wolowic	Project Manager, Strengthening Canadian Democracy, ↱SFU's Morris J. Wosk Centre for Dialogue
John Beebe	Founder, The Democratic Engagement Exchange
Jonathan Lee	Director of Global Public Policy, WhatsApp
Kathryn Ann Hill	Executive Director, MediaSmarts
Kiran Garimella	Postdoctoral Hammer Fellow, MIT Institute for Data, Systems and Society
Marcus Kolga	Senior Fellow, Macdonald-Laurier Institute and founder of disinfowatch.org
Maryam Faisal	Project Manager, Council of Agencies Serving South Asians
Michele Austin	Head of Public Policy, Twitter Canada
Mohammed Hashim	Executive Director, Canadian Race Relations Foundation
Natalie Turvey	President & Executive Director, Canadian Journalism Foundation
Priyanjana Bengani	Senior Research Fellow, Tow Center for Digital Journalism
Rafael Evangelista	Research Professor, Unicamp (Brazil)
Ryan Wai On Chan	Project Lead, Online Hate & Social Media, Chinese Canadian National Council for Social Justice
Sahar Massachi	Fellow, Berkman Klein Center for Internet & Society, Harvard University
Shireen Salti	Executive Director, Canadian Arab Institute
Sonja Solomun	Research Director, Centre for Media, Technology and Democracy, McGill University

Appendix 2: Stakeholder Workshop Discussion Questions

The following pre-determined questions were asked to workshop participants:

1. What role are private messaging apps playing in the spread of mis- and disinformation or other online harms in Canada?
2. What more do we need to research and understand?
3. What could, and should, be done to mitigate disinformation and other online harms from private messaging apps? What roles are there for government, industry, journalism and civil society?

Appendix 3: National Survey Questionnaire

Q1) How much truth do you think there is to each of the following claims about COVID-19?

- **A great deal**
- **Some**
- **Very little**
- **None**
- **Don't know or prefer not to say**

a. Bill Gates is using the coronavirus to push a vaccine with a microchip capable of tracking people
b. The coronavirus escaped from a lab in Wuhan, China
c. Gargling saltwater helps prevent the coronavirus
d. The pharmaceutical industry is involved in the spread of the coronavirus

Q2) Which of the following do you use to stay up-to-date with the news or current events? (select all that apply)

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

Q3) Have you used any of the following messaging apps in the last year?

a. WhatsApp
b. Facebook Messenger
c. WeChat/Weixin
d. Telegram
e. Signal
f. Snapchat
g. Direct messages on Instagram
h. [Viber/imo/Weibo]*
i. [LINE/Discord/Clubhouse]*
j. [QQ/Direct messages on Twitter/Direct messages on TikTok]*

** Survey respondents split into three and each asked one of each*
[if No to all, Q4-8 skipped]

Q4) About how often would you say you receive messages from your friends, family or other contacts about the news or current events on the following apps? *[chart with apps selected by respondent in Q3]*

a. Every day
b. A few times a week
c. A few times a month
d. A few times a year
e. Never
f. Don't know or prefer not to say

Q5) *[skipped if Never to all in Q5]* Compared to the following, how much trust would you say you have that the information you are sent about news or current events on all the messaging apps you use is accurate and authentic? *[options skipped if not selected in Q2]*

- **TV news**
- **News websites**
- **Social media feeds, like Facebook or Twitter**

a. More trust
b. About the same level of trust
c. Less trust
d. Don't know or prefer not to say

Q6A) Thinking about all the messaging apps you use, how often do you think you receive messages, including links, images or videos, that contain what you would consider:

i. information about the news or current events that you immediately suspect to be false
ii. information about the news or current events that you believe to be true and later find out is at least partly false
iii. hate speech that wilfully promotes hatred against an identifiable group
iv. harassment or bullying
v. a scam (e.g., phishing to provide personal information or to download malware)
vi. promoting or encouraging violence

a. Every day
b. A few times a week
c. A few times a month
d. A few times a year
e. Never
f. Don't know or prefer not to say

Q6B) [if more than Never to i. or ii.] Could you tell us more about a recent example of false information that you received? What was it about? Who sent it to you? What made you believe it was false? *[text box]*

Q7A) Have you ever done any of the following? (Select all that apply)

a. Been part of an online group message with more than 50 people
b. Fact-checked a message you received about the news or politics using another source
c. Reported someone to a messaging platform for sending illegal, hateful or harassing content
d. Reported an account to a messaging platform for being automated or sending spam
e. None of the above

Q7B) [if b through d selected] How would you rate the effectiveness of these actions, where 1 is not at all effective and 9 is very effective?

Q8) Social media platforms, like Facebook and Twitter, have started adding warning labels to some posts that contain false information about news or current events, as determined by independent fact-checkers. Some within the tech industry have suggested the same approach of adding warning labels could be applied to messaging apps when false information is sent. Would you say you would be strongly supportive, somewhat supportive, are neutral, somewhat oppose or strongly oppose of sort of fact-checking approach being applied to your private messages?

a. Strongly support
b. Somewhat support
c. Neutral
d. Somewhat oppose
e. Strongly oppose

Demographic Questions

D1) Where do you currently live? *[Province/territory list]*

D2) What is your gender?
a. Woman
b. Man
c. Non-binary/third gender
d. Prefer to self-describe [*text box*]
e. Prefer not to say

D3) How old are you? [*Drop down*]

D4) What is your personal income, before taxes and deductions, in 2020?
a. No income
b. Less than $30,000
c. $30,000 to less than $50,000
d. $50,000 to less than $70,000
e. $70,000 to less than $100,000
f. $100,000 to less than $150,000
g. $150,000 or more
h. Don't know or prefer not to say

D5) What is the highest level of education you have completed?
a. No certificate, diploma or degree
b. High school diploma or equivalency certificate
c. Certificate of Apprenticeship or Certificate of Qualification
d. College, CEGEP or other certificate or diploma
e. University degree
f. Prefer not to say

D6) Do you self-identify as: (select all that apply and/or specify, if applicable)
a. White
b. Indigenous, that is First Nations (Status/Non-Status), Metis or Inuit
c. East Asian (e.g., Chinese, Korean, Japanese, etc.)
d. South Asian (e.g., East Indian, Pakistani, Sri Lankan, etc.)
e. Southeast Asian (e.g., Filipino, Thai, Vietnamese, etc.)
f. Black
g. Latin American
h. Middle Eastern (e.g., West Asian, Iranian, Afghan, Arab, etc.)
i. Another option – please specify
j. Prefer not to say

D7) How long have you lived in Canada?
a. Born in Canada
b. Less than 2 years
c. 2 to 10 years
d. More than 10 years
e. Prefer not to say

D8) What language do you speak most often at home?

Appendix 4: National Survey Sample Demographics

BC: 13%
Alberta: 12%
Saskatchewan: 3%
Manitoba: 4%
Ontario: 38%
Québec: 23%
New Brunswick: 2%
Nova Scotia: 3%
Prince Edward Island: 1%
Newfoundland and Labrador: 1%
Territories: 0%

Woman: 51%
Man: 49%

Age 16-29: 22%
Age 30-44: 25%
Age 40-59: 24%
Age 60 and over: 30%

High school or less: 31%
College education: 37%
University education: 32%

White: 77%
East Asian: 8%
South Asian: 4%
Southeast Asian: 3%
Black: 3%
Indigenous: 2%
Middle Eastern: 2%
Latin American: 1%

Born in Canada: 74%
In Canada less than two years: 2%
In Canada 2-10 years: 5%
In Canada more than 10 years: 19%

Language Most Often Spoken at Home:
English: 73%
French: 19%
Cantonese: 2%
Mandarin: 1%
Punjabi: 1%
Other Answers: 4%

Appendix 5: Respondents' Belief in COVID-19 Misinformation

	A great deal of truth	Some truth	Very little truth	No truth	Don't know or prefer not to say
The coronavirus escaped from a lab in Wuhan, China	14%	20%	18%	35%	13%
The pharmaceutical industry is involved in the spread of the coronavirus	5%	11%	16%	60%	9%
Bill Gates is using the coronavirus to push a vaccine with a microchip capable of tracking people	5%	9%	12%	63%	12%
Gargling saltwater helps prevent the coronavirus	2%	7%	14%	68%	9%

References

[1] This figure apparently rose to $22 billion. See McLaughlin, T. (2018, 12). How WhatsApp Fuels Fake News and Violence in India. *Wired;* Newman, J. (2014, February 20). Facebook's WhatsApp Acquisition Explained. Time. https://time.com/8806/facebooks-whatsapp-acquisition-explained/

[2] Tankovska, H. (2021, January 28). Daily active users of WhatsApp Status 2019. Statista. https://www.statista.com/statistics/730306/whatsapp-status-dau/

[3] Digital 2020. (2020). We Are Social, Hootsuite. https://wearesocial-net.s3-eu-west-1.amazonaws.com/wp-content/uploads/common/reports/digital-2020/digital-2020-global.pdf; Digital 2020: Canada. (2020). *We Are Social, Hootsuite.* https://wearesocial.com/ca/digital-2020-canada/

[4] Rody, B. (2019, July 8). New Canadians consume mobile media like no others: Study. *Media in Canada.* https://mediaincanada.com/2019/07/08/new-canadians-consume-mobile-media-like-no-others-study/

[5] Anderson, J., & Rainie, L. (2017, October 19). The Future of Truth and Misinformation Online. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/

Sonnemaker, T. (2020, October 29). Facebook reported a decline of 2 million daily active users in the US and Canada. *Insider.* https://www.businessinsider.com/facebook-decline-2-million-daily-users-us-canada-q3-earnings-2020-10

[6] Zuckerberg, M. (2021, March 12). A Privacy-Focused Vision for Social Networking. *Facebook*. https://www.facebook.com/notes/2420600258234172/

[7] Schultz, A. & Parikh, J. (2020, March 24). Keeping Our Services Stable and Reliable During the COVID-19 Outbreak. *Facebook*. https://about.fb.com/news/2020/03/keeping-our-apps-stable-during-covid-19/

[8] Inspired by the Forum on Information and Democracy: https://informationdemocracy.org/working-groups/concrete-solutions-against-the-infodemic/

[9] Wisdom, J. & Creswell, J.W. (2013, March). Mixed Methods: Integrating Quantitative and Qualitative Data Collection and Analysis While Studying Patient-Centered Medical Home Models. US Department of Health and Human Services. https://pcmh.ahrq.gov/sites/default/files/attachments/MixedMethods_032513comp.pdf

[10] Berthelsen, C. B., & Hølge-Hazelton, B. (2016). An evaluation of orthopaedic nurses' participation in an educational intervention promoting research usage: a triangulation convergence model. Journal of Clinical Nursing, 25(5-6), 846-855.

[11] Canadian Commission on Democratic Expression. (2021). 2020-21 Report of the Canadian Commission on Democratic Expression; Harms Reduction: A Six-Step Program to Protect Democratic Expression Online. *Public Policy Forum*. https://ppforum.ca/wp-content/uploads/2021/01/CanadianCommissionOnDemocraticExpression-PPF-JAN2021-EN.pdf

[12] Wilson, T., & Starbird, K. (2020). Cross-platform disinformation campaigns: Lessons learned and next steps. *Harvard Kennedy School Misinformation Review*, 1(1). https://doi.org/10.37016/mr-2020-002

[13] Miller, G. (2020, October 26). As U.S. election nears, researchers are following the trail of fake news. *Science Magazine*. https://www.sciencemag.org/news/2020/10/us-election-nears-researchers-are-following-trail-fake-news

[14] Global Engagement Center. (2020). GEC Special Report: Pillars of Russia's Disinformation and Propaganda Ecosystem. U.S. Department of State. https://www.state.gov/wp-content/uploads/2020/08/Pillars-of-Russia%E2%80%99s-Disinformation-and-Propaganda-Ecosystem_08-04-20.pdf

[15] Bagherpour, A., & Nouri, A. (2020, October 11). COVID Misinformation Is Killing People. *Scientific American*. https://www.scientificamerican.com/article/covid-misinformation-is-killing-people1/

[16] Laub, Z. (2019). Hate Speech on Social Media: Global Comparisons. *Council on Foreign Relations*. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

[17] Reuters. (2021, January 9). Google suspends Parler social network app over incitement to violence. *The Guardian.* http://www.theguardian.com/us-news/2021/jan/09/google-suspends-parler-social-network-app-over-incitement-to-violence

[18] Burkell, J. & Regan, P. M. (2019). Voter preferences, voter manipulation, voter analytics: policy options for less surveillance and more autonomy. Internet Policy Review, 8(4). https://doi.org/10.14763/2019.4.1438

[19] Boutilier, A. & MacCharles, T. (2020, May 12). COVID-19 'infodemic' reaching Canadians through social media and apps, survey suggests. *Toronto Star*. https://www.thestar.com/politics/federal/2020/05/12/facebook-says-it-flagged-covid-19-misinformation-50-million-times-in-april.html

[20] Rebuilding the Public Square. (2019). *Ryerson Leadership Lab.* https://www.ryersonleadlab.com/s/Rebuilding-the-Public-Square-Report-2019

[21] Brin, C., Charlton, S. & Leclair, K. (2020). Digital News Report: Canada. Centre d'études sur les médias. https://www.cem.ulaval.ca/wp-content/uploads/2020/06/dnr20_can_eng-pdf.pdf

Canadian Journalism Foundation. (2019, April 9). Canadians like, but don't trust, social media for news, according to CJF poll conducted by Maru/Matchbox. *Newswire.* https://www.newswire.ca/news-releases/canadians-like-but-don-t-trust-social-media-for-news-according-to-cjf-poll-conducted-by-maru-matchbox-879309205.html

[22] Karine Garneau & Clémence Zossou. (2021). Misinformation during the COVID-19 pandemic. Statistics Canada. https://www150.statcan.gc.ca/n1/pub/45-28-0001/2021001/article/00003-eng.htm

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

[24] Owen, T., Loewen, P., Ruths, D., Bridgman, A., Saleem, H. M., Merkley, E., & Zhilin, O. (2020). Understanding vaccine hesitancy in Canada: Attitudes, beliefs, and the information ecosystem. *Media Ecosystem Observatory.* https://files.cargocollective.com/c745315/meo_vaccine_hesistancy.pdf

[25] Canadian Race Relations Foundation and Abacus Data. (2021, January). Online Hate and Racism: Canadian Experiences and Opinions on What to Do About It. https://www.crrf-fcrr.ca/images/CRRF_OnlineHate_Racism_Jan2021_FINAL.pdf

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

[27] COVID-19 Fueling Anti-Asian Racism and Xenophobia Worldwide. (2020, May 12). *Human Rights Watch.* https://www.hrw.org/news/2020/05/12/covid-19-fueling-anti-asian-racism-and-xenophobia-worldwide

[28] Dad, N., & Khan, S. (2020). Threats against journalists. *Digital Rights Foundation*. https://www.international.gc.ca/campaign-campagne/media_freedom-liberte_presse-2020/policy_paper-documents_orientation-journalists-journalistes.aspx?lang=eng

[29] UNDP: Governments must lead fight against coronavirus misinformation and disinformation. (2020, June 10). *UNDP*. https://www.undp.org/content/undp/en/home/news-centre/news/2020/Governments_must_lead_against_coronavirus_misinformation_and_disinformation.html

[30] Evangelista, R. & Bruno, F. WhatsApp and political instability in Brazil: targeted messages and political radicalisation. Internet Policy Review 2019; 8(4): 10.14763/2019.4.1434.; https://misinforeview.hks.harvard.edu/article/images-and-misinformation-in-political-groups-evidence-from-whatsapp-in-india/

[31] Curry, B. (2021, January 18). Liberal government revising plan to regulate social media in light of U.S. Capitol riot. *The Globe and Mail.* https://www.theglobeandmail.com/politics/article-federal-officials-revising-plan-to-regulate-social-media-in-light-of/

[32] Bannerman, S. (2019, May 1). Canada's glaring failure to regulate Facebook. *Policy Options.* https://policyoptions.irpp.org/fr/magazines/may-2019/canadas-glaring-failure-regulate-facebook/

[33] Wong, T. (2019, September 25). Majority of Canadians want government to regulate social media, poll says. *The Toronto Star.* https://www.thestar.com/news/canada/2019/09/25/majority-of-canadians-want-government-to-regulate-social-media-poll-says.html

[34] Thompson, E. (2021, January 29). Facebook calls on Canadian government to set social media rules. *CBC.* https://www.cbc.ca/news/politics/facebook-parliament-twitter-canada-1.5892592

[35] Prime Minister of Canada (2019). Minister of Canadian Heritage Mandate Letter. Office of the Prime Minister. https://pm.gc.ca/en/mandate-letters/2019/12/13/minister-canadian-heritage-mandate-letter; Patriquin, M. (2021, February 4). With new legislation, Steven Guilbeault will make few friends in Big Tech. *Financial Post.* https://financialpost.com/technology/with-new-legislation-steven-guilbeault-will-make-few-friends-in-big-tech

[36] Zwibel, C. (2021, February 10). Regulating Social Media: Into the Unknown. *CCLA.* https://ccla.org/social-media-regulation/

[37] Thompson, E. (2020, April 15). Federal government open to new law to fight pandemic misinformation. *CBC.* https://www.cbc.ca/news/politics/covid-misinformation-disinformation-law-1.5532325

[38] Privacy Policy—EEA - Revisions. (n.d.). WhatsApp. Retrieved March 25, 2021, from https://www.whatsapp.com/legal/updates/privacy-policy-eea/?lang=en

[39] Johnson, D. (2020, October 16). Is Signal secure? How the encrypted messaging app compares to other apps on privacy protection. *Business Insider.* https://www.businessinsider.com/is-signal-secure

[40] Botha, Johnny & van 't Wout, Carien & Leenen, Louise. (2019). A Comparison of Chat Applications in Terms of Security and Privacy

[41] Parsons, C. (2019). Canada's New and Irresponsible Encryption Policy: How the Government of Canada's New Policy Threatens Charter Rights, Cybersecurity, Economic Growth, and Foreign Policy. *Citizen Lab.* https://citizenlab.ca/2019/08/canadas-new-and-irresponsible-encryption-policy-how-the-government-of-canadas-new-policy-threatens-charter-rights-cybersecurity-economic-growth-and-foreign-policy/.

[42] Comey, J. B. (2015, July 8). Going Dark: Encryption, Technology, and the Balances Between Public Safety and Privacy [Testimony]. Federal Bureau of Investigation. https://www.fbi.gov/news/testimony/going-dark-encryption-technology-and-the-balances-between-public-safety-and-privacy

[43] Wardle, C., & Derakhshan, H. (2018). Thinking about 'information disorder': Formats of misinformation, disinformation, and mal-information. In Journalism, fake news & disinformation: Handbook for journalism education and training. UNESCO. https://en.unesco.org/sites/default/files/journalism_fake_news_disinformation_print_friendly_0.pdf

[44] Ibid

[45] Wardle, C. (2017, February 16). Fake news. It's complicated. *First Draft.* https://firstdraftnews.org:443/latest/fake-news-complicated/

[46] Kim, S. (2015, June 1). All the Times People Were Fooled by The Onion. *ABC News*. https://abcnews.go.com/International/times-people-fooled-onion/story?id=31444478

[47] Goodman, J., & Carmichael, F. (2020, June 5). George Floyd: Fake White House image and protest videos debunked. *BBC News*. https://www.bbc.com/news/52934672

[48] Kenya election: Fake CNN and BBC news reports circulate. (2017, July 29). *BBC News*. https://www.bbc.com/news/world-africa-40762796

[49] Burgos, P. (2019, June 27). What 100,000 WhatsApp messages reveal about misinformation in Brazil. *First Draft*. https://firstdraftnews.org:443/latest/what-100000-whatsapp-messages-reveal-about-misinformation-in-brazil/

[50] Wardle, C., Pimenta, A., Conter, G., Dias, N., & Burgos, P. (2019). Comprova: An Evaluation of the Impact of a Collaborative Journalism Project on Brazilian Journalists and Audiences. *First Draft*. https://firstdraftnews.org/wp-content/uploads/2019/06/Comprova-Full-Report-Final.pdf?x79527

[51] Burgos, What 100,000 WhatsApp messages reveal about misinformation in Brazil.

[52] Mahapatra, S., & Plagemann, J. (2019). Polarisation and Politicisation: The Social Media Strategies of Indian Political Parties. German Institute of Global and Area Studies, 3. https://www.giga-hamburg.de/en/publications/11575625-polarisation-politicisation-social-media-strategies-indian-political-parties/

[53] Uttam, K. (2018, September 29). For PM Modi's 2019 campaign, BJP readies its WhatsApp plan. *Hindustan Times*. https://www.hindustantimes.com/india-news/bjp-plans-a-whatsapp-campaign-for-2019-lok-sabha-election/story-lHQBYbxwXHaChc7Akk6hcI.html

[54] Soutik Biswas. (2018, November 30). ShareChat: India's homegrown rival to WhatsApp. *BBC News*. https://www.bbc.com/news/world-asia-india-46341633

[55] Hui, J. Y. (2020). Social Media and the 2019 Indonesian Elections: Hoax Takes the Centre Stage. Southeast Asian Affairs, 155–172

[56] Afifa, L. (2019, March 8). Hoaxes Not Affecting Electability, Analysts Say. Tempo. https://en.tempo.co/read/1183107/hoaxes-not-affecting-electability-analysts-say; Bhwana, P. G. (2019, March 11). SMRC Survey: Some Believe KPU Not Neutral in Election. *Tempo*. https://en.tempo.co/read/1183904/smrc-survey-some-believe-kpu-not-neutral-in-election.

[57] Purohit, K. (2020, March 10). Misinformation, fake news spark India coronavirus fears. *Al Jazeera*. https://www.aljazeera.com/news/2020/3/10/misinformation-fake-news-spark-india-coronavirus-fears

[58] Kirdemir, B. (2020). Exploring Turkey's Disinformation Ecosystem: An Overview. Centre for Economics and Foreign Policy Studies. https://www.jstor.org/stable/resrep26087

[59] Zainul, H. (2020). Malaysia's Infodemic and Policy Response. Institute of Strategic and International Studies. http://www.jstor.com/stable/resrep24756

[60] Sherman, J. (2020, June 26). What's behind Russia's decision to ditch its ban on Telegram? *Atlantic Council*. https://www.atlanticcouncil.org/blogs/new-atlanticist/whats-behind-russias-decision-to-ditch-its-ban-on-telegram/

[61] Velch, V. (2021, February 26). Telegram: A Growing Social Media Refuge, for Good and Ill. *Just Security.* https://www.justsecurity.org/74947/telegram-a-growing-social-media-refuge-for-good-and-ill/

[62] Goldstein, J. A., & Grossman, S. (2021, January 4). How disinformation evolved in 2020. *Brookings*. https://www.brookings.edu/techstream/how-disinformation-evolved-in-2020/

[63] Cockerell, I. (2019, Sep 5). Inside China's Massive Surveillance Operation. *Wired*. Retrieved from https://www.wired.com/story/inside-chinas-massive-surveillance-operation/

[64] Diebert, R. (2020, May 7). WeChat users outside China face surveillance while training censorship algorithms. *The Washington Post.* Retrieved from https://www.washingtonpost.com/opinions/2020/05/07/wechat-users-outside-china-face-surveillance-while-training-censorship-algorithms/

[65] Elliott, J.K. (2018, July 16). India WhatsApp Killings: Why mobs are lynching outsiders over fake videos. *Global News.* Retrieved from https://globalnews.ca/news/4333499/india-whatsapp-lynchings-child-kidnappers-fake-news//

[66] Wakefield, J. (2019, April 23). Sri Lanka attacks: The ban on social media. *BBC News*. https://www.bbc.com/news/technology-48022530

[67] Satariano, A., & Alba, D. (2020, April 11). How a 5G Coronavirus Conspiracy Theory Fueled Arson and Harassment in Britain. *New York Times*. https://www.nytimes.com/2020/04/10/technology/coronavirus-5g-uk.html

[68] Davey, J., Hart, M., & Guerin, C. (2020). An Online Environmental Scan of Right-wing Extremism in Canada. *Institute for Strategic Dialogue*. https://www.isdglobal.org/isd-publications/canada-online/

[69] Glaser, A. (2019, August 8). Telegram Was Built for Democracy Activists. White Nationalists Love It. *Slate Magazine*. https://slate.com/technology/2019/08/telegram-white-nationalists-el-paso-shooting-facebook.html

[70] Cinelli, M., De Francisci, G., Galeazzi, A., Quattrociocchi, W. & Starnini, M. The echo chamber effect of social media. Proceedings of the National Academy of Sciences Mar 2021, 118 (9) 2023301118; DOI: 10.1073/pnas.2023301118 https://www.pnas.org/content/118/9/e2023301118

Guhl, J. & Davey, J. (2020, June 26). A Safe Space to Hate: White Supremacist Mobilisation on Telegram. *Institute for Strategic Dialogue*. https://www.isdglobal.org/wp-content/uploads/2020/06/A-Safe-Space-to-Hate2.pdf

[71] Protsiuk, A. (2021, January 27). How Telegram harbours far-right groups. *The Fix*. https://thefix.media/2021/01/27/how-telegram-harbours-far-right-groups/

[72] Chau, D. (2021, January 19). Donald Trump supporters embrace Signal, Telegram and other "free speech" apps. *ABC News*. https://www.abc.net.au/news/2021-01-20/donald-trump-social-media-apps-free-speech-privacy/13071206

[73] Vanian, J. (2021, January 13). Private messaging apps Signal and Telegram are red hot after the Capitol riots. *Fortune*. https://fortune.com/2021/01/13/messaging-apps-signal-telegram-capitol-riots/

[74] Hern, A. (2021, January 24). WhatsApp loses millions of users after terms update. *The Guardian*. http://www.theguardian.com/technology/2021/jan/24/whatsapp-loses-millions-of-users-after-terms-update

[75] Tan, R. (2017, June 30). Russia wants to stop terrorists by banning their app of choice. Good luck. *Vox*. https://www.vox.com/world/2017/6/30/15886506/terrorism-isis-telegram-social-media-russia-pavel-durov-twitter; Lomas, N. (2015, November 19). After Paris Attacks, Telegram Purges ISIS Public Content. *TechCrunch*. https://social.techcrunch.com/2015/11/19/telegram-purges-isis-public-channels/; Warrick, J. (2016, December 23). The 'app of choice' for jihadists: ISIS seizes on Internet tool to promote terror. *The Washington Post*. https://www.washingtonpost.com/world/national-security/the-app-of-choice-for-jihadists-isis-seizes-on-internet-tool-to-promote-terror/2016/12/23/a8c348c0-c861-11e6-85b5-76616a33048d_story.html; Garrido, I. (2019, April 15). Simpatizantes del ISIS llaman a atentar contra la Semana Santa. *El Plura*l. https://www.elplural.com/sociedad/religion/isis-llama-a-atentar-contra-la-semana-santa_214583102

[76] Amarasingam, A., Maher, S. & Winter, C. (2021). How Telegram Disruption Impacts Jihadist Platform Migration. *Centre for Research and Evidence on Security Threats*. https://crestresearch.ac.uk/resources/how-telegram-disruption-impacts-jihadist-platform-migration/

[77] Porter, T. (2021, February 26). Facebook reported more than 20 million child sexual abuse images in 2020, more than any other company. *Business Insider*. https://www.businessinsider.com/facebook-instagram-report-20-million-child-sexual-abuse-images-2021-2

[78] Hern, A. (2021, January 21). Facebook admits encryption will harm efforts to prevent child exploitation. *The Guardian*. http://www.theguardian.com/technology/2021/jan/21/facebook-admits-encryption-will-harm-efforts-to-prevent-child-exploitation

[79] Sullivan, J. (2020, May 21). Preventing Unwanted Contacts and Scams in Messenger. *Messenger News*. https://messengernews.fb.com/2020/05/21/preventing-unwanted-contacts-and-scams-in-messenger/

[80] Avaaz. (2020). Facebook's Algorithm: A Major Threat to Public Health. *Avaaz*. https://secure.avaaz.org/campaign/en/facebook_threat_health/

[81] Menczer, F., & Hills, T. (2020, December 1). Information Overload Helps Fake News Spread, and Social Media Knows It. *Scientific American*. https://www.scientificamerican.com/article/information-overload-helps-fake-news-spread-and-social-media-knows-it/

[82] *Reuters Institute for the Study of Journalism, University of Oxford*. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-09/KM%20RISJ%20News%20in%20social%20media%20and%20messaging%20apps%20report%20_0.pdf

[83] Lewis, R. (2018). Alternative Influence: Broadcasting the Reactionary Right on YouTube. *Data & Society*. https://datasociety.net/library/alternative-influence/

[84] Mantas, H. (2021, March 25). Growing usage of encrypted messaging apps could make it harder to combat misinformation. *Poynter*. https://www.poynter.org/fact-checking/2021/growing-usage-of-encrypted-messaging-apps-could-make-it-harder-to-combat-misinformation/

[85] Gruzd, A., & Mai, P. (2020). The State of Social Media in Canada 2020. https://doi.org/10.5683/SP2/XIW8EW

Rebuilding the Public Square. (2019). *Ryerson Leadership Lab*. https://www.ryersonleadlab.com/s/Rebuilding-the-Public-Square-Report-2019

[86] Kim, L. (2018, September 20). The Top 7 Messenger Apps in the World. *Inc.Com*. https://www.inc.com/larry-kim/the-top-7-messenger-apps-in-world.html

[87] See the questionnaire for the questions, which were inspired by a similar set of questions posed in a study by Dominik Stecula, Mark Pickup and Clifton van der Linden discussed here: https://policyoptions.irpp.org/magazines/july-2020/who-believes-in-covid-19-conspiracies-and-why-it-matters/

[88] Ibid.

[89] Canadian Heritage. (2020, June 8). Guide to the Canadian Charter of Rights and Freedoms. Government of Canada. https://www.canada.ca/en/canadian-heritage/services/how-rights-protected/guide-canadian-charter-rights-freedoms.html

[90] Criminal Code, RSC 1985, c. C-46 https://laws-lois.justice.gc.ca/eng/acts/c-46/section-319.html

[91] Gill, L. (2020). The Legal Aspects of Hate Speech in Canada. *Public Policy Forum*. https://ppforum.ca/wp-content/uploads/2020/07/1.DemX_LegalAspects-EN.pdf

[92] The Digital Services Act package. (2020, June 2). [Text]. European Commission. https://ec.europa.eu/digital-single-market/en/digital-services-act-package

[93] Ibid.

[94] Beswick, E. (2020, December 15). Five reasons why the new Digital Services Act matters. *Euronews*. https://www.euronews.com/2020/12/15/five-reasons-why-the-digital-services-act-and-digital-markets-act-matter

[95] Against information manipulation. (n.d.). Gouvernement.Fr. Retrieved March 19, 2021, from https://www.gouvernement.fr/en/against-information-manipulation

LOI n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information. *Legifrance*. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037847559?r=aE9xkiXNml

[96] Ibid.

[97] Evaluierungsbericht zum Netzwerkdurchsetzungsgesetz (NetzDG) vorgelegt. (2020, September 9). Federal Ministry of Justice and Consumer Protection. https://www.BMJV.de/SharedDocs/Artikel/DE/2020/090920_Evaluierungsbericht_NetzDG.html

[98] Prime Minister of Canada. (2019). *Minister of Canadian Heritage Mandate Letter.* Office of the Prime Minister. https://pm.gc.ca/en/mandate-letters/2019/12/13/minister-canadian-heritage-mandate-letter

[99] Fact sheet—Online Harms Full Government Response. (2020, December 15). GOV.UK. https://www.gov.uk/government/publications/fact-sheet-online-harms-full-government-response

[100] Patel, P., Barr, W. P., McAleenan, K. K., & Dutton, P. (2019, October 4). Open Letter: Facebook's "Privacy First" Proposals. https://www.justice.gov/opa/press-release/file/1207081/download

[101] Continuing to Make Instagram Safer for the Youngest Members of Our Community. (2021, March 16). *Instagram*. https://about.instagram.com/blog/announcements/continuing-to-make-instagram-safer-for-the-youngest-members-of-our-community

[102] Protection from Online Falsehoods and Manipulation Act 2019. (2019). Singapore Statutes Online. https://sso.agc.gov.sg/Act/POFMA2019?TransactionDate=20191001235959

[103] Garcia, R.T. (2020, September 10). Brazil's "fake news" bill won't solve its misinformation problem. *MIT Technology Review*. https://www.technologyreview.com/2020/09/10/1008254/brazil-fake-news-bill-misinformation-opinion/

[104] Nojeim, G. (2020, July 24). Update on Brazil's Fake News Bill: The draft approved by the senate continues to jeopordize users' rights. *Centre for Democracy & Technology.* Retrieved from https://cdt.org/insights/update-on-brazils-fake-news-bill-the-draft-approved-by-the-senate-continues-to-jeopardize-users-rights/

[105] Tiwari, U., & Ben-Avie, J. (2020, June 29). Mozilla's analysis: Brazil's fake news law harms privacy, security, and free expression. *Mozilla*. https://blog.mozilla.org/netpolicy/2020/06/29/brazils-fake-news-law-harms-privacy-security-and-free-expression; Alimonti, V. (2020, June 7). New Hasty Attempt to Tackle Fake News in Brazil Heavily Strikes Privacy and Free Expression. *Electronic Frontier Foundation*. https://www.eff.org/deeplinks/2020/06/new-hasty-attempt-tackle-fake-news-brazil-heavily-strikes-privacy-and-free

[106] Lee, D., Maldoff, G., & Wimmer, K. (2020, March). Comparison: Indian Personal Data Protection Bill 2019 vs. GDPR. *IAPP*. https://iapp.org/resources/article/comparison-indian-personal-data-protection-bill-2019-vs-gdpr/

[107] Basu, A., & Sherman, J. (2020, January 23). Key Global Takeaways From India's Revised Personal Data Protection Bill. *Lawfare*. https://www.lawfareblog.com/key-global-takeaways-indias-revised-personal-data-protection-bill

[108] Our initial comments on the Personal Data Protection Bill, 2019. (2020, January 17). *Dvara Research Blog*. https://www.dvara.com/blog/2020/01/17/our-initial-comments-on-the-personal-data-protection-bill-2019/

[109] H.R.1865 - Allow States and Victims to Fight Online Sex Trafficking Act of 2017. U.S. Congress. https://www.congress.gov/bill/115th-congress/house-bill/1865

[110] Halpern, S. (2020, December 4). How Joe Biden Could Help Internet Companies Moderate Harmful Content. *The New Yorker*. https://www.newyorker.com/tech/annals-of-technology/how-joe-biden-could-help-internet-companies-moderate-harmful-content

[111] Rosen, G. (2020, August 11). Community Standards Enforcement Report, August 2020. *Facebook*. https://about.fb.com/news/2020/08/community-standards-enforcement-report-aug-2020/

[112] Tobin, A., Varner, M., & Angwin, J. (2017). Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up. *ProPublica*. https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes

[113] Paul, K. (2020, May 12). Facebook reports spike in takedowns of hate speech, terrorism. *Reuters.* https://www.reuters.com/article/us-facebook-enforcement-idUSKBN22O2TE

[114] Wyciślik-Wilson, M. S. (2020, November 5). WhatsApp is changing the way you report abuse. *TechRadar.* https://www.techradar.com/news/whatsapp-is-changing-the-way-you-report-abuse

[115] Safi, Michael. (2019, February 6). WhatsApp 'deleting 2m accounts a month' to stop fake news. *The Guardian*. https://www.theguardian.com/technology/2019/feb/06/whatsapp-deleting-two-million-accounts-per-month-to-stop-fake-news

[116] Editorial Staff. (n.d.). What are Telegram Admins and their Rights? Telegram Guide. Retrieved March 19, 2021, from https://telegramguide.com/telegram-admin-rights/

[117] Snap Inc. (2020, December 18). Transparency Report. https://www.snap.com/en-US/privacy/transparency/canada

[118] WhatsApp Help Center—About forwarding limits. (n.d.). WhatsApp. Retrieved March 19, 2021, from https://faq.whatsapp.com/general/chats/about-forwarding-limits/?lang=en

[119] Introducing a Forwarding Limit on Messenger. (2020, September 3). *Facebook*. https://about.fb.com/news/2020/09/introducing-a-forwarding-limit-on-messenger/

[120] Owen, L. H. (2019, September 27). WhatsApp's message forwarding limits do work (somewhat) to stop the spread of misinformation. *Nieman Lab*. https://www.niemanlab.org/2019/09/whatsapps-message-forwarding-limits-do-work-somewhat-to-stop-the-spread-of-misinformation/

[121] WhatsApp Help Center—How to add and remove group participants. (n.d.). WhatsApp. Retrieved March 19, 2021, from https://faq.whatsapp.com/android/chats/how-to-add-and-remove-group-participants/?lang=en

[122] How many people can I message at once on Facebook? | Facebook Help Center. (n.d.). *Facebook*. Retrieved March 19, 2021, from https://www.facebook.com/help/131313586947248

[123] Group chats. (n.d.). Signal Support. Retrieved March 25, 2021, from https://support.signal.org/hc/en-us/articles/360007319331-Group-chats

[124] Telegram FAQ. (n.d.). *Telegram*. Retrieved March 25, 2021, from https://telegram.org/faqq

[125] Beta testing the new GV2 groups: Big Signal group with as many volunteering forum members as possible? (2020, September 25). *Signal Community.* https://community.signalusers.org/t/beta-testing-the-new-gv2-groups-big-signal-group-with-as-many-volunteering-forum-members-as-possible/17030/20

[126] Kalogeropoulos, A. (2019). Groups and Private Networks – Time Well Spent? Reuters Institute Digital News Report 2019. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/inline-files/DNR_2019_FINAL.pdf

[127] Forum on Information & Democracy. (2020, November). Working Group on Infodemics: Policy Framework. https://informationdemocracy.org/wp-content/uploads/2020/11/ForumID_Report-on-infodemics_101120.pdf

[128] Ovadya, A. (2021, March). 'Contextualization Engines' can fight misinformation without censorship. *Medium*. https://aviv.medium.com/aa3023b0ae24

[129] Lange, E., & Lee, D. (2020, November 23). How One Social Media App Is Beating Disinformation. *Foreign Policy*. https://foreignpolicy.com/2020/11/23/line-taiwan-disinformation-social-media-public-private-united-states/

[130] Sanz, C. (2020, November 4). Twitter and Facebook slap labels on Trump's "misleading" election posts. *ABC News*. https://abcnews.go.com/Technology/twitter-facebook-slap-labels-trumps-misleading-election-posts/story?id=74020537; Roth, Y., & Pickles, N. (2020, May 11). Updating our approach to misleading information. *Twitter.* https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html

[131] Twitter Moments guidelines and principles. (n.d.). *Twitter*. Retrieved March 19, 2021, from https://help.twitter.com/en/rules-and-policies/twitter-moments-guidelines-and-principles

[132] Portnoy, E. (2019, November 1). Why Adding Client-Side Scanning Breaks End-To-End Encryption. *Electronic Frontier Foundation*. https://www.eff.org/deeplinks/2019/11/why-adding-client-side-scanning-breaks-end-end-encryption

[133] Rosenzweig, P. (2020, August 20). The Law and Policy of Client-Side Scanning. *Lawfare*. https://www.lawfareblog.com/law-and-policy-client-side-scanning

[134] Porter, J. (2021, March 10). TikTok will warn users before posting 'inappropriate or unkind' comments. *The Verge*. https://www.theverge.com/2021/3/10/22322814/tiktok-inappropriate-or-unkind-comments-warning-pop-up-anti-bullying

[135] Ferrer, C. C., Pflaum, B., Dolhansky, B., Bitton, J., Pan, J., & Lu, J. (2020, June 12). Deepfake Detection Challenge Results: An open initiative to advance AI. *Facebook AI*. https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/

[136] Dufour, N., & Gully, A. (2019, September 24). Contributing Data to Deepfake Detection Research. *Google AI Blog*. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[137] *Lytvynenko*, J. [JaneLytv] (2021, February 26). Those Tom Cruise deepfakes on TikTok are deeply unsettling. Let's run them through @sensityai 's new deepfake detector tool to see how they do *Down pointing backhand index* [Tweet]. https://twitter.com/JaneLytv/status/1365362169827762184

[138] Ledford, H. (2019). Millions of black people affected by racial bias in health-care algorithms. Nature, 574(7780), 608–609. https://doi.org/10.1038/d41586-019-03228-6

[139] Khatib, H. A., & Kayyali, D. (2019, October 23). Opinion | YouTube Is Erasing History. *The New York Times*. https://www.nytimes.com/2019/10/23/opinion/syria-youtube-content-moderation.html

[140] Trump, K.S., Rhody, J., Edick, C., & Weber, P. (2018). Social Media and Democracy: Assessing the State of the Field and Identifying Unexplored Questions. Conference Paper from Social Media and Democracy: Assessing the State of the Field and Identifying Unexplored Questions.

[141] Goldman, E. (2018, October 3). Good News! USMCA (a/k/a NAFTA 2.0) Embraces Section 230-Like Internet Immunity. *Technology & Marketing Law Blog*. https://blog.ericgoldman.org/archives/2018/10/good-news-usmca-a-k-a-nafta-2-0-embraces-section-230-like-internet-immunity.htm; Geist, M. (2018, October 1). From Copyright Term to Super Bowl Commercials: Breaking Down the Digital NAFTA Deal. Michael Geist. https://www.michaelgeist.ca/2018/10/from-copyright-term-to-super-bowl-commercials-breaking-down-the-digital-nafta-deal/

[142] Online Harms White Paper. (2020, December 15). GOV.UK. https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper

[143] About Christchurch Call. (n.d.). Christchurch Call. Retrieved March 19, 2021, from https://www.christchurchcall.com/call.html

[144] Canada's Digital Charter: Trust in a digital world - Innovation for a better Canada. (2021, January 12). Government of Canada; Innovation, Science and Economic Development Canada. https://www.ic.gc.ca/eic/site/062.nsf/eng/h_00108.html

[145] Canada Declaration on Electoral Integrity Online. (2019, May 27). Government of Canada. https://www.canada.ca/en/democratic-institutions/services/protecting-democracy/declaration-electoral-integrity.html

[146] Canadian Commission on Democratic Expression, 2020-21 Report of the Canadian Commission on Democratic Expression

[147] Ibid, 31.

[148] May, P., Koski, C., & Stramp, N. (2016). Issue expertise in policymaking. *Journal of Public Policy*, 36(2), 195-218.

[149] Casey, B.J., Farhangi, A., & Vogl, R. (2018). Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise. *Berkeley Technology Law Journal*, 34, 143.

[150]Barbosa, S., & Milan, S. (2019). Do Not Harm in Private Chat Apps: Ethical Issues for Research on and with WhatsApp. *Westminster Papers in Communication and Culture*, 14, 49–65. https://doi.org/10.16997/wpcc.313; Sehat, C.M., Prabhakar, T. & Kaminski, A. (2021, March 15). Ethical Approaches to Closed Messaging Research: Considerations in Democratic Contexts. Misinfocon and the Carter Center. https://electionstandards.cartercenter.org/verifying-elections-misinfocon2020/ethical-approaches-to-closed-messaging-research-considerations-in-democratic-contexts/

[151] Eshet, Y. (2004). Digital Literacy: A Conceptual Framework for Survival Skills in the Digital era. *Journal of Educational Multimedia and Hypermedia*, 13(1), 93–106

[152] Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. Proceedings of the National Academy of Sciences, 117(27), 15536. https://doi.org/10.1073/pnas.1920498117

[153] Patil, S. (2019, April 29). Opinion | India Has a Public Health Crisis. It's Called Fake News. *The New York Times*. https://www.nytimes.com/2019/04/29/opinion/india-elections-disinformation.htmlPatil, S. (2019, April 29). Opinion | India Has a Public Health Crisis. It's Called Fake News. *The New York Times*. https://www.nytimes.com/2019/04/29/opinion/india-elections-disinformation.html

[154] Jazynka, K. (2017, April 6). Colleges turn 'fake news' epidemic into a teachable moment. *Washington Post*. https://www.washingtonpost.com/lifestyle/magazine/colleges-turn-fake-news-epidemic-into-a-teachable-moment/2017/04/04/04114436-fd30-11e6-99b4-9e613afeb09f_story.html

[155] WhatsApp and NASSCOM collaborate to teach about fake news. (2019, March 19). *India Today*. https://www.indiatoday.in/education-today/news/story/whatsapp-and-nasscom-collaborate-to-teach-about-fake-news-1481882-2019-03-19

[156] Aiello, R. (2019, February 27). Feds unveil plan to tackle fake news, interference in 2019 election. *CTV News*. https://www.ctvnews.ca/politics/feds-unveil-plan-to-tackle-fake-news-interference-in-2019-election-1.4274273

Department of Finance Canada. (2019, March 19). Budget 2019. Retrieved from https://www.budget.gc.ca/2019/docs/plan/toc-tdm-en.html

[157] International Centre for Journalists (2019). The State of Technology in Global Newsrooms. International Centre for Journalists.

[158] Rowley, J. D. (2018, February 8). Messaging Apps: Where Cryptocurrency and Conversation Collide. *Crunchbase News*. https://news.crunchbase.com/news/messaging-apps-cryptocurrency-conversation-collide/

[159] Pymnts. (2017, August 25). Mark Cuban unveils new blockchain project: mercury protocol. Retrieved from https://www.pymnts.com/blockchain/2017/mark-cuban-blockchain-mercury-protocol/

[160] Decentralized messaging apps (2021) – ySign Blog. (2020, December 29). *YSign*. https://blog.ysign.app/index.php/2020/12/29/decentralized-messaging-apps/

[161] Marr, B. (2019). What Is Homomorphic Encryption? And Why Is It So Transformative? [Map]. *Forbes*. https://www.forbes.com/sites/bernardmarr/2019/11/15/what-is-homomorphic-encryption-and-why-is-it-so-transformative/

[162] Covid-19 Triggers Wave of Free Speech Abuse. (2021, February 11). *Human Rights Watch*. https://www.hrw.org/news/2021/02/11/covid-19-triggers-wave-free-speech-abuse

[163] Cambodia expands monitoring of "fake news." (2021, January 28). *UCA News*. https://www.ucanews.com/news/cambodia-expands-monitoring-of-fake-news/91186

[164]Cambodia: Covid-19 Spurs Bogus 'Fake News' Arrests. (2020, April 29). *Human Rights Watch*. https://www.hrw.org/news/2020/04/29/cambodia-covid-19-spurs-bogus-fake-news-arrests

[165] Anderson, J., & Rainie, L. (2017, October 19). The Future of Truth and Misinformation Online. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/

[166] Canadian Commission on Democratic Expression, 2020-21 Report of the Canadian Commission on Democratic Expression https://ppforum.ca/wp-content/uploads/2021/01/CanadianCommissionOnDemocraticExpression-PPF-JAN2021-EN.pdf