Public Inquiry Into Foreign Interference in Federal Electoral Processes and Democratic Institutions | Enquête publique sur l'ingérence étrangère dans les processus électoraux et les institutions démocratiques fédéraux

Summary Report

Author: Vivek Krishnamurthy, Associate Professor, University of Colorado Law School

Panel Theme: Disinformation, Digital Space, and Democratic Processes

Date: October 30, 2024

Key Issues

This report addresses the significant challenges public and private entities face in identifying and responding to misinformation, disinformation, and malinformation (MDM) on the large online platforms that host and curate so much of the world's expression. It explains the challenges associated with determining the truth or falsity of online content at the scale and speed that the contemporary online information ecosystem operates. It also explores difficulties in ascertaining intent online and in deciding how to act once MDM is identified. The report concludes by examining potential regulatory responses that may help address these challenges in Canada and throughout the broader democratic family of nations.

Assessment

Adjudicating Truth and Intent

A first-order challenge in addressing MDM's threat to democracy lies in identifying it. This is no easy task; it requires adjudicating the truth or falsity of statements made online and the intent behind them. Further complicating matters is that these assessments must happen at the speed and scale at which the modern information environment operates.

Consider the accepted definitions of disinformation, misinformation, and malinformation. Disinformation and misinformation are "verifiably false claims" shared with or without intent to deceive, respectively. Malinformation, by contrast, is "information that stems from the truth but is exaggerated or used out of context to mislead and cause harm."

We can visualize the sum total of information shared online as a "normal distribution" that takes the familiar shape of a bell-shaped curve. At one end of the curve, we have verifiable truths (e.g., "the Earth is a sphere"); at the other, obvious falsehoods (e.g., "the Earth is flat”).[1] In the vast middle lies content whose truth is harder to ascertain, as is the intent behind its creation and distribution online.

Consider, for example, the following statement: “Many doctors believe that COVID-19 vaccines cause infertility.” Some may say the statement is true in the narrow sense that a number of medical professionals sincerely hold this view despite the overwhelming weight of scientific evidence to the contrary.[2] To determine whether this statement is malinformation, we must consider the speaker’s intent—which may or may not be discernible based on the context of the statement.

Yet others might argue that this statement is misleading because the word “many” implies a significant number of medical professionals hold this belief, when it is, in fact, a small minority. On this view, the statement could be misinformation or disinformation, so again we must determine the speaker’s intent to decide how to respond.

It can take our legal system months or years to determine truth and intent in fraud or defamation cases. The companies that operate online platforms, by contrast, often have minutes to make such decisions to prevent MDM-related harms. Social media content’s “half-life”—the time by which half of its audience has seen it—is measured in hours if not minutes, creating a narrow window for proactive action.[3] Otherwise, we are left with *ex post* responses and remedies to falsehoods that may well have spread around the world.

Challenges of Scale

The challenges of detecting MDM are compounded by the sheer volume of online content. Every minute of every day, some 510,000 comments are posted to Facebook[4] and over 500 hours of video are uploaded to YouTube.[5] Consequently, companies that operate large platforms rely on automated systems to detect problematic content. Yet, as with any large-scale decision-making system—from trial courts to radiologists—even the best automated content moderation systems make errors with some regularity. Even if error rates are low, the sheer quantity of content that is shared online means that such systems generate a

[1] And yet, five centuries after the first circumnavigation of the Earth by Ferdinand Magellan, the belief that the Earth is flat persists. *See* Rob Picheta, “The flat-Earth conspiracy is spreading around the globe. Does it hide a darker core?”, (18 November 2019), online: *CNN.com* <https://www.cnn.com/2019/11/16/us/flat-earth-conference-conspiracy-theories-scli-intl/index.html>.

[2] Sahana Sule et al, “Communication of COVID-19 Misinformation on Social Media by Physicians in the US” (2023) 6:8 JAMA Network Open e2328928.

[3] Scott M Graffius, “Lifespan (Half-Life) of Social Media Posts: Update for 2024” (2024), online: <https://rgdoi.net/10.13140/RG.2.2.21043.60965>.

[4] Susie Marino, “What Happens in an Internet Minute? [2024 Statistics]”, online: *LocaliQ* <https://localiq.com/blog/what-happens-in-an-internet-minute/>.

[5] “YouTube for Press”, online: *blog.youtube* <https://blog.youtube/press/>.

vast number of mistakes. For instance, a 0.1% error rate in Facebook's content moderation systems would produce over 700,000 errors a day or some 286 million over the course of a year.

Those who study large-scale systems distinguish between Type I and Type II errors. Type I errors, or false positives, occur when a system misclassifies a phenomenon as meeting certain criteria when it does not. Conversely, Type II errors, or false negatives, occur when the system fails to identify phenomena that meet the criteria. In the context of a COVID-19 test, a false positive would show infection where there is none, while a false negative would miss an actual infection.

Both types of errors are inevitable in large-scale systems, and their consequences can be significant in their effects— especially before high-stakes events like elections. A false positive could remove a critical statement by a candidate from the public sphere, while a false negative might allow foreign threat actors to target voters with manipulative content.

Yet in matters of free expression, the legal system treats Type I errors as more consequential than Type II errors. Type I errors suppress legitimate speech, the archetypal harm that free expression law aims to prevent. Conversely, constitutional protections for free expression assume that in the "marketplace of ideas," the truth will prevail, making the law more reluctant to regulate content based on its apparent falsity.[6]

Content moderation systems may improve as technology evolves, but our constitutional commitments to free expression nonetheless require us to tolerate a wide range of expression that might be false, distasteful, or even harmful. Consequently we must focus on how best to mitigate the harms to democracy that the continued prevalence of MDM online poses, which militates in favour of an "all of society" approach to this problem.

Determining the Effectiveness of Interventions

Deciding what to do about MDM once it is identified presents yet another challenge. Private companies enjoy more latitude than governments to act against harmful content (such as MDM) in view of the constitutional constraints faced by the latter. Of course, companies are expected under emerging international norms to respect human rights—including the

[6] As McLachlin J. (as she then was) notes in her dissenting reasons in *Keegstra,* the fact that "there is no guarantee that the free expression of ideas will in fact lead to the truth" does not "negate the essential validity of the notion of the value of the marketplace of ideas." As she further explains, "[w]hile freedom of expression provides no guarantee that the truth will always prevail, it still can be argued that it assists in promoting the truth in ways which would be impossible without the freedom." *R. v. Keegstra,* [1990] 3 S.C.R. 697 at 803 (McLachlin J., dissenting).

right to free expression.[7] Even so, there is broad agreement that the range of interventions available to private actors to deal with MDM is significantly wider than what is available to governments.[8]

Even so, this does not answer the question of how exactly companies should respond to MDM, whether it is a one-off occurrence or part of a larger pattern of "coordinated inauthentic behavior."[9] Options range from doing nothing to permanently banning offending accounts. They also extend to labeling content as false, referring users to independent fact-checks, downranking content in algorithmically curated feeds, or demonetizing it to prevent those spreading MDM from profiting from it.

We can have normative debates on what companies *should* do, but as in the health sciences, empirical evidence on what works best would be better. Yet such evidence is sorely lacking not simply because of the complexities of studying these phenomena, but also because the politicization of efforts to counter MDM has led some platforms to curtail their efforts to counter these phenomena.[10] This is especially regrettable given what we now know about the second-order effects of certain actions to combat MDM that seemed sensible just a few years ago.

Consider the decision by many social media platforms to ban former U.S. President Donald Trump after the insurrection at the U.S. Capitol on January 6, 2021.[11] These decisions initially garnered widespread support, and logic seemed to suggest that they would serve to constrain Mr. Trump's efforts to undermine the results of the 2020 U.S. presidential election. What we did not foresee at the time, however, was that Mr. Trump had the resources to create his own social media platform that lacks the safeguards found on more

[7] The United Nations Guiding Principles on Business and Human Rights developed by the late Professor John Ruggie explains that the concept of business respect for human rights requires them to "they should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved." Office of the United Nations High Commissioner for Human Rights, "Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework," March 21, 2011, principle 11.

[8] *See, e.g.,* Martha Minow & Vicki Jackson, "Facebook Suspended Trump. The Oversight Board Shouldn't Let Him Back.", (8 March 2021), online: *Lawfare* <https://www.lawfaremedia.org/article/facebook-suspended-trump-oversight-board-shouldnt-let-him-back>.

[9] "Coordinated inauthentic behavior" has been defined as "a manipulative communication tactic that uses a mix of authentic, fake, and duplicated social media accounts to operate as an adversarial network (AN) across multiple social media platforms." *See* Monica Murero, "Coordinated inauthentic behavior: An innovative manipulation tactic to amplify COVID-19 anti-vaccine communication outreach via social media" (2023) 8 Front Sociol 1141416.

[10] *See, e.g.,* Sheera Frenkel & Mike Isaac, "How Meta Distanced Itself From Politics", *The New York Times* (24 September 2024), online: <https://www.nytimes.com/2024/09/24/technology/meta-election-politics.html>.

[11] Melina Delkic, "Trump's banishment from Facebook and Twitter: A timeline.", *The New York Times* (10 May 2022), online: <https://www.nytimes.com/2022/05/10/technology/trump-social-media-ban-timeline.html>.

established platforms.[12] This complicates our assessment of whether companies made the right decision back to 2021 to "de-platform" Mr. Trump.

Recommendations

In light of these complexities, how should Canadian policymakers address MDM? While a whole-of-society approach that includes civic education and increased institutional resilience is essential, my focus here is on addressing the role online intermediaries play in distributing MDM.

First, it is crucial for regulators, academics, journalists, the public, and the platforms to gain a better understanding of how MDM is created and disseminated online. Just as medical research has transformed our understanding of diseases, we need deeper insights on how MDM and other harmful content spreads through different online platforms. Bill C-63, which would provide researchers with more access to data from online platforms, is a promising first step in this regard. In so doing, however, we must make sure that the rights of social media users to privacy and anonymity are not compromised, and that researcher access to social media data does not itself contribute to the harms we are trying to prevent—as was the case in the lead-up to the 2016 "Cambridge Analytica" scandal.[13]

Second, smart regulations can incentivize platforms to develop better approaches to combat MDM. Many jurisdictions, including the EU, have embraced risk-based[14] technology sector regulation in laws such as the General Data Protection Regulation (GDPR) and the Digital Services Act (DSA). These laws require companies to assess risks to fundamental rights, implement mitigations, evaluate effectiveness, submit to independent audits, and to improve their measures over time. Canada's proposed Online Harms Act (Bill C-63) and Artificial Intelligence and Data Act (Bill C-27) both follow this regulatory approach.

[12] *See, e.g.,* Neil Bedi et al, "How Trump Is Using Truth Social to Concoct and Spread Conspiracy Theories", *The New York Times* (29 October 2024), online: <https://www.nytimes.com/interactive/2024/10/29/us/politics/trump-truth-social-conspiracy-theories.html>.

[13] Cambridge Analytica obtained more than 50 million Facebook profiles as part of its efforts to help conservative U.S. political candidates with reaching voters in the 2016 U.S. elections from Dr. Aleksandr Kogan, an academic researcher who had obtained the data under false pretenses. *See* "FTC Sues Cambridge Analytica, Settles with Former CEO and App Developer", (23 July 2019), online: *Federal Trade Commission* <https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-sues-cambridge-analytica-settles-former-ceo-app-developer>.

[14] Julia Black, "Risk-based Regulation: Choices, Practices and Lessons Being Learnt" in *Risk and Regulatory Policy* OECD Reviews of Regulatory Reform (OECD, 2010) 185.

Third, beyond risk-based regulation, policymakers could encourage more experimentation in anti-MDM interventions. The technology sector frequently uses “A/B testing,” where two versions of a feature are trialed, and the better performer is deployed. Policymakers might consider incentives for platforms to test MDM interventions in this manner to discover more effective solutions.

Finally, the most effective interventions against MDM may not directly target the content itself but could address related factors. For example, to address viral disinformation linked to mob violence in Brazil, India, and Indonesia, WhatsApp limited the number of people to whom users could forward content, hence slowing its spread.[15] This design-based solution appears effective and illustrates the potential of design tweaks and other creative solutions to confront the challenges posed by MDM.

[15] *See, e.g.,* Philipe de Freitas Melo et al, *Can WhatsApp Counter Misinformation by Limiting Message Forwarding?* (arXiv, 2019) arXiv:1909.08740.