PUBLIC ORDER
EMERGENCY
COMMISSION

COMMISSION
SUR L'ÉTAT
D'URGENCE

Commissioned Paper:

Mis- Dis- and Mal-Information and the Convoy: An Examination of the Role and Responsibilities of Social Media

Prepared by: Emily B. Laidlaw

Note to Reader

Pursuant to Rules 5-10 of the Commission's Policy Phase Rules of Practice and Procedure, the Commissioner may, in his discretion, engage external experts to produce discussion, research and policy papers, known as "Commissioned Papers".

Any views expressed in a Commissioned Paper are those of the author(s) and do not necessarily reflect the views of the Commissioner. Statements of fact contained in a Commissioned Paper do not necessarily represent the Commissioner's views. The Commissioner's findings of fact are based on the evidence presented during the Commission's hearings.

Parties and members of the public may provide written comments to the Commission in response to this paper. Information about the process for filing comments, including deadlines, are set out in the Commission's *Notice re Policy Phase of the Commission*, which is available on the Commission's website.

Mis- Dis- and Mal-Information and the Convoy: An Examination of the Roles and Responsibilities of Social Media

Dr. Emily B. Laidlaw, Canada Research Chair in Cybersecurity Law, Associate Professor, Faculty of Law, University of Calgary
emily.laidlaw@ucalgary.ca

A paper prepared for the Public Order Emergency Commission
September 2022

Table of Contents

Summary

This paper supports the Public Order Emergency Commission in the examination of "the impact, role and sources of misinformation and disinformation, including the use of social media".[1] The term social media is used broadly in this paper to refer to applications that are designed to enable third parties to interact, create and share content, including messaging, video, audio, and images.

This paper does not make factual findings as to online information manipulation and the Convoy. Rather, the purpose of this paper is to deepen understanding of the information environment of mis-, dis- and mal-information, how it is regulated and how this intersects with the Convoy. Social media was the central nervous system of the Convoy, and exploration of its role crosses numerous domains, such as law, psychology, history, sociology, and public policy, to name a few. Even within law, the applicable laws (and significant gaps in law) are too numerous to explore in detail. To the extent that I can provide more detail for interested readers, I do so in the footnotes, and I also encourage readers to peruse the many resources cited in this paper.

The paper is structured as follows. Part I examines the various social media used in the Convoy, the meaning of mis-, dis- and mal-information, how it spreads, the psychology and impact. Parts II and III examine how information manipulation on social media is regulated. There are two angles to regulation that are relevant. First, what laws regulate users and other entities that consume or spread mis-, dis- or mal-information? This is the question of whether an individual, for example, commits a crime or can be civilly liable for spreading false information. A necessary part of this analysis is the right to freedom of expression: its value, application, and limits. This aspect of regulation is examined in Part II. Second, what are social media providers legal and governance responsibilities to address mis-, dis- and mal-information? This is examined in Part III and entails analysis of the laws that regulate social media companies and how they self-regulate through content moderation.[2]

Part I The Convoy and the Online Information Environment

Social Media and the Convoy

[1] See (a)(ii)(c) https://publicorderemergencycommission.ca/files/documents/Order-in-Council-De%CC%81cret-2022-0392.pdf

[2] I want to thank my research assistants Akinkunmi Akinwunmi and Sylvana Crosby for their excellent work in support of this paper.

Social media provided the network that gave the Convoy, a movement that was otherwise "loose-knit" and "decentralized",[3] a shape and voice. As Stephanie Carvin comments, online activism was "the lifeblood of the convoy movement."[4] This was a Canadian movement, amplified first by Canadian social media influencers, then amplified by media and US influencers and exploited by other global actors.[5] On social media, leaders and influencers used a variety of social media, including video, audio and messaging applications, to spread their messages and engage with their followers.[6] They include Facebook, Twitter, TikTok, YouTube, Rumble, BitChute, Odysee, Telegram and Zello.

A few things are key about the role of social media in the Convoy. First, the movement started long before January 2022. The initial organization of the Convoy was through a Facebook group Canada Unity, which led a "United We Roll" convoy in 2019.[7] Before the Convoy, the content posted to Canada Unity had themes of anti-vaccination and anti-lockdown.[8] Many of the accounts and influencers of the Convoy are reported to have ties to far right groups such as the Canadian Yellow Vests and conspiracy theories.[9] Second, the movement was built "almost entirely by sharing video links."[10] Video-sharing platforms used include YouTube, and alt-YouTube platforms Rumble, BitChute and Odysee, livestreaming on Facebook and Twitter, and TikTok.[11] As will be explored below under the Psychology and Dangers of Information Manipulation, video and images are particularly powerful vectors to influence users. Third, posts were across various platforms, which has implications for content moderation. For example, videos posted on Facebook crowdsourced fundraising for the Convoy and directed users to GoFundMe. A video uploaded to Rumble, with less restrictive content moderation, was

[3] CBC, "How anger, faith and conspiracy theories fuelled the trucker convoy" (February 24, 2022) *The Fifth Estate*.

[4] Stephanie Carvin, "How the Freedom Convoy was fuelled by online activism" (March 5, 2022) *National Post,* online: https://nationalpost.com/opinion/stephanie-carvin-how-the-freedom-convoy-was-fuelled-by-online-activism.

[5] *Ibid.* Scam and hacked Convoy accounts were created and removed by Facebook: Elizabeth Culliford, "Meta says it removed scammers' Canada convoy Facebook group" (February 7, 2022) *Reuters*, online: https://www.reuters.com/technology/meta-says-it-removed-scammers-canada-convoy-facebook-groups-2022-02-08/; Anya van Wagtendonk *et al*, "The hacked account and suspicious donations behind the Canadian trucker protests" (February 8, 2022) *Grid*, online: https://www.grid.news/story/misinformation/2022/02/08/the-hacked-account-and-suspicious-donations-behind-the-canadian-trucker-protests/.

[6] The list does not cover all social media used by organizers and influencers of the convoy, but it gives an idea of the main services used. See this summary: Maggie Parkhill, "Who is who? A guide to the major players in the trucker convoy protest" (February 22, 2022) *CTV News*, online: https://www.ctvnews.ca/canada/who-is-who-a-guide-to-the-major-players-in-the-trucker-convoy-protest-1.5776441.

[7] Ryan Broderick, "How Facebook Twisted Canada's Trucker Convoy into an International Movement" (February 19, 2022) *The Verge*, online: https://www.theverge.com/2022/2/19/22941291/facebook-canada-trucker-convoy-gofundme-groups-viral-sharing.

[8] *Fifth Estate*, *supra* note 3.

[9] Broderick, *supra* note 7.

[10] Broderick expands on some of his research on his substack garbageday: "Freedom Convoy Facebook Content Is Coming From YouTube" (February 9, 2022), online: https://www.garbageday.email/p/boomers-are-weird-and-obsessive-posters

[11] Parkhill, *supra* note 6; Press Progress, "Meet the Extremists and Social Media Influencers at the Centre of the Far-Right Siege of Ottawa" (February 8, 2022), online: https://pressprogress.ca/meet-the-extremists-and-social-media-influencers-at-the-centre-of-the-far-right-siege-of-ottawa/.

shared on Facebook and on messaging app Telegram. Content moderation by these platforms is explored in Part III.

The Convoy was initially organized through the Facebook group Canada Unity. When the exemption from the vaccine mandate for truck drivers was set to end in January 2022, Canada Unity began posting about truckers. Then a Facebook group “Freedom Convoy 2022” was created. At the time of writing, the Canada Unity Facebook page has 79,877 followers and 34,161 likes.[12] On Twitter, some of the earliest posts date to January 12, 2022.[13] On January 18, 2022, the Convoy’s spread across Facebook gained momentum when a video about the protest was posted on Rumble.[14] The video, titled “Freedom Convoy 2022” has garnered 60,588 views on Rumble.[15] It provided links to groups on Facebook, Telegram, GoFundMe and Change.org, and to Canada Unity’s website.

In addition to mainstream social media, organizers used messaging app Telegram, walkie-talkie app Zello and video-sharing platforms, Rumble, BitChute and Odysee.[16] Telegram is a messaging app comprised of groups and channels. Groups are spaces for members to chat, while channels enable one-to-many broadcasting of messages. Groups and channels can be public or private. However, even private groups can have up to 200,000 members, and channels can have unlimited subscribers.[17] Telegram supports what it calls “secret chats”, which uses end-to-end encryption, meaning that Telegram does not see the content of these groups chats, nor stores it on their servers. Only participants in the group know what is discussed and shared.[18]

The app Zello was used by leaders to coordinate meeting locations.[19] Zello is a walkie-talkie app where users can set up public or private channels for communication. Private channels are end-to-end encrypted. Organizers primarily used public channels, which are capped at 7,000 users.[20] Once in a channel, users can hear and talk to everyone else in the channel. The organizer might set up several channels and can broadcast messages to all channels at once. Users can be connected to multiple channels at the same time.[21] Users can also send texts and images. With

[12] Here is the URL for Canada Unity’s Facebook group, online: https://www.facebook.com/CanadaUnity/.

[13] See posts on Twitter by Canadian for Freedom (January 12, 2022): https://twitter.com/CanFreedomLover/status/1481340478247346179?s=20&t=j-FIYPbSX217iiw4Z4DKPw, Dr. Ezra Kaxah (January 13, 2022), https://twitter.com/EzraKahan/status/1481760964043325448?s=20&t=j-FIYPbSX217iiw4Z4DKPw and (January 17, 2022) https://twitter.com/EzraKahan/status/1483134186919706626, and Fringe-Juli (January 13, 2022), https://twitter.com/Juliz1lb/status/1481767182577160193?s=20&t=j-FIYPbSX217iiw4Z4DKPw.

[14] Ryan Broderick, *supra* note 7.

[15] This number is as of August 31, 2022.

[16] TVO Today, “How does Social Medial Fuel Protest?” (February 18, 2022), online: https://www.tvo.org/video/how-does-social-media-fuel-protest>.

[17] Telegram FAQ, online: https://telegram.org/faq#q-what-39s-the-difference-between-groups-and-channels

[18] Telegram Privacy Policy, online: https://telegram.org/privacy.

[19] Demar Grant, “What is Zello? Inside the app that helped organize “freedom convoy” blockades” (February 11, 2022) Toronto Star, online: https://www.thestar.com/news/canada/2022/02/11/what-is-zello-inside-the-app-that-helped-organize-freedom-convoy-blockades.html.

[20] *Ibid.*

[21] Zello Channels, online: https://zello.com/product/features/channels/.

Zello, the organizers set up multiple channels for communication. For example, for the Ambassador Bridge blockade, a channel called "Windsor convoy 2" was used. Counter protestors joined some of these channels and disrupted communications.[22] Throughout, organizers communicated with supporters on various social media livestreaming events, posting videos, memes and text, and supporters participated by commenting or sharing posts, or creating their own content.

We cannot point to any one social media that was pivotal to the Convoy. Rather, all social media acted as the nervous system of the Convoy, as it has for many movements in the digital age. As the following paper explores, a variety of factors converged to give the Convoy momentum, in particular amplification and influence. The design of social media, including its recommender, advertising and content moderation systems, amplify certain content, enabling the creation and spread of influencers (individual and media). In the case of the Convoy, influencers primarily posted short videos, which are effective vectors for information manipulation, and used messaging apps, which can be less moderated. The cross-platform nature of communications means that cutting off one arm meant that communication could re-route elsewhere. At the application level, one of the only avenues through this is cross-platform collaboration.[23] At an infrastructure level, this is the open design of the Internet.[24]

Defining Mis-, Dis- and Mal-information

Information manipulation, however scoped, is a wide and varied landscape, some of which is "an old story fuelled by new technology".[25] Propaganda, hoaxes and campaigns of social smearing has been around since the earliest records. And to the extent that new technologies can serve to enable and amplify these narratives, they have been used for these purposes, such as Gutenberg's printing press, newspapers, and the radio.[26] The key difference now is the

[22] Grant, *supra* note 18.

[23] But see Evelyn Douek, "The Rise of Content Cartels" (February 11, 2020) *Knight First Amendment Institute*, online: https://knightcolumbia.org/content/the-rise-of-content-cartels.

[24] It is important to value the internet's original architecture and understand the internet stack. There is no one model of the internet stack. Most enduring, and fulsome, would be the OSI model developed by the International Organization for Standardization. There are other intermediaries than those explored thus far that are increasingly targets for content regulation, in particular arising from crisis incidents involving violent and extremist content and information manipulation. At the top of the internet stack is the application or content layer. Most debates about content regulation concern this layer because this is the layer of social media. Those higher up the internet stack rely on technologies deeper in the stack to operate and are impacted by their actions. Moving down the internet stack are technology like web hosting providers (e.g. WordPress) and cloud services (e.g. Amazon Web Services), and beneath them network infrastructure providers (e.g. domain name registries, internet service providers, and cloud delivery networks). The key thing to understand is that the deeper one moves in the stack, the more blunt, imprecise and less visible are their regulatory actions: Georgia Evans, "Down the Stack: Power and Accountability in Internet Intermediaries' Content Moderation Decisions" (July 9, 2021) *Kroeger Policy Review,* online: https://www.kroegerpolicyreview.com/post/down-the-stack-power-and-accountability-in-internet-intermediaries-content-moderation-decisions.

[25] Cherilyn Ireton *et al*, "Journalism, Fake News & Disinformation" (2018) *UNESCO* at 15.

[26] *Ibid* at 15-19. See Heidi J.S. Tworek, *News From Germany: The Competition to Control World Communications, 1900-1945* (Harvard University Press, 2019).

affordability of social networking, the speed, reach and precision with which a message can be communicated and spread, and access to cheap editing and publishing tools.[27]

The global communications space "is a common good of humankind".[28] Within this space, mis-, dis- and mal-information operate as a "complex system", wherein seeds are planted, then amplified to a wider audience and spread into the wider information ecosystem.[29] There is no uniform definition of mis- ,dis- and mal-information. This reflects the complexity of these concepts and the contextual nature of their application.[30] I recommend using UNESCO's definitions:

- Disinformation is information that is false, and the person who is disseminating it knows it is false. It is a deliberate, intentional lie, and points to people being actively disinformed by malicious actors.
- Misinformation is information that is false, but the person who is disseminating it believes that it is true.
- Malinformation is information, that is based on reality, but used to inflict harm on a person, organisation or country.[31]

[27] Samantha Bradshaw *et al*, "Industrialized Disinformation
2020 Global Inventory of Organized Social Media Manipulation" (2020) *Computational Propaganda Research Project*, online: https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/ at 11; Claire Wardle and Hossein Derakhshan, "Information Disorder: Toward an interdisciplinary framework for research and policy making" (2017) Council of Europe Report DGI(2017)09, online: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html at 11-12.

[28] Reporters Without Borders, *Global communication and information space: a common good of humankind*, online: https://rsf.org/en/global-communication-and-information-space-common-good-humankind.

[29] Canadian Security Intelligence Service, "Who Said What? The Security Challenges of Modern Disinformation" (December 5, 2016), online: https://www.canada.ca/content/dam/csis-scrs/documents/publications/disinformation_post-report_eng.pdf at Chapter 1.

[30] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan, "Disinformation and freedom of opinion and expression" (April 13, 2021), A/HRC/47/25 at para 9. Ronan Ó Fathaigh *et al*, "The perils of legally defining disinformation" (2021) 10(4) Internet Policy Review at 3.

[31] Ireton, *supra* note 25 at 44.

Broadly, these are types of "information disorder".[32] Other terms include viral deception,[33] information chaos, propaganda,[34] influence operations[35] and computational propaganda, which refers to the mix of platforms, algorithms, big data, and artificial intelligence that shape information flows and manipulate public opinion.[36] Often disinformation is used as the umbrella term, although that is not done in this paper to avoid confusion. For ease, when using a broad term, I will refer to information manipulation.

Definitions even vary for the terms mis-, dis- and mal- information. For example, the definition of malinformation provided above focuses on information rooted in reality that is intentionally shared to inflict harm.[37] Other definitions do not turn on intention to harm, concerned simply with the sharing of accurate information in a context that is misleading.[38] Still other definitions focus on the intentional sharing of private information.[39] These differences matter. If both disinformation and malinformation entail an intention to inflict harm, what is the line between accurate but misleading and false information? Is malinformation another term for doxing, narrowly concerned with public disclosure of private information? Claire Wardle and Hossein Derakhshan, for example, include hate speech as a form of malinformation, which fits uneasily with any of the above definitions, even if it is logical to include hate speech somewhere in this

[32] Wardle and Derakhshan, *supra* note 27. See figure at 20. These categories are not mutually exclusive: Dhanaraj Thakur and DeVan L. Hankerson, "Facts and their Discontents: A Research Agenda for Online Disinformation, Race, and Gender" (2021) *Center for Democracy & Technology*, online: https://cdt.org/insights/facts-and-their-discontents-a-research-agenda-for-online-disinformation-race-and-gender/ at 8.

[33] Khan, *supra* note 30 at para 13; Camille Francois, "Actors, Behaviors, Content: A Disinformation ABC" (September 20, 2019), *Transatlantic Working Group,* online: https://science.house.gov/imo/media/doc/Francois%20Addendum%20to%20Testimony%20-%20ABC_Framework_2019_Sept_2019.pdf at 1.

[34] Ireton *et al* said that propaganda is different than disinformation: "[t]he Term propaganda is not synonymous with disinformation, although disinformation can serve the interests of propaganda. But propaganda is usually more overtly manipulative than disinformation, typically because it traffics in emotional rather than informational messaging": *supra*, note 25 at 45. See also Yochai Benkler, Robert Faris and Hal Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford University Press, 2018) which discuss our constant connection as a propaganda rich environment.

[35] This term seems most prevalent in national security literature. See CSIS, *supra* note 29. Wardle and Derakhshan define it as "actions taken by governments or organized non-state actors to distort domestic or foreign political sentiment, most frequently to achieve a strategic and/or geopolitical outcome": *supra* note 27 at 16. The term 'fake news' is not used in this paper, because it is vague and politically loaded and is therefore unhelpful to the analysis: *Ibid* at 15-16; Alice Marwick and Rebecca Lewis, "Media Manipulation and Disinformation Online" (May 15, 2017) *Data & Society Research Institute*, online: https://datasociety.net/library/media-manipulation-and-disinfo-online/ at 44.

[36] See the Oxford Internet Institute, computational propaganda project, online: https://www.oii.ox.ac.uk/research/projects/computational-propaganda/.

[37] See Ireton *et al*, *supra* note 25. Kate Jones uses a similar definition, "Online Disinformation and Political Discourse: Applying a Human Rights Framework" (November 2019) *Chatham House The Royal Institute of International Affairs* at 2.2.

[38] Thakur and Hankerson, *supra* note 32 at 7.

[39] This is the definition referenced in Faithagh *et al*, *supra* note 30 at 4.

framework.[40] As Samantha Bradshaw and co-authors identified, harassment is increasingly the tool used to silence the press and political dissent, such as net centers in Guatemala using fake accounts to target individuals and journalists with labels such as terrorist and enemies of the state.[41]

Similarly, definitions of disinformation differ on critical points. For example, definitions of falsity vary, to include "information that is false", "verifiably false",[42] misleading[43] or inaccurate.[44] Further, the harm and intent required vary. Some definitions refer to harm to the public, while others refer to harm to people, social groups, or states. Some definitions include economic gain, meaning that the spread of false information for a commercial purpose would meet the definition of disinformation.[45]

As Fathaigh and co-authors state, these definitions are unfit legal categories although they serve the policy domain.[46] For the purpose of understanding the Convoy and the role of social media I recommend the following three categories. For the terms disinformation and misinformation, the UNESCO definitions above are instructive. Disinformation refers to the intentional spread of false information, and the person or entity spreading it knows the information is false. This is a category for malicious actors, such as a state-sponsored disinformation campaigns or an individual creating a subscriber-based website to intentionally spread false health information for economic gain. Misinformation refers to the intentional spread of false information by a person or entity that believes it to be true. A large swatch of false information shared on social media are misinformation, and there is certainly overlap between mis- and dis-information, especially concerning misleading content.

The third category is what I call the "everything else" bucket. Hate speech, harassment, defamation, violent and extremist content, trolling, and so on, are forms of expression or attack vectors that are fomented by mis- and dis-information, and foundational to it. Consider gamergate, the attack on female games developers. It exemplifies a mix of attacks, including doxing, where private information is obtained about an individual (whether through hacking or otherwise) and shared publicly, brigading, which is coordinated harassment, lies and gendered-based attacks.[47] To the extent malinformation is referred to in this paper, it is in the "everything else" bucket, and I use the UNESCO definition that it refers to information based in reality spread to inflict harm. The best lies are truth adjacent, and malinformation captures that.

[40] Wardle and Derakhshan, *supra* note 27 at 20. No attempt is made to define hate speech here, as there are a wide variety of definitions. See *Criminal Code*, RSC 1985, c C-46, s. 319 and the International Covenant on Civil and Political Rights, 1966, Article 20.
[41] Bradshaw, *supra* note 27 at 13.
[42] Fathaigh *et al*, *supra* note 30 at 4.
[43] Ireton *et al* give the example of cropping photos, or selecting quotes out of context as misleading, what is known as Framing Theory: *supra* note 25 at 47.
[44] Fathaigh *et al*, *supra* note 30 at 5.
[45] *Ibid* at 5-7.
[46] *Ibid.*
[47] Marwick and Lewis, *supra* note 35 at 27.

The ABC-D of the Information Environment

A useful framework for understanding the environment of information manipulation is the ABC-D Framework:[48]

- A is for manipulative actors who knowingly spread disinformation;
- B is for the behavioural techniques used to spread disinformation;
- C is for harmful content; and
- D is for the digital architectures of social media and how it impacts information distribution.

The authors created this framework to identify where to target solutions, but it is also a useful framework for making sense of the information space.

A is for manipulative actors

These actors knowingly and covertly launch a disinformation campaign. In this respect, many scholars separate state-backed disinformation from other actors. Research indicates that actors that produce disinformation are motivated by "ideology, money, and/or status and attention."[49] Recall that if the issue is misinformation, the actor does not knowingly spread false information. Indeed, one of the great challenges with information manipulation is that eventually the information seeds to humans, who believe it to be true and amplify it through their networks. In terms of strategy, many disinformation campaigns target key online influencers, who then spread the content to their networks. This was observable in a study of anti-vaccine content, which primarily spread through 12 key online influencers.[50]

Further, media plays a role in the spread of disinformation. Alice Marwick identifies a spectrum of media manipulation. At one end are websites intentionally created to deceive readers. The websites are designed to look like reputable sources and the stories are sensational to draw in readers and make money. In the middle of the spectrum are media, often ideologically driven, that publish a mix of true and false stories. At the other end of the spectrum are mainstream media, which might use clickbait-styled headlines that are sensationalist and misleading to increase readers, and might report on false news stories thus inadvertently amplifying the such stories.[51] Some state-backed media, such as Russia Today (RT), are well known for their role in spreading disinformation. This led major social media platforms to block RT from their services during the Russia-Ukraine war, and prompted the European Union (EU) to direct platforms to

[48] Francois, *supra* note 33.

[49] Marwick and Lewis, *supra* note 35 at 7-9, 27-29.

[50] The Center for Countering Digital Hate and Anti-Vax Watch, "Disinformation Dozen: the Sequel – How Big Tech is Failing to Act on Leading Anti-Vaxxers Despite Bipartisan Calls from Congress" (2021), online: https://counterhate.com/research/the-disinformation-dozen/. See also, for example, targeting of influencers in the Latinx community in the lead up to the 2020 election: Thakur & Harkeson, *supra* note 32 at 13-15.

[51] Marwick and Lewis, *supra* note 35 at 44-45.

block access to RT.[52] However, partisan news also plays a role in spreading mis- and dis-information, with misleading headlines and captions.[53] As Stephanie Carvin explains, American right-wing media had a greater impact on the US election in 2016 than disinformation.[54] As a result, journalists are often a key target of producers of disinformation.[55]

A question for the Commission would be who the key social media actors were who seeded the Convoy movement, both individual and media, and who amplified it. Various news reports identify some of the key actors, here and abroad.[56] A further question is whether there was – or the extent of- any foreign influence via social media, whether state, media or individual.[57] For example, 88% of donated funds through GoFundMe are reported to have originated in Canada.[58] Facebook also removed some Convoy groups, pages and accounts, that lured users to off-platforms websites with pay-per-click ads, hate groups and conspiracy groups.[59]

B is for deceptive behaviour

Behaviour refers to the techniques used by actors to spread information. The techniques include, among others:

- Automated tools, such as bots, which are algorithms that scrape data from the internet and then spread messages through networks and help boost the virality of content.[60] Not all bots are bad. A Vatican bot posts reflections. News organizations use bots to post breaking news, and so on. However, bots can also be sock puppets (false online identities), designed to impersonate another person and manipulate opinion. They are

[52] Council of EU Press Release, "EU imposes sanctions on state-owned outlets RT/Russia Today and Sputnik's broadcasting in the EU" (March 2, 2022), online: https://www.consilium.europa.eu/en/press/press-releases/2022/03/02/eu-imposes-sanctions-on-state-owned-outlets-rt-russia-today-and-sputnik-s-broadcasting-in-the-eu/. See Wardle and Derakhshan, *supra* note 27 at 13.

[53] See study discussed in Wardle and Derakhshan *supra* note 27 at 36-37.

[54] Stephanie Carvin, *Stand on Guard: Reassessing Threats to Canada's National Security* (University of Toronto Press, 2020) at 223.

[55] Thakur and Hankerson, *supra* note 32 at 8.

[56] See Parkhill, *supra* note 6. For example, one of Russell Brand's videos has 1,252,343 views as of September 9, 2022, "Truckers Convoy: Why The Mainstream Media Blackout?!" (January 27, 2022) *YouTube*, online: https://www.youtube.com/watch?v=itbSIqY4Nnw. Tucker Carlson also drew attention to the Convoy: Tucker Carlson, "Tucker Carlson: What's happening to truckers in Canada reveals the future of the United States" (February 21, 2022) *Fox News*, online: https://www.foxnews.com/opinion/tucker-carlson-truckers-canada-future-united-states.

[57] Carvin, *supra* note 4.

[58] Sarah Turnbull, "GoFundMe head testifies over Freedom Convoy fundraising, says most donors were Canadian" (March 3, 2022) *CTV*, online: https://www.ctvnews.ca/politics/gofundme-head-testifies-over-freedom-convoy-fundraising-says-most-donors-were-canadian-1.5804094.

[59] Culliford, *supra* note 5.

[60] "Bots are automatic posting protocols used to relay content in a programmatic fashion", Marco Bastos and Dan Mercea, "The Accountability of Social Platforms: Lessons from a Study of Bots and Trolls in the Brexit Campaign" (2018) 376(2128) Philosophical Transactions.

not as easy to detect as one might imagine and are designed to "fly under the radar".[61] Text prediction tools are improving that can produce content at scale. A sample text that reflects the ideological perspective the author wants distributed is used to generate unlimited articles in a similar vein, all appearing to be originals.[62]

- Image deception, such as the creation of deepfakes, which are altered audio, video or images that are hyper realistic.[63] For example, a deepfake video of Ukraine President Zelensky surrendering was circulated early in the conflict with Russia, although was quickly debunked.[64] Most often, image deception is a much simpler tactic of using out-of-context images. For example, old photos are presented as evidence of a new event, such as a photo posted after a global warming protest in London to show evidence of trash, but some of the photos were from Mumbai.[65] Memes are powerful tools of deception, because the visual image combined with short, emotional statements, are particularly good at influencing public opinion.[66] Memes are such effective tools of persuasion that the term "memetic warfare" was coined to refer to the key role memes play as a strategy in influence operations.[67]
- Manual trickery wherein humans participate online to shape internet flows. They might be a troll farm, who are paid or otherwise organized.[68] The purpose might be harassment, to shift political opinions or generally sow distrust in institutions and democracy. Some of these have become well-known, such as the Russian troll farms used to interfere with the 2016 US Election.[69] However, private firms are also regularly hired by companies to launch influence campaigns for their products and services, making disinformation for profit an industry.[70] Humans are more effective in the early

[61] Samuel Woolley, "The Business of Computational Propaganda Needs to End" (September 20, 2021) *Centre for International Governance Innovation*, online: https://www.cigionline.org/articles/the-business-of-computational-propaganda-needs-to-end/; Bastos and Mercea, *supra* note 56.

[62] See Sarah Kreps, "The Role of Technology in Online Misinformation" (June 2020) *Foreign Policy at Brookings*, online: https://www.brookings.edu/wp-content/uploads/2020/06/The-role-of-technology-in-online-misinformation.pdf at 4.

[63] Bobby Chesney and Danielle Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security" (2019) 107 CLR 1753.

[64] Tom Simonite, "A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be" (March 17, 2022), *Wired,* online: https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/.

[65] Lisa Fazio, "Out-of-context photos are a powerful low-tech form of misinformation" (February 14, 2020) *The Conversation*, online: https://theconversation.com/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation-129959.

[66] *Ibid.*

[67] See CSIS, *supra* note 29 at 23.

[68] Bradshaw *supra* note 27 refers to cyber troops: "government or political party actors tasked with manipulating public opinion online" at 2.

[69] Report of the Select Committee on Intelligence, United States Senate, on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election.

[70] Bradshaw, *supra* note 27 at 9. Samuel Woolley talks about this as a key part of computational propaganda. One question is the line between this and marketing. Woolley describes these practices as "manufacturing consensus", and although less harmful, includes e.g. fake likes on a client post. More insidious are practices of paying trolls to harass journalists or use bots to boost things like anti-vaccine content or pay influencers to spread political messages: Woolley, *supra* note 61 at 2.

part of a disinformation campaign in targeting and undermining any skeptics of the truth of the information.[71]

- Hybrids of the above, such as human driven disinformation campaigns using some automated tools. A campaign often uses multiple techniques, such as creation of fake accounts (impersonation or hacked/stolen credentials), use of bots, manual posts, paid advertising to micro-target users and so on.[72] For example, those seeking to disrupt first analyze points of social and political division, the groups occupying particular perspectives in those debates and the types of content that would be most polarizing. They then select tools to generate and distribute this polarizing content, often using artificial intelligence (AI) tools that can work at scale.[73]
- Virtual reality. The performance art of short TikTok videos has become an effective vector for the spread of disinformation.[74] The new battleground for information manipulation is virtual reality. Metaverse's immersive world has already been home to harassment, mis- and dis-information and hate speech.[75]

A challenge with mis- and dis-information is that it often entails activities across multiple platforms. For example, the pattern of online behaviour after mass shooting events follows one of seeding, amplifying, and spreading. The theories are planted in less visible forums like Reddit, 4Chan and Discord, then amplified to more mainstream platforms like Twitter and Facebook, and then spread to the wider ecosystem that interacts with these platforms.[76] This same pattern of harnessing influencers is evident in any disinformation campaign, such as finding alignment with groups with similar ideologies to support and amplify a cause,[77] or a message becoming de-contextualized as it is posted across different platforms.[78] Cast widely, extremist content and other online harms are commonly posted across various platforms, creating a challenging environment for regulation. For example, the Buffalo Attacker explored extremist content on 8chan, wrote a diary on a private Discord server which he later invited users to join, posted a manifesto to Google docs and then to 8chan "moe" and 4chan, livestreamed the attack on Twitch, which was then copied and posted or linked to on other social media.

[71] Wardle and Derakhshan, *supra* note 27 at 31-32.

[72] Bradshaw, *supra* note 27 at 2, 11. Wardle and Derakhshan, *supra* note 27 at 38.

[73] Kreps, *supra* note 62 at 4.

[74] It was recently called "WarTok" for the disinformation spread concerning the Russia Ukraine war: Alex Cadier *et al*, "WarTok: TikTok is feeding war disinformation to new users within minutes — even if they don't search for Ukraine-related content" (March 2022) *NewsGuard*, online: https://www.newsguardtech.com/misinformation-monitor/march-2022/.

[75] Jillian Deutsh *et al*, "Misinformation Has Already Made Its Way to the Metaverse" (December 15, 2021) *Bloomberg,* online: https://www.bloomberg.com/news/articles/2021-12-15/misinformation-has-already-made-its-way-to-facebook-s-metaverse. See this study: Adrian Verhulst *et al*, "Impact of Fake News in VR compared to Fake News on Social Media, a pilot study" (May 2020) IEEE *Xplore*, online: https://ieeexplore.ieee.org/document/9090558.

[76] CSIS, *supra* note 29 at 17.

[77] Bradshaw, *supra* note 27 at 9.

[78] Thakur and Hankerson, *supra* note 32 at 9.

The Convoy movement was similarly cross-platform and the campaign human-driven. What is not known is the extent to which any automated tools were used, images or videos manipulated, advertisements purchased and targeted, and any monetization by key influencers who financially benefitted from amplification of their content.

C is for harmful content

Content is most often the target of regulation, in part, because regulating the actors and deceptive behaviours is more difficult and because harmful content is what is visible. Camille Francois argues that to be effective, regulation should focus more on AB and less on C.[79] It is a strong argument that technical, legal and policy solutions should better target actors, behaviours and distribution. However, one can never escape an analysis of content, because ultimately a post is generated and this implicates the right to freedom of expression. To put it another way, there is always a free expression element to the regulation of any activities on social media. Questions raised include:

- Do users have a right to freedom of expression on social media? This includes the right to seek, receive and impart information and ideas.
- Does social media have a right to freedom of expression that is implicated?
- Is the content posted potentially *illegal* in some form, whether hate propaganda, terrorist propaganda, defamation, or an invasion of privacy, and so on? I use that language loosely as it does not reflect the steps in a legal analysis. Rather, I offer it here to remind the reader that ultimately the content posted might be unlawful, whether criminally and civilly. It may be a different matter to ask who might be liable for the content, particularly if the content is generated by an algorithm.
- If the focus of regulation is on the actors, behaviour, or methods of distribution, does this indirectly regulate expression and what is the legality of this?[80]

Current law and governance are primarily focused on content regulation and explored in Parts II and III.

D is for distribution

Alexandre Alaphilippe added D to this framework to refer to distribution. As he explains, the distribution of disinformation depends on the platform's architectural design. Recommender systems and paid advertising play important roles in the spread and monetization of mis- and dis-information.[81] This fits with the concept of computational propaganda discussed above, that

[79] Francois, *supra* note 33.

[80] This has been a focus of my research recently. See also Daphne Keller, "Amplification and its Discontents" (June 8, 2021) *Knights First Amendments Institute at Columbia University*, online: https://knightcolumbia.org/content/amplification-and-its-discontents.

[81] Alexandre Alaphilippe, "Adding a D to the ABC disinformation framework" (April 27, 2020) *Brookings Institute*, online: https://www.brookings.edu/techstream/adding-a-d-to-the-abc-disinformation-framework/.

if one can game the system of distribution through use of automation and algorithms, it is an effective attack vector to manipulate public opinion.[82] This is also the reason why some argue that the key to addressing disinformation is regulation of business models.[83]

Critical to the spread of disinformation is the design of the platforms, which determine the posts that are recommended and advertised to users, and the content that is thereby amplified.[84] Nathalie Maréchal and her co-authors at the New America Foundation identify two types of algorithms: content-shaping and content-moderating algorithms.[85] Content-shaping algorithms determine the content that users see when they use a company's services. It might be the Newsfeed on Facebook, or recommender system on YouTube or TikTok's ForYou page. It also includes the advertising that micro-targets users. Some law reform proposals have focused on mandating neutrality of content shaping algorithms.[86] Neutrality is a red herring. Content curation is key to manage incident response and demote or remove harmful messages, such as hateful posts or eating disorder content, and to target advertising to user preferences.[87] Rather, algorithmic accountability through transparency reporting, researcher access to data to monitor compliance, mandatory third-party audits, and creation of a regulator, are all better routes to improve conduct, transparency and trust in the systems.[88]

The other type of algorithm is content moderation. As Tarleton Gillespie comments, content moderation is not ancillary to the functioning of platforms, but rather their defining feature.[89] Content moderation algorithms analyse content to determine if a post violates the terms and conditions of that service and decide whether to action that content, with or without human reviewers in the loop.[90] Content moderation will be examined in Part III.

A question concerning the Convoy is how the recommender systems and other design features of social media shaped what users saw, the extent that advertising was purchased on Convoy-related matters, by whom, and the metrics used to micro-target users. While content

[82] See Bradshaw, *supra* note 27; Woolley, *supra* note 61.

[83] Nathalie Maréchal *et al*, "Getting to the Source of Infodemics: It's the Business Model" (May 2020) *New America Foundation*, online: https://www.newamerica.org/oti/reports/getting-to-the-source-of-infodemics-its-the-business-model/. But see interview with Jonathan Stray by Evelyn Douek and Quinta Jurecic, "What We Talk About When We Talk About Algorithms" *The Lawfare Podcast*, online: https://www.lawfareblog.com/lawfare-podcast-what-we-talk-about-when-we-talk-about-algorithms.

[84] Maréchal *et al*, *supra* note 83: "We cannot clean up downstream pollutants like misinformation or dangerous speech without tackling upstream processes - targeted advertising and algorithmic systems - that make this speech so damaging to our information environment in the first place": at 10.

[85] *Ibid*.

[86] For example, the *Protecting Americans from Dangerous Algorithms Act*, H.R. 2154, 117th Cong. (2021-2022) proposes an exception to immunity from liability for social media (see discussion in Part III Legal Overview) if algorithms are used to curate content, unless it is done in a way that is obvious, understandable, and transparent.

[87] See Keller, *supra* note 80. Advertising and marketing are premised on the idea that it wants to get to know consumer preferences and advertise to those preferences: Douek and Jurecic interview, *supra* note 83.

[88] See Part III, Law Reform.

[89] Tarleton Gillespie, *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media* (Yale University Press, 2018) at 21.

[90] Maréchal *et al*, *supra* note 83.

moderation is explored in Part III, questions frontloaded here include, the extent of automated and human moderation, the number and types of content actioned, how quickly, the number of complaints, and reasons for decisions.

The Psychology and Dangers of Information Manipulation

The study of the impact of mis- and dis-information, and therefore its dangers, is challenging. Without consensus on definitions or taxonomy, or the specific problem being studied, it is difficult to measure the impact. The datasets tend to be one-offs and small in scale, or the data is diffuse across various locations, or draws from various disciplines for which it is difficult to reconcile.[91] The result is that "disinformation is often linked to broad goals, the impacts may be diffuse and not targeted, making it harder to find evidence of harm."[92]

In studying the impact, it is important to avoid falling into the trap of the hypodermic needle theory, that users are passive receivers of messages injected by the media.[93] This theory has been refuted but still holds sway. Research is ongoing, but studies have found that false information can impact mental health, leading to stress, fatigue, anger and panic.[94] Exposure to disinformation can lead to belief echoes, meaning that a person knows the information is false but their attitudes are nonetheless shaped by it.[95] And users with conservative ideologies are more likely to follow disinformation accounts than liberal users.[96] In extremism research, while no causal or direct link can be found between online radicalisation and real world violence, there is a link, what the scholars call decision-shaping, not decision-making.[97]

Some studies point to the disruptive impact of disinformation in undermining faith in institutions and democracy, sowing cynicism and doubt. In one study commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs, researchers found that the degree of impact of disinformation depended on the plurality of

[91] Eleni Kapantai *et al*, "A systematic literature review on disinformation: Toward a unified taxonomical framework," (2020) 23(5) New Media & Society; Duncan J. Watts *et al*, "Measuring the news and its impact on democracy," (2021) 118(15) PNAS at 2-5.

[92] Thakur and Hankerson, *supra* note 32 at 8.

[93] Ahmed Al-Rawi *et al*, "What the Fake? Assessing the extent of networked political spamming and bots in the propagation of #fakenews on Twitter" (2019) 43(1) Online Information Review at 65.

[94] Yasmin Mendes Rocha *et al*, "The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review" (2021) 43(2) Journal of Public Health: From Theory to Practice.

[95] Emily Thorson, "Belief Echoes: The Persistent Effects of Corrected Misinformation," (2015) 33(3) Political Communication.

[96] Frederik Hjorth and Rebecca Adler-Nissen, "Ideological Asymmetry in the Reach of Pro-Russian Digital Disinformation to United States Audiences," (2019) 69(2) Journal of Communication, online: https://doi.org/10.1093/joc/jqz006 at 169-170.

[97] Ghayda Hassan *et al,* "Exposure to Extremist Online Content Could Lead to Violent Radicalization: A Systematic Review of Empirical Evidence" (July 2018) 12(7) International Journal of Developmental Sciences 1, online: https://www.researchgate.net/publication/326384034_Exposure_to_Extremist_Online_Content_Could_Lead_to_Violent_Radicalization_A_Systematic_Review_of_Empirical_Evidence. See also Craig Forcese and Kent Roach, "Criminalizing Terrorist Babble: Canada's Dubious New Terrorist Speech Crime" (2015) 53(1) Alberta L Rev 35.

media and how organized the disinformation campaign was.[98] Another study identified the way that this doubt can be used strategically. Spencer McKay and Chris Tenove studied the 2016 US election and explained that Russian operatives focused on undermining credibility in institutions, in part, through creation of faux institutions and competing narratives.[99] Often the dangers of mis- and dis-information are linked to broader human rights values, and the injury caused to human dignity, autonomy, freedom of expression and opinion, and privacy.[100]

Two psychological biases have been identified as key to believing and spreading false information. First, confirmation bias means that we seek out information that reinforces existing beliefs, and that we interpret information to align with our beliefs. Second, motivated cognition and information processing means that we tend to seek out information that reinforces our cultural outlook. The result is that sources that reinforce our pre-existing world views are interpreted as more credible.[101]

Wardle and Derakhshan identify four characteristics that make messages resonate the most with users: (1) it prompts an emotional response (2) it is visual (3) the narrative is strong and (4) repetition.[102] Social media can tick all these boxes. They enable the sharing of the kind of emotional content that is so appealing to users. It is well documented that engagement with our posts acts like a dopamine hit, encouraging sharing of posts that conform with the majority and will be more likely to be liked and shared.[103] Further, obtaining news through social media feeds the ritualistic act of communication; that we read news not to obtain new information, but because we like the ritual, like "attending a church service". We are not there to obtain new information, but to reinforce our beliefs.[104] Engagement online is relational. One reason that these messages are believed is that stories posted online draw from multiple sources, so the focus of readers is not on the source to assess credibility, but rather on the story itself, and credibility is then determined by their networks endorsing the stories.[105] Further, repetition is easy online and the more repetitive the message, the more likely it is that it will be believed.[106]

[98] Judit Bayer *et al*, "Disinformation and propaganda – impact on the functioning of the rule of law in the EU and its Member States" (2019) the *European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs*.

[99] Spencer McKay and Chris Tenove, "Disinformation as a Threat to Deliberative Democracy," (2020) 74(3) Political Research Quarterly, online: https://doi.org/10.1177/1065912920938143.

[100] Bayer *et al* go in this direction, *supra* note 98 at 76, and it is evident in the analysis in Part II on freedom of expression.

[101] Rebecca K Helm and Hitoshi Nasu, "Regulatory Responses to 'Fake News' and Freedom of Expression: Normative and Empirical Evaluation" (February 2021) 21 Human Rights Law Review 302 at 305-306.

[102] Wardle, *supra* note 27 at 38-39

[103] *Ibid* at 13.

[104] *Ibid* at 14-15.

[105] *Ibid* at 12. Wardle also later discusses research that people are more likely to believe a message to be true if it is from someone they know: at 50.

[106] See discussion *ibid* at 45-56. The authors discuss six mental shortcuts used to evaluate whether a message is credible: (1) reputation (familiarity); (2) endorsement; (3) consistency (repetition); (4) expectancy violation (whether website looks and acts as expected); (5) self-confirmation (confirmation of beliefs); (6) persuasive intent (intention of the message creator): at 45.

Visuals are an effective form of information manipulation. People are more likely to believe a statement is true if it is accompanied by an image because it provokes an emotional response,[107] and it can alter the memory of the news e.g. an event accompanied by a dramatic image will impact the memory of the event.[108] Memes are particularly effective, because in addition to the image and short, easily consumable message, they are often humorous. Humour is the playbook for sharing extremist content online and avoiding the moderator, because it "masquerade[s] as medium-specific parody."[109] For example, the Daily Stormer style guide recommended the use of coded words and humour to spread their message.[110] Memes are a way to create social belonging and to mock outsiders for taking it too seriously. Blyth Crawford explains:

> Thus, these memes profit from the inherent ambiguity of online interactions, as is outlined in Poe's Law, creating what Milner has termed a "Logic of Lulz" where it is never possible to discern the intended tone of an online post with any certainty, thereby rendering all participants of an online space perpetually vulnerable to trolling. This way, extreme views are allowed to thrive as memes, enriched by a surrounding culture of troll sensibility and ambiguity.[111]

The ambiguity is strategic. As Alice Marwick explains, "[a]mbiguity is, itself, a strategy; it allows participants to dissociate themselves with particularly unappetizing elements while still promoting the overall movement."[112] Thus, white supremacy messages are pushed through the lens of irony and meme culture, both by alternative media or in groups online. Meme culture has infiltrated warfare, with discussion of "memetic warfare", referencing the "social media battlefield" over "narratives, ideas and social control."[113] Memes are thus part and parcel of information operations to win a competitive advantage over the adversary. Memes and trolling are essentially the new forms of propaganda in warfare.[114]

[107] Eryn J Newman *et al*, "Nonprobative photographs (or words) inflate truthiness" (August 2012) 19 Psychonomic Bulletin & Review 969, as discussed in Fazio *supra* note 61.

[108] Fazio, *supra* note 65. The thinking is that visuals are more easily retrieved from memory, make it easier to imagine an event, serve as proof of the event and get our attention.

[109] Blyth Crawford, "The Influence of Memes on Far-Right Radicalisation" (June 9, 2020), Centre for Analysis of the Radical Right, online: https://www.radicalrightanalysis.com/2020/06/09/the-influence-of-memes-on-far-right-radicalisation/.

[110] Andrew Marantz, "Inside the Daily Stormer's Style Guide" (January 15, 2018) *New Yorker*, online: https://www.newyorker.com/magazine/2018/01/15/inside-the-daily-stormers-style-guide.

[111] Crawford, *supra* note 109. And see Ryan M. Milner, "FCJ-156 Hacking the Social: Internet Memes, Identity Antagonism, and the Logic of Lulz" The Fiberculture Journal, online: http://twentytwo.fibreculturejournal.org/fcj-156-hacking-the-social-internet-memes-identity-antagonism-and-the-logic-of-lulz/. Poe's law is that it is impossible to know if something is a joke.

[112] Marwick and Lewis, *supra* note 35 at 11.

[113] Jeff Giesea, "It's Time to Embrace Memetic Warfare" online: https://stratcomcoe.org/pdfjs/?file=/publications/download/jeff_gisea.pdf?zoom=page-fit at 69.

[114] *Ibid:* "Memetic warfare can be useful at the grand narrative level, at the battle level, or in a special circumstance. It can be offensive, defensive, or predictive. It can be deployed independently or in conjunction with cyber, hybrid, or conventional efforts" at 69.

A richly explored question is the existence and impact of filter bubbles and echo chambers on users.[115] This is the concept that our online experiences are now personalized and trap us in a chamber where we hear, read, and interact with the same people and views. Our newsfeed on Facebook is personalized. We select the rooms to join on the audio app Clubhouse. We select the individuals and entities we follow on Twitter. We message with individuals and groups of our choosing on Telegram. We thus live in a bubble of our making and our views are never challenged or expanded. However, scholarship on the filter bubble is mixed. Several scholars now argue that the phenomenon has been seriously overstated.[116] Axel Bruns, for example, posits that the filter is not from failing to see content that opposes our worldview, but rather the filter in our head, which leads individuals to take an oppositional stance to information.[117]

While the impact of information manipulation in the Convoy may be difficult to asse ss, the nature of the content that was influencing supporters can be analyzed. There was an existing audience for many of Convoy influencers discussing themes of anti-vaccination and anti-lockdown.[118] A question is the extent to which false information was shared in these spaces, and if any of it was intentionally shared (disinformation) to audiences that believed and re-shared it (misinformation). Another question is the extent that there was content in the "everything else" bucket of hatred, extremism, doxing, harassment and so on.[119] Some of the accounts and influencers of the Convoy were reported to have ties to far-right groups.[120]

Is there a solution to online information manipulation? Law and governance are complicated, and briefly explored in Parts II and III. There seems to be consensus that the solution is multi-dimensional, involving multiple strategies that, together, counteract and manage the risks of

[115] Axel Bruns defines the echo chambers as "when a group of participants choose to preferentially *connect* with each other, to the exclusion of outsiders" and filter bubbles as "when a group of participants choose to preferentially *communicate* with each other, to the exclusion of outsiders": "Filter Bubble" (November 29, 2021) Internet Policy Review, online: https://policyreview.info/concepts/filter-bubble.

[116] See e.g. Axel Bruns, *ibid*; Amy Ross Arguedas *et al*, "Echo chambers, filter bubbles, and polarisation: a literature review" (January 19, 2022) Reuters Institute and University of Oxford, online: https://ora.ox.ac.uk/objects/uuid:6e357e97-7b16-450a-a827-a92c93729a08.

[117] Bruns, *supra* note 115. The question, he asks, is "why and how different groups in society come to develop such highly divergent personal readings of the same information."

[118] *Fifth Estate*, *supra* note 3.

[119] Doxing is problematic from a legal, ethical and cybersecurity perspective, and it was used by those for and against Convoy. There was a data breach of GiveSendGo and donors were doxed: Tanya Basu, "Online activists are doxxing Ottawa's anti-vax protesters" (February 11, 2022) *MIT Technology Review,* online: https://www.technologyreview.com/2022/02/11/1045281/ottawa-antivax-protests-doxxing/. See "Letter sent to parliamentarians warning of doxing ahead of trucker convoy: 'Go somewhere safe'" (January 28, 2022) *City News Ottawa*, online: https://ottawa.citynews.ca/local-news/letter-sent-to-parliamentarians-warning-of-doxing-ahead-of-trucker-convoy-go-somewhere-safe-5002917.

[120] Broderick, *supra* note 7.

harm of information manipulation.[121] For example, education is a key "inoculation"[122] to help the public identify false information, dubious sources, synthetic accounts and so on.[123] Transparency is also important. Social media reports about advertising, content moderation and privacy, for example, enable users to evaluate the veracity of what they consume. Education and transparency only work if there are trusted media sources. Therefore, supporting a diverse and sustainable media is crucial to combatting the effect of disinformation.[124] Technology is also core to combatting information manipulation,[125] such as identifying, flagging, demoting or otherwise limiting the visibility or virality of content, although they are not (nor ever will be) perfect instruments of regulation. Technical solutions are explored in Part III.

Part II Freedom of Expression and User Rights and Responsibilities

The remainder of this paper will explore three legal and governance angles to the question of regulation of information manipulation. The first concerns the right to freedom of expression. As noted above, policy has shifted from a narrow focus on content regulation, to regulating actors, behaviours, and methods of distribution. This wider scope is important, but free expression is a relevant legal conundrum regardless of the regulatory strategy undertaken. A shift away from content regulation only makes the analysis indirect rather than direct. The second concerns the responsibilities of users for the spread mis-, dis- and mal-information. The third focuses on the legal obligations of social media and content moderation. The latter is a space of tremendous legal change, globally and in Canada.

Freedom of Expression

Central to any discussion of information manipulation is freedom of expression. It is examined here narrowly as it relates to content regulation. Two issues emerge. First, regulating disinformation tasks a court or decision-making body to label a message as disinformation. Courts are fact-finding bodies and so in principle labeling content as true or false is not a hurdle, and many areas of law that intersect with freedom of expression might entail a finding of truth. For example, truth is a defence to a defamation claim and a court would be tasked with determining whether the defendant has met the burden of establishing truth on a balance of probabilities, if pled.

[121] Report of the independent High level Group on fake news and online disinformation, "A multi-dimensional approach to disinformation" (2018) *European Commission*, online: https://coinform.eu/wp-content/uploads/2019/02/EU-High-Level-Group-on-Disinformation-A-multi-dimensionalapproachtodisinformation.pdf at 4. The expert advisory group on online harms, appointed by Heritage Canada, discussed the importance of a multi-dimensional approach to tackling information manipulation: see worksheets online: https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content.html.

[122] Helm, *supra* note 101 at 318.

[123] Kreps, *supra* note 62 at 6; European Commission, *supra* note 121 at 25-27.

[124] The European Commission identifies the need for the state to protect freedom of expression, a free press, and media pluralism to address media challenges: *ibid* at 29.

[125] See Kreps, *supra* note 62 at 6-7.

However, the line between truth and falsity can be more complicated in the sphere of information manipulation. Wardle and Derakhshan's taxonomy of information disorder evidences this dilemma. In their taxonomy of types of mis- and dis-information, they list satire or parody, misleading content, imposter content, fabricated content, false connection, false context and manipulated content.[126] Albert Zhang and co-authors give the example of a post by a spokesperson for China's foreign ministry of an Australian soldier holding a knife to a child purportedly as commentary about Australia's war crimes inquiry in 2020. Australia's Foreign Minister denounced the image as disinformation. The image was artwork and not necessarily created to deceive.[127]

This connects to a second issue with labelling content as disinformation. All definitions of disinformation hinge on the intention to harm (in contrast to misinformation). Proving intention to harm (and to harm what?) is a difficult exercise as it requires the motivations of the actors to be discerned, which is particularly challenging in light of the strategic ambiguity of many posts.[128] Further, the road that leads to the creation of a disinformation campaign may be based on true information and honestly held opinions.[129] Information labelled as false may later be found to have a kernel or possibility of truth, such as the COVID-19 lab leak theory that was initially dismissed by the scientific community, and posts removed by many social media platforms, and later investigated by the U.S. Secret Service at the direction of President Biden.[130]

The second issue is more fundamental to free expression itself: namely what does it mean in law to say that we value and protect freedom of expression and how does this intersect with information manipulation?[131] Under international human rights law, freedom of expression is protected under Article 19 of the International Covenant on Civil and Political Rights (ICCPR),[132] to which Canada is a party:[133]

Article 19

[126] Wardle and Derakhshan, *supra* 27 at 17.

[127] Alberta Zhang *et al*, "Submission to the UN Special Rapporteur on disinformation and freedom of opinion and expression", online: https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/disinformation/2-Civil-society-organisations/Australian-Strategic-Policy-Institute.pdf, at 3-4.

[128] See discussion above about memes in Marwick and Lewis, *supra* note 235 a 11.

[129] Ryan Calo *et al*, "How do you solve a problem like misinformation?" (2021) 7(5) Science Advances at 1.

[130] See, Stephan Lewandowsky, "The Lab-Leak Hypothesis Made It Harder for Scientists to Seek the Truth" (March 1, 2022) *Scientific American,* online: https://www.scientificamerican.com/article/the-lab-leak-hypothesis-made-it-harder-for-scientists-to-seek-the-truth/.

[131] See Khan, *supra* note 30 at Part B.

[132] *Supra* note 40. The Universal Declaration of Human Rights, 1948 (UDHR) is the anchor of international human rights, and its Article 19 provides: "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."

[133] See Government of Canada, "Reports on United Nations human rights treaties", online: https://www.canada.ca/en/canadian-heritage/services/canada-united-nations-system/reports-united-nations-treaties.html.

> 1. Everyone shall have the right to hold opinions without interference.
> 2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.
> 3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:
> (a) For respect of the rights or reputations of others;
> (b) For the protection of national security or of public order (ordre public), or of public health or morals.[134]

Freedom of expression is protected in the *Canadian Charter of Rights and Freedoms* (*Charter*)[135] under s. 2(b):

> 2 Everyone has the following fundamental freedoms:
> ---
> (b) freedom of thought, belief, opinion and expression, including freedom of the press and other media of communication;[136]

Freedom of expression is a fundamental right in a democratic society and central to our search for truth, democracy, and self-discovery and fulfilment.[137] It is a fulsome right, including the right to seek, receive and impart information and ideas regardless of frontiers[138] and it includes the right not to express oneself.[139] The *Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda* (*Joint Declaration*),[140] confirms that the right is not limited to correct statements, and that it includes the right to "shock, offend and disturb".[141] Any limitation on the right to freedom of expression must comply with the test under international law and embedded in s. 1 of the *Charter*, that the restriction is provided by law, services a legitimate aim and is necessary and proportionate to that interest.

A few aspects of Article 19 are notable as it relates to information manipulation. First, the right to freedom of expression is the only human right in the ICCPR that carries with it "special duties

[134] Article 19, ICCPR, *supra* note 40.

[135] Part 1 of the *Constitution Act*, 1982, being Schedule B to the *Canada Act 1982* (UK), 1982, c 11, s. 8.

[136] *Ibid* at s. 2(b). Justifiable limits of s. 2(b) are set out in s. 1.

[137] The Supreme Court of Canada commented: "[f]reedom of expression is seen as worth preserving for its own intrinsic value": *R v Keegstra*, [1990] 3 SCR 697 at 881. There is significant debate about the meaning and value of freedom of expression, but they are not explored in this paper.

[138] ICCPR, *supra* note 40, Article 19.

[139] Khan, *supra* note 30 at para 35.

[140] (2017) FOM.GAL/3/17.

[141] *Ibid.* They are drawing from oft-quoted *Handyside v UK*, [1976] ECHR 5 *(*and quoted in *Irwin Toy Ltd. v Québec (Attorney General)*, [1989] 1 SCR 927). It is worth noting that in *Handyside,* the Court goes on to state: "[s]uch are the demands of that pluralism, tolerance, and broadmindedness without which there is no "democratic society". This means, amongst other things, that every "formality", "condition", "restriction" or "penalty" imposed in this sphere must be proportionate to the legitimate aim pursued." In Canada, the quote is often from *Irwin Toy*: freedom of expression "ensure[s] that everyone can manifest their thoughts, opinions, beliefs, indeed all expressions of the heart and mind, however unpopular, distasteful or contrary to the mainstream" at 968.

and responsibilities".[142] While discussions often centre on limitations to the right, it is significant that the ICCPR emphasizes the unique rights *and responsibilities* that are the building blocks of the right to free expression. Second, the right to hold opinions is an absolute right in the ICCPR.[143] In practice, our thoughts and opinions are influenced by all kinds of people and media. Marketing and advertising, for example, are designed to influence our consumer behaviour. The question is the line between legitimate and unlawful forms of manipulation.[144] Arguably disinformation campaigns unjustifiably interfere with one's autonomy to form an opinion free from manipulation, and social media surveillance and profiling implicate the right not to reveal one's thoughts.[145] Third, the rights of everyone, including individuals that believe and share misinformation, are undermined by disinformation campaigns. As the *Joint Declaration* affirms, disinformation interferes with various aspect of the right to free expression, including the right to know, to seek, receive and impart information and ideas.[146] Disinformation can cause harm to individual reputations and privacy, and to national security, which can be the basis for legitimate restriction of the right to free expression.[147] It can also advocate hatred that incites violence, discrimination or hostility, prohibited in Article 20 of the ICCPR.[148]

Misinformation as a target of regulation is particularly problematic. Rumours and gossip are part of the rituals of human interaction.[149] For those that are innocent receivers and distributors of such information, they are arguably engaged in the search for truth, one of the philosophical values that underpin the right to free expression.[150] There are many reasons to question the search for truth as sufficient foundation to protect freedom of expression in these circumstances. It is premised on the idea that by leaving it to the marketplace of ideas, the truth will surface.[151] It also fails to account for the unequal burden experienced by marginalized and racialized groups. This theory also assumes that we have equal access to and experiences with free expression, and studies show that women, racialized and LGBTQ+ individuals, in

[142] Article 19, ICCPR, *supra* note 40. Francesca Klug illuminated this point in an address she gave to Intelligence Squared and the London Jewish Cultural Centre public debate Royal Geographical Society "Freedom of Expression Must Include the Licence to Offend" (June 2016) I Religion and Human Rights 225.

[143] Article 19(1), ICCPR, *supra* note 40. Susie Alegre identifies three elements to the right to hold opinions: the right to not reveal one's thoughts or opinion, the right to not have them manipulated, and to not be penalised for one's thoughts: Susie Alegre, "Rethinking Freedom of Thought for the 21st Century" (2017) 3 European Human Rights Law Review 221 at 225; Evelyn Marie Aswad, "Losing the Freedom to be Human" (2020) 52(1) Columbia Human Rights Law Review 306; Khan, *supra* note 24.

[144] Alegre, *supra* note 143 at 227.

[145] Khan, *supra* note 30 at para 34-36. Alegre, *supra* note 143 at 225.

[146] Joint Declaration, *supra* note 140.

[147] See ICCPR, Article 19(2), *supra* note 40.

[148] See Joint Declaration, *supra* note 140. As Jones explores, *supra* note 37, Article 20 shows that disinformation is not new and concerns about its widespread use in World II were addressed in the ICCPR, at 41.

[149] Robert Post, "The Social Foundations of Defamation Law: Reputation and the Constitution" (1986) 74 Cali L Rev 691.

[150] The other main theories embraced by the Supreme Court of Canada are self-fulfilment and democracy: see e.g. *Irwin Toy Ltd v Québec (Attorney General),* [1989] 1 SCR 927.

[151] See John Stuart Mills, *On Liberty* (1859) and the dissent of Justice Holmes in *Abrams v US* (1919) 250 U.S. 616: "the best test of truth is the power of the thought to get itself accepted in the competition of the market".

particular intersectional individuals, are the primary targets of abuse, and driven from participation online.[152] In short, the right to freedom of expression is a right that is often only fully enjoyed by privileged groups.

This is not to say that freedom of expression should be undermined to protect individuals from offence. A properly run system of free expression expects us to put up with a lot based on the ideal of free expression. And the free flow of information is central to the right.[153] However, there is more to the analysis than putting up with offence. Harm facilitated by disinformation and through social media impacts the right to equality, including equality of expression.[154] For example, researchers identified the use of racially targeted disinformation campaigns in the US to suppress voter turnout from communities of colour.[155] Another example is the use of memes to spread extremist ideologies by playing on humour and familiar racist tropes, which then shifts the boundaries of acceptable discourse and normalizes and embraces racism.[156]

It is difficult to design a human rights compliant law that targets the creators of disinformation. Any right must be broadly enjoyed, and exceptions narrowly construed.[157] Several laws in other jurisdictions exemplify the risks in passing broadly scoped laws that prohibit disinformation.[158] They risk incentivizing systems of content filtering or takedown, including internet shutdowns, and enable state control and removal of dissenting voices (sometimes in places where states also sponsor their own disinformation). Even states with a strong commitment to human rights have faced unintended consequences. Of concern are laws that criminalize disinformation and do not have sufficiently precise definitions of false information and/or harm. Broadly framed laws have been used by Governments against civil society, journalists, and political opponents.[159] Civil laws may be legitimate but must be narrowly tailored such as our defamation laws, where the defendant is given a full suite of defences aimed at protection of

[152] See Jon Penney and Danielle Citron, "When Law Frees Us to Speak" (2019) 87 Fordham Law Review.

[153] Khan, *supra* note 30 at para 38.

[154] See *Keegstra*, *supra* note 137. Dickson C.J. for the majority reasoned:

> Indeed, expression can be used to the detriment of our search for truth; the state should not be the sole arbiter of truth, but neither should we overplay the view that rationality will overcome all falsehoods in the unregulated marketplace of ideas. There is very little chance that statements intended to promote hatred against an identifiable group are true, or that their vision of society will lead to a better world. To portray such statements as crucial to truth and the betterment of the political and social milieu is therefore misguided.

See also Cynthia Khoo "Deplatforming Misogyny" (2021) LEAF, online: https://www.leaf.ca/wp-content/uploads/2021/04/Full-Report-Deplatforming-Misogyny.pdf.

[155] Thakur and Hankerson, *supra* note 32 at 10-11, and for further examples.

[156] Crawford, *supra* note 109.

[157] Khan, *supra* note 30 at para 39.

[158] See Ruth Levush, "Government Responses to Disinformation on Social Media Platforms" (2019) Library of Congress, online: https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1180&context=scholcom.

[159] See discussion Joint Declaration, *supra* note 140 at 53-54. The Joint Declaration goes so far as to say that criminal defamations laws should be abolished at para 2(b), which puts Canada out of step with international human rights: *Criminal Code*, *supra* note 40, ss. 297-316; *R v Lucas,* [1998] 1 SCR 439.

freedom of expression, including truth, fair comment, and responsible communication in the public interest.[160]

Canadian Disinformation Laws

There are a variety of different laws that apply to individuals who communicate false statements. However, there are no laws that directly target individuals who communicate mis- or dis-information as contemplated and explored in this paper.

In 1992, in *R v Zundel*,[161] the Supreme Court of Canada (SCC) struck down the provision in the *Criminal Code* that prohibited the spread of false news. Section 181 of the *Criminal Code* provided:

> 181. Every one who wilfully publishes a statement, tale or news that he knows is false and that causes or is likely to cause injury or mischief to a public interest is guilty of an indictable offence and liable to imprisonment for a term not exceeding two years.[162]

In a split 4/3 judgment, the majority held that s. 181 infringed the right to freedom of expression in s. 2(b) of the *Charter* and was not justified under s. 1. The majority emphasized the severity of criminal sanctions and the importance of "liberty of speech".[163] They viewed the criteria that the false statement causes or likely caused "injury or mischief to a public interest" as vague and overbroad, posing great danger to minority groups and their full participation in society. A key split between the majority and minority was how to characterize false expression. The majority stated the provision required a court to decide the meaning that was to be judged true or false: "[d]ifferent people may draw from the same statement different meanings at different times."[164] In their view, truth is a matter of perception, and prohibiting the spread of false news would enable dominant groups to impose their perception of truth on the minority. The dissenting judgment framed false information differently, concluding that there are provable facts and that the criminal provision was narrowly concerned with deception. In their view, the intention to deceive through the sharing of provably false and harmful information undermines the value of free expression.[165]

Despite *Zundel*, the Supreme Court upheld as constitutional criminal and civil defamation laws[166] because false information obstructs the search for truth and does not enjoy the same level of protection as political speech, although later broadened the civil defence for matters of

[160] *Ibid*. See *WIC Radio v Simpson,* 2008 SCC 40 and *Grant v Torstar*, 2009 SCC 61. For country specific examples, see Daniel Funke and Daniela Flamini, "A guide to anti-misinformation actions around the world" *Poynter*, online: https://www.poynter.org/ifcn/anti-misinformation-actions/

[161] *R v Zundel,* [1992] 2 SCR 731.

[162] *Criminal Code*, *supra* note 40.

[163] *Zundel*, *supra* note 161.

[164] *Ibid*.

[165] *Ibid*.

[166] *Lucas*, *supra* note 159.

public interest.[167] These cases wrestle with the spectrum of low-value expression that can be characterized as far from the core rationales for the protection of expression.[168] There are also other *Criminal Code* provisions that criminalize an aspect of falsity, in particular hate propaganda,[169] counselling terrorism[170] and fraud.[171] There are also several civil causes of action that have an element of falsity, in particular defamation and false light.[172] To the extent that malinformation may be captured by a civil provision, intentional infliction of mental suffering and public disclosure of private embarrassing facts may apply.[173]

In 2018, the Government of Canada amended s. 91(1) of the *Canada Elections Act*[174] to remove the word "knowingly" from a provision that prohibited making or publishing false statements about a candidate's personal character or conduct during an election period. The constitutional challenge of the provision turned on the removal of the term "knowingly" and whether this meant that intention was no longer a requirement of the offence.[175] The Ontario Superior Court concluded that the amendment prohibited the spread of accidental or unknown false information, as in misinformation, and that this was an unjustifiable limit on the right to freedom of expression. The provision was held to be of no force or effect.[176]

Competition law applies to an aspect of disinformation as advertising. The *Competition Act*[177] prohibits false or misleading representations and deceptive practices to promote a product,

[167] *Ibid*; *Hill v Church of Scientology*, [1995] 2 SCR 130; *Grant*, *supra* note 160.

[168] The Supreme Court of Canada has consistently held that not all expression is treated equally, rather the justification for infringement of the right is a spectrum from low to high value expression that contribute to the search for truth, democracy, and self-fulfilment. See e.g. *Keegstra*, *supra* note 137; *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11; *Lucas*, *supra* note 159 ("the negligible value of defamatory expression"); *Hill*, *supra* note 167 ("defamatory statements are very tenuously related to the core values which underlie s. 2(b). They are inimical to the search for truth" at para 106). But then see *Grant*, *supra* note 160, in which the Supreme Court stated that Hill "must be read in the context of that case" (para 57), and in adopting a new defence of responsible communication in the public interest: "The law of defamation currently accords no protection for statements on matters of public interest published to the world at large if they cannot, for whatever reason, be proven to be true. But such communications advance both free expression rationales mentioned above — democratic discourse and truth-finding — and therefore require some protection within the law of defamation" at para 65.

[169] *Criminal Code*, *supra* note 40 at s 319.

[170] *Ibid* at s 83.221.

[171] *Ibid* at s 320.

[172] *Yenovkian v Gulian,* 2019 ONSC 7279.

[173] See *Jane Doe 72511 v NM,* 2018 ONSC 6607.

[174] SC 2000, c 9, concerning amendment 2018, c 31, s 61.

[175] *CCF v Canada (AG)*, 2021 ONSC 1224.

[176] See Eve Gaumond, "Why a Canadian Law Prohibiting False Statements in the Run-Up to an Election Was Found Unconstitutional", (March 16, 2021), online: https://www.lawfareblog.com/why-canadian-law-prohibiting-false-statements-run-election-was-found-unconstitutional.

[177] RSC 1985, c C34. See also the *Elections Modernisation Act*, SC 2018, c 31, which required creation of political ad registries.

service, or business interest.[178] This includes, for example, misleading consumers to obtain their data.[179] To the extent that disinformation is for one of these promotional purposes, this may be investigated by the Competition Bureau. For example, false and misleading claims were made about cures to COVID-19, which were investigated by the Bureau.[180] Influencers must also disclose if their posts are sponsored, whether through payments, discounts, free products, and services, and similar.[181]

As the above brief overview shows, there is currently no law in Canada that directly targets mis- and dis-information. In criminal law post-*Zundel*, the prosecution would be driven by another higher order wrong. For example, a false statement that is hate, terrorism or fraudulent. In civil law, a similar dynamic is observable. The conduct might be actionable if the false statement impacts reputation (defamation) or presents an individual in a false light (in Ontario). If malinformation is understood as essentially doxing, then the common law public disclosure of private embarrassing facts may be applicable in some provinces.[182] That there is no law that directly applies to mis- or dis-information may be appropriate. Connecting the conduct to another higher order wrong enables prosecution or a civil action in the most egregious circumstances. However, it is arguable that if *Zundel* were decided today, the result might be different.

Part III Social Media Law and Governance

The question that follows is what law and governance mechanisms are available to hold social media and other online services accountable for harmful content posted using their services? This is a rich area of study, and a detailed analysis is beyond the scope of this paper, although readers are encouraged to review the resources cited in this paper for more information.[183]

Regulation ultimately is concerned with "how much involvement government actually devolves to private actors."[184] There are three types of regulation at play for social media services. First

[178] *Competition Act*, *supra* note 177 at s 52 and Part 74.01. There are both criminal and civil adjudicative regimes. See explanation: Government of Canada, "False or Misleading Representations and Deceptive Marketing Practices", online: https://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/eng/03133.html.

[179] *Ibid*.

[180] Competition Bureau of Canada, "Competition Bureau cracking down on deceptive marketing claims about COVID-19 prevention or treatment" (May 6, 2020), online: https://www.canada.ca/en/competition-bureau/news/2020/05/competition-bureau-cracking-down-on-deceptive-marketing-claims-about-covid-19-prevention-or-treatment.html.

[181] Government of Canada, "Influencer Marketing and the *Competition Act*", online: https://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/eng/04512.html.

[182] *Jones v Tsige*, 2021 ONCA 312. It may also be an invasion of privacy under statute in some provinces.

[183] For a broad overview of Canadian intermediary liability see Emily B. Laidlaw, "Mapping Current and Emerging Models of Intermediary Liability" (2019) prepared for the Broadcasting and Telecommunications Review Panel, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3574727. On content moderation see Evelyn Douek, "Content Moderation as Administration" *forthcoming* 136 Harvard Law Review, draft at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4005326.

[184] Chris T. Marsden, *Internet Co-Regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace* (Cambridge University Press, 2011) at 53.

are the laws that apply to social media in terms of content regulation. The area of law is known as intermediary liability, because of the go-between role of these companies, linking content creators with content consumers. Their role, traditionally, can be understood as facilitative and secondary to that of content creators, and therefore less morally culpable.[185] Often the term 'platform' is now used to refer to intermediaries with particular social or cultural power in the marketplace.[186] Another area of law that impacts online safety is private sector privacy law. User data is the core of social media functionality and profitability, and these companies have privacy obligations to users to protect their personal information. This area of the law is not helpful to address the legality of individual posts but whether social media design and specific data transactions sufficiently protect user privacy.

The second type of regulation is co-regulation, which is a form of government-backed self-regulation, such as codes of practice and industry bodies.[187] Co-regulation is collaborative and helps fill the gap between legal obligation and voluntariness and has been central to internet governance since the internet's commercialization. This type of regulation is only briefly touched on herein. The third type of regulation is self-regulation, or in the case here, content moderation by social media.[188] The absence of federal intermediary liability laws in Canada means that content moderation was the primary regulatory force with the Convoy.

Legal Overview

There is no comprehensive federal intermediary liability law in Canada in contrast to Europe[189] and the United States of America.[190] In the US, section 230 of the *Communications Decency Act* (CDA)[191] provides broad immunity from liability to intermediaries for the content posted by third parties, except for federal criminal law, intellectual property law or electronic communications privacy, and a recent, widely criticized, amendment to address human

[185] The OECD definition of intermediaries is that they "bring together or facilitate transactions between third parties on the Internet": OECD, "Economic and Social Role of Internet Intermediaries" (April 2010), online: https://www.oecd.org/internet/ieconomy/44949023.pdf at 9.

[186] There are several more layers to exploring different types of platforms beyond the scope of this paper . See e.g. Tarleton Gillespie, "Platforms are not Intermediaries", (2018) 2 Geo L Tech Rev. 198 209; José van Dijck, Thomas Poell and Martijn De Waal, *The Platform Society* (Oxford University Press, 2019); Robert Gorwa, "What is Platform Governance?" (2019) 22(6) Information, Communication & Society 854.

[187] There is more to regulation than these categories, although these are the three main ones. See Emily B. Laidlaw, "The Challenge Designing Intermediary Liability Laws" in Catherine Easton and David Mangan, *The Philosophical Foundations of Information Technology Law* (Oxford University Press, forthcoming 2023).

[188] As Chris Marsden comments, self-regulation usually does not exist in pure form as government is often in the shadows pressuring companies to act: *supra* note 184 at 48.

[189] *Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market* (ECD) and *Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*, COM(2020) 825 (DSA).

[190] *Communications Decency Act,* 47 USC 230 (CDA). There are, of course, other jurisdictions to consider, but I mention the US and EU as they are similar legal systems and were two of the first jurisdictions to implement broadly scoped intermediary liability laws.

[191] *Ibid*.

trafficking.[192] The result is that intermediaries have a safe harbour from liability for mis- or dis-information that users post that might be illegal, because it is e.g. defamatory or reveals private information. A few things are key to tease out about s. 230. First, the immunity concerns decisions to both leave content up and take it down. Thus, social media are protected to develop content moderation practices that are stricter than the law and align with their values. The problem with s. 230 is that it does not incentivize responsibility and companies can, and have, embraced the protection of s. 230 for leaving illegal content up without corresponding steps to implement moderation practices.[193]

The EU, in contrast, adopted a conditional safe harbour model with the *Electronic Commerce Directive* (ECD).[194] Under this model, an intermediary that hosts third party content is provided with a conditional immunity from liability for unlawful content posted by users. However, the intermediary risks losing the immunity if it obtains knowledge of the unlawful content and fails to act to disable access to it. Thus, it operates as a notice and takedown regime revolving around *knowledge* of the unlawful activity and an obligation to *action* illegal content by removing it.

For clarity, intermediary obligations under the ECD would only be triggered for unlawful content and a significant portion of information manipulation is lawful but awful. Thus, the European Commission led drafting of a voluntary industry *Code of Practice on Disinformation*.[195] Content removal is important for unlawful content,[196] but conditional safe harbour models are more problematic to deploy than they appear at first blush. They tend to incentivize removal of content without corresponding protection of lawful content or decisions by companies that are more human rights sensitive.[197]

[192] The provision was poorly drafted and backfired. It broadly made vulnerable sex workers, incentivized removal of legal speech and encouraged lack of oversight by platforms of dangerous activities: see Daphne Keller, "SESTA and the Teachings of Intermediary Liability" (November 2, 2017), online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3121296 .

[193] Also, the broad legal protection afforded to free expression under the First Amendment means that content that would be considered hate propaganda in Canada is protected speech in the USA and thus would not invite constitutional or section 230 scrutiny. Section 230 has had global impact, with controversy when it conflicts with laws in other jurisdictions: see for example: *Google Inc v Equustek Solutions Inc*: 2017 SCC 34; *Google Inc v Equustek Solutions Inc*, 2017 WL 5000834 (ND Cal Nov 2, 2017). Then see *Equustek Solutions Inc v Jack*, [2018] BCSC 610; Michael Geist, "The *Equustek* Effect: A Canadian Perspective on Global Takedown Orders in the Age of the Internet" in Giancarlo Frosio, ed., *The Oxford Handbook of Online Intermediary Liability* (Oxford University Press, 2020).

[194] As a Directive, there was significant variation in the ways that it was implemented by Member States, one of the reasons for the DSA: *supra* note 189. See also the *Digital Millennium Copyright Act*, Pub. L. No. 105-304, 112 Stat. 2860 (1998).

[195] See "2018 Code of Practice on Disinformation", online: https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation and "Strengthened Code of Practice on Disinformation" (June 2022), online: https://ec.europa.eu/commission/presscorner/detail/en/ip_22_3664.

[196] But see discussion of effectiveness in Part III, Content Moderation.

[197] David Kaye *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, A/HRC/32/38 (2016) at paras 43-44.

There is a current trend in law reform to shift to a due diligence model for intermediaries, including in Europe which has supplemented the ECD with the *Digital Services Act* (DSA).[198] These are variously framed as duty of care, risk assessment and due diligence. At its core, these models shift from a binary leave up/takedown model to task intermediaries with managing the risks of harm of their services. These have the advantage that they are not limited to content regulation and can be more focused on actor, behaviour, and distribution. There are significant variations in the way these might be implemented, with risks associated with this being poorly done. I am broadly favourable of a risk management approach to intermediary responsibility, although the devil is in the details. Due diligence models will be explored below in the section on Law Reform.

As explained, Canada does not have any federal intermediary liability law similar to the US or Europe. The *Criminal Code* provides an avenue for content takedown. A court may order online content removal of terrorist or hate propaganda, child pornography, voyeurism, and non-consensual disclosure of intimate images.[199] Québec is the only province that has a broad intermediary liability law, which provides a safe harbour on the condition that if an intermediary becomes aware of illicit activity on their services, they act promptly to block access to the content.[200]

Canadian intermediary liability laws relevant to information manipulation have developed primarily in defamation law.[201] In practice, it operates similarly to the ECD as a notice and takedown regime. If TikTok, for example, obtains knowledge that it is hosting a video with defamatory content, it is obligated to disable access to the video or risk liability for the underlying wrong.[202] Defamatory content only captures a fraction of the forms of information manipulation at issue, specifically false information that lowers the reputation of an individual or entity.[203] Mis and dis-information often relate to broad topics, such as health. Also, the

[198] DSA, *supra* note 189.

[199] *Criminal Code*, *supra* note 40, ss 320.1(5), 83.223, 164.1(5).

[200] *Act to establish a legal framework for information technology*, CQLR c C-1.1. Two differences between the ECD and Québec Act are notable. First, the Québec Act refers to illicit activity, a broader concept that unlawful content. Pierre Trudel explains that while this catches lawful but awful content, constitutional constraints mean that it would be narrowly construed. Second, s. 22 provides that the intermediary "may incur responsibility", which means that the analytical test is whether the intermediary behaved diligently in the circumstances: Pierre Trudel, "Liability of Platforms: The Law of Québec" (on file with author) at 2-3.

[201] For intermediary liability laws in copyright see *Copyright Act*, RSC 1985, c C-42, ss 41.25-41.27.

[202] See e.g. *Weaver v Corcoran*, 2015 BCSC 165. On publication see *Crookes v Newton*, 2011 SCC 47. Further, in defamation the intention need only be to distribute the information, with the result that individuals can be unintentionally defamed: *Hulton v Jones*, [1910] AC 20. Therefore, misinformation could be actionable in defamation (intentional distribution of false information you believe to be true), although note th e defences of fair comment and responsible communication in the public interest: *WIC Radio, supra* note 160; *Grant*, *supra* note 160.

[203] *Hill*, *supra* note 167. Various defences are important to broadly protect expression that might nonetheless harm reputation, but this is not explored here. Note that whether an intermediary might be liable under the umbrella of one of the privacy torts is untested in law, although presumably a court would draw from defamation principles.

Government of Canada is limited concerning any intermediary liability laws it can introduce because of its trade commitments in the Canada-US-Mexico Trade Agreement (CUSMA).[204]

Private sector privacy legislation (federally and provincially)[205] are foundational to protection of privacy, and the moral culpability of companies is more direct than the area of intermediary liability. Privacy law requires that organizations are accountable for personal information about an identifiable individual that they collect, use, or disclose in the course of commercial activities.[206] What is complex, given the opacity of many social media business practices, is identifying information flows to nail down precisely what social media collect, use and disclose, and the various third parties with which they transact. This is best exemplified by the *Joint investigation of Facebook, Inc. by the Privacy Commissioner of Canada and the Information and Privacy Commissioner for British Columbia* (Joint Investigation)[207] concerning the Cambridge Analytica scandal. Cambridge Analytica used data from an app *This is Your Digital Life*, which collected data about Facebook users to build psychological profiles of users, which were then used to send targeted ads to users to influence voters in various elections, most notably the US republican presidential nomination race and subsequent election. The Joint Investigation concluded that Facebook (now Meta) violated privacy law, because it failed to be accountable to give meaningful effect to protection of privacy, did not obtain meaningful consent from users, nor had in place adequate security safeguards.[208]

[204] See Article 19.17, Canada-US-Mexico Trade Agreement, online: https://www.international.gc.ca/trade-commerce/trade-agreements-accords-commerciaux/agr-acc/cusma-aceum/index.aspx?lang=eng. Article 19.17 of the CUSMA introduces a broad safe harbour from liability for intermediaries styled on CDA s. 230, although it is a matter of debate whether it goes as far as s. 230. Article 19.17 prohibits treatment as an "information content provider in determining liability", which leaves scope for equitable remedies and the duty of care/risk management models: see Vivek Krishnamurthy *et al*, "CDA 230 Goes North American? Examining the Impacts of the USMCA's Intermediary Liability Provisions in Canada and the United States" (July 7, 2020) CIPPIC, online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3645462. The Government of Canada's "Canadian Statement on Implementation", online: https://www.international.gc.ca/trade-commerce/trade-agreements-accords-commerciaux/agr-acc/cusma-aceum/implementation-mise_en_oeuvre.aspx?lang=eng, indicates that Article 19.17 means that intermediaries are not to be held civilly liable for user posts nor for actions taken to moderate such posts, which is consistent with s. 230. The CUSMA came into force on July 1, 2020.

[205] Our federal private sector privacy law is the *Personal Information Protection and Electronic Documents Act*, SC 2000, c 5. Various provincial legislation have been deemed substantially similar, such as Alberta's *Personal Information Protection Act*, SA 2003, c P-6.5.

[206] PIPEDA, *supra* note 205 at s 4(1). The person must be identifiable, but the Office of the Privacy Commissioner interprets that broadly.

[207] Office of the Privacy Commissioner of Canada, *Joint investigation of Facebook, Inc. by the Privacy Commissioner of Canada and the Information and Privacy Commissioner for British Columbia* (April 25, 2019), online: https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2019/pipeda-2019-002/.

[208] The Competition Bureau also investigated Facebook concerning this matter for false or misleading representations. Facebook settled for $9.2 million: Government of Canada, "Facebook to pay $9 million penalty to settle Competition Bureau concerns about misleading privacy claims" (May 19, 2020), online: https://www.canada.ca/en/competition-bureau/news/2020/05/facebook-to-pay-9-million-penalty-to-settle-competition-bureau-concerns-about-misleading-privacy-claims.html.

At the time of writing, the Government of Canada has introduced Bill C-27 to reform PIPEDA and expand privacy protection.[209] Analysis is beyond the scope of this paper, but readers are encouraged to review the Bill through the lens of social media responsibility and protection from online harms. The bedrock of information manipulation is data. Thus, privacy law is a natural framework to address protection of users from general systems of surveillance,[210] to provide oversight of the business models that make users vulnerable,[211] to introduce measures for algorithmic accountability and regulation of artificial intelligence, to complement deceptive marketing laws with data protection specific obligations, and to strengthen the rights and obligations concerning the kinds of data practices that are acceptable in an information environment we cannot realistically argue we can opt out of.

Where does that leave us? Social media are certainly regulated in Canada, but there are significant gaps in our laws concerning intermediary liability for online harms, and specifically for information manipulation. Generally, the most viable route to intermediary liability is a claim in defamation law, but only some forms of information manipulation are defamatory. Outside of intermediary liability, data protection is an important legal tool to address privacy aspects of platform responsibility, but it is one piece of the pie and does not directly regulate information manipulation or online harms more generally.

This turns attention to the governance frameworks that exist beyond traditional law, namely companies content moderation practices. It is notable that even in countries with comprehensive intermediary liability laws, content moderation policies play an important role. There are a variety of reasons for this. Companies are incentivized to moderate content to address even the lawful but awful, although the lack of content moderation is the business model of some of the platforms. These companies are global and policies provide an avenue to create standards of wide application. The other reason is that few cases make it to court. In civil law, litigation is too expensive and slow to be worthwhile for most litigants. In the area of criminal law, online harms are notoriously underreported and under investigated. Vulnerable groups are often wary of complaining to the police, and law enforcement do not all have the resources or special knowledge necessary to investigate some complaints, and there are reports of not taking such complaints seriously.[212]

Content Moderation by Social Media

[209] Bill C-27, *An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts*, 1st Session, 44th Parliament, 2022.

[210] Intermediary laws, such as the ECD, provide that there is no general obligation to monitor: see ECD, *supra* note 189 at Article 15.

[211] This is where competition law might play a key role here, especially to address micro-targeting and its role in information manipulation.

[212] Some of this will be explored in a forthcoming report by the Canadian Council of Academics on *Public Safety in the Digital Age*, online: https://cca-reports.ca/reports/public-safety-in-the-digital-age/.

The argument is often made that users have 'rights' as against social media companies, that a platform has violated their *Charter* right to free expression because of a content moderation decision. The *Charter* does not apply to the activities of private companies unless these companies undertake a governmental action.[213] This means that, in general, users do not have a right to free expression on social media, because they are privately owned. This does not mean that the right to free expression has no legal significance concerning content moderation by platforms. For example, governments must comply with the *Charter* in the enactment of any laws. Thus, whatever online harms bill is introduced by the Government of Canada will need to be *Charter* compliant.[214]

If there is some unease that rights seem privatized in digital spaces, there is legitimacy to this concern. The way a platform interprets free expression, for example, whether based on corporate values, domestic law (often the First Amendment) or international human rights, is a system of private governance of their design without any of the normal features of accountability we expect of state-run systems.[215] The decision of various social media to deplatform former President Trump, for example, is important from a private governance perspective, invites scrutiny of who sets the terms of moderation and highlights the tremendous power of these platforms. Facebook has a formal appeal mechanism through the Oversight Board, which reviewed the decision to deplatform President Trump and concluded the decision to restrict access was appropriate, but the penalty of indefinite suspension was not.[216]

A similar phenomenon is observable in what Elena Chachko calls "national security by platform".[217] As she explains, platforms are now central to geopolitics and security.[218] Many social media collaborate with governments, employ foreign policy directors, incident response teams, formal content moderation and appeals mechanisms, trust and safety teams and policies directed at national security issues, such as mis- and dis-information, election integrity, terrorism and violent extremism. For most social media, they have taken on this role out of

[213] See ss 1 and 32 of the *Charter*, *supra* note 135. The *Charter* applies to the legislative, judicial, and executive government and in instances where government has delegated authority to a private party or that party acts as an agent of the state. There has been some scholarly exploration of this, but not to the extent of analysis of the indirect horizontal application of the *European Convention for the Protection of Human Rights and Fundamental Freedoms*, 1950. Online harms legislation will need to be scrutinized to the extent that there is a deliberate delegation of power of the government as that might invite a different type of *Charter* scrutiny to the actions of platforms.

[214] Courts must also develop the common law in line with *Charter* values. For example, the right to privacy and to free expression influenced the development of the torts of defamation and privacy: *Hill*, *supra* note 167; *Jones*, *supra* note 182.

[215] By privatization I mean that a private party performs a function normal the reserve of government. This has been described as tilting, where human rights were initially structured as a relationship be tween citizens and the state, and with digital technologies, we now exercise and experience our rights as between users and technology companies: see Emily B. Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press, 2015), chapter 6.

[216] Oversight Board, Case decision 2021-001-FB-FBR https://www.oversightboard.com/decision/FB-691QAMHJ.

[217] Elena Chachko, "National Security by Platform" (2021) 25 Stanford Technology Law Review 55.

[218] *Ibid*.

necessity, because of the ways that their platforms have been used and exploited. However, some platforms espouse political ideologies that influence their design and content moderation. A key risk is that since this arrangement is often indirect and informal, a private actor can "choose what functions they wish to fulfill."[219] Indeed, platforms can choose whether to engage with it at all.[220] This destabilizes national security protection, because there is minimal oversight by government of what platforms do, a novelty to the issues and tremendous discretion for platforms to decide on a course of action, if any.[221] This is observable with Facebook's response to the spread of violent and extremist content, and mis- and dis-information, in Myanmar, and its contribution to violence against the Rohingya. The *Human Rights Council Report of the independent international fact-finding mission of Myanmar* went so far as to call Facebook's response "slow and ineffective."[222]

Content moderation is not entirely voluntary. Rather, it is an important step to fulfil businesses responsibility to respect human rights per the *United Nations Guiding Principles on Business and Human Rights* ("UN Guiding Principles").[223] The UN Guiding Principles impose due diligence obligations on businesses concerning their human rights impact. Namely, businesses should avoid adversely impacting human rights, monitor their compliance and work to prevent and mitigate harm, and provide access to a remedial mechanism. The Guiding Principles is the blueprint for companies to embed human rights into their operations and to hold them accountable, but it still relies on good faith commitment, and many of the content moderation policies of social media used in the Convoy make no reference to or reflect the Guiding Principles.[224]

Further, company efforts are often collaborative and/or spurred by government, as a form of co-regulation, such as the EU *Code of Practice on Disinformation* discussed above.[225] Another example is The Global Internet Forum to Counter Terrorism (GIFCT), which is a collaboration between various online services to address terrorism and violent extremism. It was created through cooperation among various stakeholders beyond its industry founders to include academia, civil society, and bodies such as the United Nations Counter-Terrorism Executive

[219] *Ibid* at 125.
[220] *Ibid* at 127. This problem of indirection and informality is observable in technology regulation generally, an issue flagged by technology regulatory scholars for years.
[221] *Ibid.*
[222] Human Rights Council, "Report of the independent international fact-finding mission on Myanmar" (2018) A/HRC39/64, online: https://www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf at para 74.
[223] See Office of the High Commissioner for Human Rights, ""Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework" (2011), HR/PUB/11/04, online: https://www.ohchr.org/sites/default/files/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf.
[224] The Guiding Principles are rooted in a company's social licence to operate. They were endorsed by the Human Rights Council, which elevated it from guidance to a system of governance. See John Ruggie, *Just Business: Multinational Corporations and Human Rights* (Norton, 2013); David Kaye, *Speech Police: The Global Struggle to Govern the Internet* (New York: Columbia Global Reports, 2019). See Meta, "Corporate Human Rights Policy", online: https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf; Google, "Human Rights", online: https://about.google/human-rights/.
[225] Codes of practice, *supra* note 195.

Directorate and the EU.[226] GIFCT works to develop both preventative and incident response mechanisms.

There are two aspects to content moderation to be examined. First, social media companies usually deploy technology, in some form, to regulate harmful content. While these are often framed as technical solutions, they are systems of governance technically deployed. Second, social media regulate users through their content moderation policies. This will be examined through the lens of the social media used in the Convoy.

Content Moderation Technology

Technology is key to combatting information manipulation.[227] However, it is not the panacea to harmful content. It can be blunt, lack the finesse necessary to assess ambiguous content accurately and contextually, and is shaped by the mindset (and potential bias) of the dataset creator with minimal oversight external to the organization.[228]

There are numerous automated tools used for content moderation to filter, classify, curate, and organize content. Many are driven by artificial intelligence.[229] For example, technology can help identify inauthentic accounts.[230] Perceptual hashing, such as Microsoft's PhotoDNA, is a digital fingerprint used to identify harmful images and videos, such as child sexual abuse, terrorist and violent extremist content, or copyright infringing content.[231] GIFCT spearheaded a shared databased for its members.[232] Project Arachnid uses hashing to identify child sexual abuse material.[233] Other tools include image recognition to prioritize content for human review, and natural language processing techniques to detect hate speech and extremist content.[234]

Technology is also used to nudge behavioural changes in users.[235] Currently, such strategies are

[226] See online: https://gifct.org/; Chachko, *supra* note 217 at 89.
[227] See Kreps, *supra* note 62 at 6-7.
[228] See Spandana Singh, "Everything in Moderation" (July 22, 2019), *New American Foundation*, online: https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/the-limitations-of-automated-tools-in-content-moderation.
[229] See limits of AI in Alex Feerst, " The Use of AI in Online Content Moderation or, the tech sector invested in automation and all I got was this questionable adjudication" (September 2022) Digital Governance Working Group, https://platforms.aei.org/wp-content/uploads/2022/09/The-Use-of-AI-in-Online-Content-Moderation.pdf.
[230] It is the AI text generator for false information that can also be used to identify it. Kreps discusses Grover, a model which both generates and identifies "fake news", *supra* note 62 at 6-7.
[231] See YouTube Content ID (adapted from PhotoDNA) and Google Drive.
[232] It is limited to terrorist entities on the United Nations designated terrorist groups lists. GIFCT is an NGO founded in 2017 by Facebook (now Meta), Microsoft, Twitter, and YouTube, and has expanded its membership since: *supra* note 226.
[233] See online: https://www.projectarachnid.ca/en/.
[234] Singh, *supra* note 228.
[235] Nudging is the theory by Richard Thaler and Cass Sunstein that indirect, subtle choice architecture are effective to prompt behaviour changes, such as mandating choice for organ donation with driver's license renewal: *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Penguin Book, 2008).

necessarily experimental, because the research on their effectiveness is in the early stages.[236] For example, a common tool used by online services to address information manipulation is information correction. Information correction has appeal, because the interference with the right to freedom of expression is relatively minor.[237] Users still have access to the information, but the content is flagged as false information or a synthetic account. Is information correction effective? The research is mixed. The strongest argument against information correction was set out in a 2010 article by Brendan Nyhan and Jason Reifler, which they labelled "backfire effect", that debunking information as false is ineffective and can have the opposite effect and further entrench readers misperceptions.[238] However, the backfire effect has since been shown to be overstated, and the problem more subtle, and further research is needed to measure the effectiveness of information correction.[239] For example, information correction can have the effect of "belief echoes".[240] And there is an illusion of believability to any information shared, thus there is a risk that debunking information gives the story the illusion of truth.[241] The answer for information correction might be how softly it is delivered, such as Facebook's change from information correction to providing a diverse array of news stories on the topic,[242] or timing the correction at the tipping point of public awareness that it is untrue.[243]

Similarly, content removal and blocking are sometimes used. I include it in the discussion of technical solutions, although it is often a mix of automated and human actioning. One question is the effectiveness of content removal. It is not clear that it is effective in changing beliefs, because there is no new information to replace what was taken down.[244] Further, problematic content is shared almost instantaneously after posting, and therefore the content is rarely actually removed from circulation. The livestream of the Buffalo attack was removed by Twitch within two minutes, but the video had already been copied and reposted across various

[236] One can see platforms make changes in response to new research: Tessa Lyons, "Replacing Disputed Flags With Related Articles" (December 20, 2017) *Meta,* online: https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/.

[237] Helm, *supra* note 101 at 315-318.

[238] Brendan Nyhan and Jason Reifler, "When Corrections Fail: The Persistence of Political Misperceptions" (2010) 32 Political Behaviour 303. Helm discusses, *supra* note 101 at 315-318; Wardle and Derakhshan, *supra* note 27 at 45; Timothy Caulfield, "Does Debunking Work? Correcting COVID-19 Misinformation on Social Media" in Colleen Flood *et al*, *Vulnerable: The Law, Policy and Ethics of COVID-19* (University of Ottawa Press, 2020) at 188-193.

[239] See exploration of literature by Caulfield, *supra* note 238. He concludes "while a backfire effect may occur in some circumstances – this is an area where more research would be helpful – it certainly isn't such a robust and measurable phenomenon that it should stop us from mounting efforts to counter misinformation on social media": at 190.

[240] Thorson, *supra* note 95.

[241] Caulfield, *supra* note 238 at 190-191.

[242] See Lyons, *supra* note 236.

[243] Caulfield, *supra* note 238 at 191-192. Caulfield lists various principles to maximize the impact of information correction: use facts; communicate clearly and simply; use trusted sources, although trust is a challenge; identify that there is scientific consensus if applicable; be kind and authentic; write in a storytelling style; use rational argument emphasizing gaps in logic etc; frame the debunking to emphasize the facts not the misinformation; audience should be the general public not believers of the misinformation: at 193-198.

[244] See discussion, Helm, *supra* note 101 at 321.

platforms.[245] Removal can also act as a beacon drawing more attention to the post or reinforce beliefs.[246] Deplatforming accounts can have some success disrupting the momentum of a group. They lose followers and fans and struggle to gain news ones, and it disrupts monetization. They may jump to alternative platforms, but it does not return the group to its previous level.[247] Content removal might serve a different function than changing beliefs, serving an expressive role reinforcing what is acceptable conduct, although care must be taken to balancing various rights.[248]

From a law and governance perspective, we are in a period of technical and regulatory experimentation. There is no consensus on the strategies that will work to address information manipulation. Therefore, solutions change and sometimes backfire. Because of the social power of some platforms, the backfire can be monumental. The result in a mixed bag of interventions. For example, the messaging app WhatsApp recently updated technical features to improve user privacy, to enable users to leave groups without notifying channels, hide that they are online and block screenshotting of messages intended to be viewed once.[249] These are positive solutions specific to that app. Twitter embeds nudge theory using technical tools,[250] Twitter uses this technique by prompting users to reconsider posting tweets that contain harmful language and to read articles before sharing them.[251]

In contrast, Facebook's tweak to its algorithm to improve user well-being backfired. Around 2017-18, Facebook changed its engagement ranking algorithm[252] to boost meaningful social interactions (MSI). Popular posts and those by friends and family were amplified and professional news was de-amplified. The Facebook Files leaked by Frances Haugen also revealed that part of the algorithmic tweaks entailed boosting posts that generate strong

[245] Mia Sato, "How the Buffalo shooting livestream went viral" (May 17, 2022) *Verge*, online: https://www.theverge.com/2022/5/17/23100579/buffalo-shooting-twitch-livestream-viral-content-moderation.

[246] This is known as the Streisand effect. Also discussed in Helm, *supra* note 101 at 321-322.

[247] See literature review discussion by Amarnath Amarasingam: "Does Deplatforming Work? A quick survey of literature in the wake of the Capitol Hill Attack" (January 12, 2021) *Intrepid*, online: https://www.intrepidpodcast.com/blog/2021/1/12/does-deplatforming-work-a-quick-survey-of-literature-in-the-wake-of-the-capitol-hill-attack. See monetization of YouTube vs Rumble vs BitChute vs Odysee. The more generous monetization strategies of some new video-sharing platforms are worth studying for their impact on mainstream video-sharing platforms.

[248] Drawing from the argument that a purpose of the law is to reinforce or change norms. Content removal should be based on human rights principles.

[249] Michelle Toh, "WhatsApp is going to stop letting everyone see when you're online" (August 9, 2022) *CNN,* online: https://www.cnn.com/2022/08/09/tech/whatsapp-privacy-changes-meta-intl-hnk/index.html.

[250] Thaler and Sunstein, *supra* note 235.

[251] Twitter Safety, "How Twitter is nudging users to have healthier conversations" (June 1, 2022), online: https://blog.twitter.com/common-thread/en/topics/stories/2022/how-twitter-is-nudging-users-healthier-conversations.

[252] Engagement ranking is controversial and was the focus of some of Frances Haugen's testimony. Some of the controversy is that increasing engagement is viewed as self-serving in that it keeps users on Facebook: Jeremy B. Merrill and Will Oremus, "Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation" (October 26, 2021) Washington Post, online: https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/.

emotional emoji reactions. Love, laughter, wow, sad, and angry emoji reactions receive five times the value as the like emoji.[253] Internal research showed that posts that elicited angry emoji reactions were "disproportionately likely to include misinformation, toxicity and low-quality news."[254] As a result, Facebook's algorithm fostered the spread of mis- and dis-information.[255]

Content Moderation Policy

Turning to the Convoy, the conversations that fueled it started long before January 2022 across various mainstream and alternative social media, in groups discussing vaccine mandates, COVID restrictions and conspiracy theories, and amplified by influencers and alternative news organizations. There was a built-in audience and participants. Key to social media responsibilities is the systemic risks associated with their services. How are the designed? How does the algorithm work? How do they monitor the impact of their services, and do they action their findings? Except for privacy, there is no legal obligation to manage their systemic risk, particularly as much of the content is lawful. The opacity of social media, beyond transparency reports, which have not yet matured and standardized for this industry, means that this aspect of what fueled the Convoy is a question of self-regulation by social media. Further, while the monetization structure of various social media is not explored in depth in this paper, the Commission may consider enquiring further as to the financial drive of some influencers and monetization practices of social media used in the Convoy. For example, under YouTube's Partner Programme, a creator would make money off the ads that surround their videos on YouTube. The content produced by these influencers might then be subject to review under community guidelines and other monetization policies.[256]

The content moderation policies of social media used in the Convoy vary widely. As we explored, Convoy supporters and organizers tended to use Facebook, Twitter, TikTok, YouTube, Rumble, Telegram, Zello, BitChute, Odyssey, GoFundMe and GiveSendGo. Mainstream platforms have relatively developed content moderation policies, although this analysis puts aside the effectiveness or legitimacy of their approaches, including enforcement.[257] By developed, I mean simply that they have a policy or policies that substantively address the risks of harm of their services, a system to action content that infringes these policies, including a mechanism for users to report content, and an appeal mechanism. Some social media might

[253] See Facebook files, Wall Street Journal, online: https://www.wsj.com/articles/the-facebook-files-11631713039.

[254] Merrill and Oremus, *supra* note 252.

[255] I reference both mis- and dis-information, because Facebook's algorithm has been exploited for information operations, such as Russian originated ads on Meta, "An Update On Information Operations On Facebook" (September 6, 2017), online: https://about.fb.com/news/2017/09/information-operations-update/.

[256] YouTube, "Monetisation Policies", online: https://www.youtube.com/howyoutubeworks/policies/monetization-policies/.

[257] For a thorough analysis of content moderation and human rights law, see Mackenzie Common, *Rule of law and human rights issues in social media content moderation* (2020) PhD thesis, London School of Economics and Political Science; Douek, *supra* note 183.

notionally tick these boxes, but moderate minimally. And a key difference across social media used in the Convoy is the extent to which they proactively moderate harmful content.[258]

Facebook and Twitter, for example, have various policies on hate speech, terrorism and violent extremism, violence, manipulated media, synthetic accounts and similar.[259] The subject matters covered are similar, but their policies are not, reflecting the differences in their platforms, but also different ethos about where they land on particular issues.[260] Notably, these policies serve to reflect and reinforce what is illegal (e.g. uttering threats),[261] but they also set rules about acceptable expression above and beyond the law. Therefore content moderation is key to addressing legal but offensive expression. On hate speech, their policies set the bar for acceptable expression higher than criminal or human rights law.[262] For example, Facebook defines hate speech attacks as "violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation".[263] Facebook treats misinformation as an issue of integrity and authenticity. It has a misinformation policy, which targets types of misinformation: physical harm and violence, health, voter, and census interference and manipulated media.[264] Twitter introduced a crisis misinformation policy in May 2022.[265] Disinformation is dealt with separately under policies addressing e.g. spam, coordinated accounts or inauthentic behaviour.

Some of the softer mechanisms used by social media are important because they move beyond the leave up/takedown binary. As explored above, information correction or diversity, demotion or limiting visibility, warnings, labels, and invitations to rethink or read before sharing are all forms of strategic friction. For more formal content moderation, Facebook uses a hybrid approach using a mix of algorithmic and human review, and proactive and complaints-based review. When a user submits content to post, it is screened to identify if the content matches hash databases of CSAM and terrorism content. If there is a match, the content is blocked from being posted. Once content is online, Facebook monitors content using algorithms to identify objectionable content based on a variety of metrics, including words, images or behaviours that are viewed as commonly associated with the different types of objectionable content, the

[258] A significant issue is the need for proactive content moderation, but it is difficult to square that with the equal need not to mandate general systems of monitoring or undermine encryption, which can be an invasion of privacy. The line tends to be found between general and specific monitoring e.g. proactively searching for hashed content that is specifically CSAM, terrorist or violent extremist content, but there is a lot of content in the grey zone that manipulates users.

[259] See Twitter's "The Twitter Rules", online: https://help.twitter.com/en/rules-and-policies/twitter-rules; Meta's "Facebook Community Standards", online: https://transparency.fb.com/policies/community-standards/.

[260] For a long time Twitter was resistant to impose stronger content moderation in faithfulness to a First Amendment approach, but in recent years has begun to develop more comprehensive moderation practices.

[261] *Criminal Code*, *supra* note 40, s 264.1.

[262] *Ibid*, s 319, which prohibits public incitement or wilful promotion of hatred, or wilful promotion of antisemitism; *Keegstra*, *supra* note 137; *Whatcott*, *supra* note 168.

[263] Meta, "Hate Speech", online: https://transparency.fb.com/policies/community-standards/hate-speech/.

[264] Meta, "Misinformation", online: https://transparency.fb.com/policies/community-standards/misinformation/.

[265] Yoel Roth, "Introducing our crisis misinformation policy" (May 19, 2022), online: https://blog.twitter.com/en_us/topics/company/2022/introducing-our-crisis-misinformation-policy.

poster's identity, the context of the post and comments. If the algorithm determines that the content clearly violates the Community Standards, it is removed, but if the algorithm is unclear, a human moderator reviews the content. Users can also report content as violating the Community Guidelines. Penalties for violations range from warnings, to restricted access to certain Facebook features such as live streaming, to disabling accounts permanently or temporarily. Users can appeal the decision, including to the Facebook Oversight Board.[266] Facebook removed some Convoy-related Facebook groups, pages, and accounts, such as spammers, which capitalized on the Convoy to draw users to off-platforms websites with pay-per-click ads, hate groups and conspiracy groups, such as QAnon.[267]

Twitter uses a variety of friction techniques, but like Facebook it also uses a hybrid approach relying on automation and human review. Users can report content that violates Twitter's Rules. Enforcement is scaled to include specific actioning of tweets (labels, visibility, removal), restricted messaging, or account level enforcement (read-only mode, verification, suspension). Users can appeal a locked or suspended account.[268] Twitter permanently suspended a Convoy account and that of an influencer.[269]

TikTok also uses friction techniques, and a hybrid system of automation and human review, and similar to the above, uses a scalable approach of warnings, temporary and then permanent account suspensions, with the opportunity to appeal. For some content, such as CSAM, it has a zero-tolerance policy.[270] TikTok prohibits harmful disinformation that causes significant harm.[271] Several influencers made regular use of TikTok, but at the time of writing, I am not aware that there has been any actioning of Convoy related accounts.

Content moderation is considerably different for what has been described as alternative social media. This can lead to a whac-a-mole game as users jump to less moderated platforms, whether to avoid moderation rules or because their account has been suspended. This is

[266] Oversight Board, "Appeal to shape the future of Facebook and Instagram", online: https://www.oversightboard.com/appeals-process/. Generally see, online: https://transparency.fb.com/.

[267] Culliford, *supra* note 5.

[268] Twitter, "Our range of enforcement options", online: https://help.twitter.com/en/rules-and-policies/enforcement-options.

[269] Kevin Jiang, "Ontario MPP Randy Hillier 'permanently suspended' from Twitter" (March 8, 2022), *Toronto Star,* online: https://www.thestar.com/politics/provincial/2022/03/08/ontario-mpp-randy-hillier-suspended-from-twitter.html.

[270] TikTok, "Content violations and bans", online: https://support.tiktok.com/en/safety-hc/account-and-user-safety/content-violations-and-bans.

[271] Per TikTok's definition in its "Community Guidelines", significant harm relates to "harm to individuals, our community, or the larger public regardless of intent. Significant harm includes serious physical injury, illness, or death; severe psychological trauma; large-scale property damage, and the undermining of public trust in civic institutions and processes such as governments, elections, and scientific bodies. This does not include simply inaccurate information, myths, or commercial or reputational harm", online: https://www.tiktok.com/community-guidelines#37.

observable with the suspension by GoFundMe of the Convoy campaign, which then moved to GiveSendGo.[272]

As noted, the Convoy was a movement that thrived off videos. YouTube was used by organizers and supporters of the Convoy, but users also moved to alternative video-sharing platforms, primarily BitChute, Rumble and Odysee. Like other mainstream social media, YouTube has community guidelines that broadly prohibit deceptive practices, harassment, hate speech and harmful or dangerous content or similar.[273] They enforce their guidelines using a mix of automated and human review, and flagging by users.[274] In contrast, BitChute has Community Guidelines, and a method for reporting content and appealing decisions.[275] However, moderation is primarily based on user reports and does not seem to include proactive measures. Further, and crucially, the Community Guidelines primarily only prohibit illegal content with the result that BitChute has become a haven for extreme expression. It only restricts hateful content if it is illegal as incitement to hatred. It narrowly prohibits terrorist and extremist content of designated entities under counterterrorism legislation, which is underinclusive of white nationalist groups and the broader harms that concern social media moderation.[276] There is no policy concerning mis- and dis-information or other forms of information manipulation, including manipulated media or synthetic accounts, although it prohibits spamming and brigading.

Rumble, which hosted one of the videos that helped spark momentum for the Convoy,[277] takes a similar approach to content moderation as BitChute and as a result has also become popular for posting extremist content. Rumble's CEO describes the platform as "different from YouTube and Facebook because it uses far fewer algorithms for recommending and reviewing

[272] Amanda Connolly, "GoFundMe, GiveSendGo defend handling of convoy blockade fundraising campaigns" (March 3, 2022) *Global News*, online: https://globalnews.ca/news/8656947/gofundme-givesendgo-convoy-blockade-campaigns/.

[273] YouTube, "Community Guidelines", online: https://www.youtube.com/intl/ALL_ca/howyoutubeworks/policies/community-guidelines/?gclid=Cj0KCQjw39uYBhCLARIsAD_SzMT8vv1X75lRt3vTw4in65q_TUDuN1-shNv9PQvP8UFfG2YKBSGTOksaAoKIEALw_wcB, and "Policies Overview", online: https://www.youtube.com/howyoutubeworks/policies/overview/.

[274] YouTube, "How does YouTube enforce its Community Guidelines?", online: https://www.youtube.com/intl/ALL_ca/howyoutubeworks/policies/community-guidelines/?gclid=Cj0KCQjw39uYBhCLARIsAD_SzMT8vv1X75lRt3vTw4in65q_TUDuN1-shNv9PQvP8UFfG2YKBSGTOksaAoKIEALw_wcB#enforcing-community-guidelines.

[275] BitChute, "Content Moderation Policy", online: https://support.bitchute.com/policy/content-moderation#flagging--reporting and "Community Guidelines", online: https://support.bitchute.com/policy/guidelines/.

[276] Terrorist designation has not included alt right groups, although Canada has added e.g. Proud Boys and Blood & Honour in recent years: Public Safety Canada, "Current Listed Entities", online: https://www.publicsafety.gc.ca/cnt/ntnl-scrt/cntr-trrrsm/lstd-ntts/crrnt-lstd-ntts-en.aspx. Bitchute also keeps its own list of prohibited entities, but there are only two on the list: "Prohibited Entities List", online: https://support.bitchute.com/policy/prohibited-entities-list.

[277] Broderick, *supra* note 7.

content."[278] Videos are displayed in chronological order to users based upon who they follow on the platform. Rumble does not use algorithms to proactively filter high risk content. While the Terms and Conditions prohibit more than illegal content, the bar set is not much higher.[279] There is no policy on mis- or dis-information.

Odysee's Community Guidelines broadly prohibit incitement of hatred or violence, promotion of terrorism and/or criminal activity and violence that is not newsworthy. Mis- and dis-information, hateful content and extremism is permissible. Users can report content and enforcement includes content removal, blocking comments or filtering a user channel.[280] However, the structure of Odysee is unique. Videos are not stored on a centralized server, but instead decentralized across a network using blockchain technology.[281] This means that the videos cannot be permanently deleted – even by the user that uploaded it, although Odysee can block access to the content via the app.[282]

Telegram was actively used to organize and generate support for the Convoy. Telegram minimally moderates its service. As a messaging app it is different than any of the above social media. Many forms of moderation used by mainstream platforms create significant privacy risks if used to moderate private messaging.[283] However, as examined in Part I, Telegram private groups can have up to 200k members (while Instagram and iMessage cap group chats at 32 people) and their channels allow broadcasting to unlimited subscribers.[284] It is difficult to characterize these as private.[285] Telegram does not moderate its private groups or channels,

[278] Fizza Kulvi, "Meet Rumble, Canada's new 'free speech' platform — and its impact on the fight against online misinformation" (July 8, 2021) *The Conversation*, online: https://theconversation.com/meet-rumble-canadas-new-free-speech-platform-and-its-impact-on-the-fight-against-online-misinformation-163343.

[279] Rumble, "Website Terms and Conditions of Use and Agency Agreement", online: https://rumble.com/s/terms. See discussion by Kevin Newman, "Investigating Canadian YouTube rival Rumble and its growing popularity among the world's far right" (February 19, 2022) *CTV News*, online: https://www.ctvnews.ca/w5/investigating-canadian-youtube-rival-rumble-and-its-growing-popularity-among-the-world-s-far-right-1.5787533.

[280] Odysee, "Community Guidelines", online: https://odysee.com/@OdyseeHelp:b/Community-Guidelines:c and "Report content:, online: https://odysee.com/$/report_content?claimId=166ec880e443d4e1bca31dbd142bdf2a4a8aa61f&sunset=lbrytv.

[281] Eviane Leidig, "Odysee: The New YouTube for the Far-Right" (February 17, 2021) Global Network on Extremism and Technology, online: https://gnet-research.org/2021/02/17/odysee-the-new-youtube-for-the-far-right/#:~:text=Odysee's%20community%20guidelines%20state%20that,not%20allowed%20on%20the%20platform

[282] Odysee explains "We cannot remove published content from the blockchain itself, although we can block content accessed via our app or other services on top of the blockchain": "Terms of Service" online: https://odysee.com/$/tos. Blockchain, as an immutable ledger, means that the data on the blockchain cannot be changed: Eileen Brown, "Blockchain-based Odysee keeps your social media content online" (April 8, 2021) *Zdnet*, online: https://www.zdnet.com/finance/blockchain/blockchain-based-odysee-keeps-your-social-media-content-online/.

[283] This paper also does not explore the privacy and security risks of undermining encryption, but it is an important issue when examining what kind of a duty a messaging app should have to regulate content.

[284] Telegram "FAQ", online: https://telegram.org/faq#q-what-39s-the-difference-between-groups-and-channels; Sam Andrey, Alexander Rand and Karim Bardeesy, "Rebuilding Canada's Public Square" (September 2021), online: https://static1.squarespace.com/static/5e9ce713321491043ea045ef/t/615478c6a74009181c27d15e/1632925924146/RebuildingCanada%27sPublicSquare.pdf.

[285] This is a difficult issue. Facebook groups are also private and have no limit in size.

except reports of spam. Therefore, hate speech, terrorist and violent extremist content, mis- and dis-information, graphic content, CSAM and similar are unmoderated. In public channels and groups, the Terms of Service only prohibit promotion of violence and illegal pornographic content.[286]

The walkie-talkie app Zello was used to organize the blockades. Its Terms of Service and Community Guidelines broadly prohibit harmful behaviour,[287] including anything that a Zello representative considers objectionable.[288] It prohibits promotion of violent extremism, but narrowly defines terrorism as related to organizations on sanctions lists.[289] Mis- and dis-information is not specifically addressed. Users can report violations, and there is no other information on the assessment process or whether Zello proactively moderates any content. Penalties for violations include suspensions or termination of accounts.

There is no legal obligation for these companies to go beyond the law in setting the conditions for use of their spaces, and there is reasonable criticism of social media that go too far.[290] The Guiding Principles provide the blueprint for developing content moderation policies, but there is no enforcement mechanism and thus relies on good faith implementation by corporate actors. To the extent social media act, it is driven by market incentives, social responsibility, and public pressure. As we saw, much of the conduct that seeds movements like the Convoy are a slow burn until something sparks it into action, and much of the content the fuels the slow burn is legal or in the grey zone.

Law Reform

We are in a period of rapid change in law reform to address online harms and intermediary liability.[291] Over the past several years there has been growing awareness about the central role of social media and other intermediaries, and technical and regulatory experimentation beyond the takedown models. Germany's restrictive notice and action regime with the *Network Enforcement Act*[292] seems to be the high-water mark for intermediary liability in western states. More recently, there has been a paradigmatic shift in the approach to law reform proposals

[286] Telegram, "Terms of Service", online: https://telegram.org/tos/terms-of-service-for-telegram-premium.

[287] Zello, "Terms of Service", online: https://zello.com/legal/terms/: "unlawful, harmful, threatening, abusive, harassing, tortious, excessively violent, defamatory, vulgar, obscene, nude, partially nude, or sexually suggestive, pornographic, libelous, invasive of another's privacy, hateful racially, ethnically or otherwise objectionable".

[288] *Ibid* and "Community Guidelines", online: https://zello.com/community/user-guidelines/.

[289] Zello, *supra* note 287.

[290] It is beyond the scope of this paper to explore how users' rights to free expression does *and should* operate in Canada as to access to social media. This is an area I am currently researching.

[291] Law reform is evident broadly across various areas of technology regulation. I am focused here on specific online harms and intermediary liability frameworks, although law reform in privacy, AI regulation and competition law are also relevant to addressing information manipulation.

[292] Network Enforcement Act (Netzwerkdurchsetzungsgesetz, NetzDG) (2017).

evidencing creative and promising solutions.[293] While a detailed review is outside the scope of this paper, I will sketch four main themes and flag some of the more problematic areas.

First, the most important development in law reform is the shift from a focus purely on content regulation to managing systemic risk. In Canada, the Commission on Democratic Expression proposed a duty to act responsibly.[294] Heritage Canada is currently exploring a risk management approach to online harms, which was the focus of workshops with the expert advisory group. How to address mis- and dis-information was a matter of debate.[295] In my view, if social media and other online services are tasked with managing their systemic risks, then this would naturally include mis- and dis-information. However, there should not be an obligation to action content that would fall in the category of lawful but awful. Ultimately risk management is not about actioning individual pieces of content, but this would need to be clear in legislation, particularly for lawful content. A model for such an approach is the EU's DSA.[296]

The EU passed the DSA in 2022, which imposes risk management obligations on "very large platforms".[297] These platforms must identify systemic risks related to the dissemination of illegal content, any negative impacts on human rights, and intentional manipulation of their services. In particular, platforms must take into account the impact of the "content moderation systems, recommender systems and systems for selecting and displaying advertisement".[298] Platforms must then mitigate these risks and conduct independent audits for compliance.[299] Other key provisions include user controls of recommender systems, advertising transparency and research access to data to monitor compliance.[300] The threat of mis- and dis-information is discussed throughout the recitals, but it is not specifically referenced in the body of the DSA. Rather, soon after the DSA was passed, a new *Code of Practice on Disinformation* was introduced.[301] A flaw of the DSA is the focus of risk management on very large platforms. While special obligations for these platforms might be appropriate, risk management is equally important for other online services, but the capacity of small and medium sized companies must be considered. The Convoy illustrates that both mainstream and alternative social media, and cross-posting, were used extensively, and risk management by just a few of the major mainstream platforms would not do much in the way of addressing online harms.

[293] To be certain, some misguided proposals have been made, which lack the delicate balancing this paper tries to show is necessary in the area of online harms.

[294] Canadian Commission on Democratic Expression, "How to Make Online Platforms More Transparent and Accountable to Canadian Users" (May 2022) *Public Policy Forum*, online: https://ppforum.ca/wp-content/uploads/2022/05/DemX-2-English-May-4-1.pdf. The duty to act responsibly is like the duty of care proposed in the United Kingdom, but the Commission sought to separate the concept from jurisprudence in negligence law.

[295] See worksheets, *supra* note 121.

[296] DSA, *supra* note 189.

[297] *Ibid* at Articles 25-33.

[298] *Ibid* at Article 26.

[299] *Ibid* at Articles 27-28.

[300] *Ibid* at Articles 29-31

[301] *Supra* note 195.

The focus on risk management captures the due diligence foundation of the Guiding Principles.[302] A variation would be a duty of care model, which has been proposed in the United Kingdom's *Online Safety Bill* (OSB).[303] It is unclear what the fate will be of the OSB as it is currently on pause, but it exemplifies legislation that got lost in the complexity of online harms regulation.[304] Regulated content, services, and the nature of the obligations varied so widely that if implemented, it will be difficult for most online services to comply. Complexity is a risk for any legislation to address online harms if the goal is to impose human rights sensitive obligations that differentiate between different types of content and online services. Most controversial, the OSB sought to directly regulate lawful but offensive expression and created specific offences related to disinformation.[305]

Second, transparency reporting is a crucial component of monitoring intermediary compliance. Both the DSA and OSB impose transparency reporting obligations.[306] As I have discussed, it is hard to do transparency reporting well and it is new for online services, in particular social media services. We might not have firm knowledge about what it is, the metrics of success or even precisely what we want social media to be transparent about, but it seems clear that transparency is central to the future of online harms regulation.[307]

Third, law reform is consistently focused on the creation of independent regulators for online safety. These are crucial to improve access to justice and to advance the necessary co-regulatory approach to online harms. Australia is the first jurisdiction to create an eSafety Commissioner with a research, education, investigation and enforcement mandate.[308] The remit started as protection of children from bullying and non-consensual sharing of intimate images, and abhorrent violent material, and has expanded to protection of adults, and broader regulatory power.[309] As a regulatory body, it is set up for actioning content, but they work closely with industry and have incorporated safety by design into their work.[310] Mis- and dis-information is not in the remit of the Commissioner. The UK has selected its telecommunications and broadcasting regulator, OFCOM, to be the OSB regulator.[311] The DSA

[302] *Supra* note 223.
[303] *Online Safety Bill*, 2022-2023, HC Bill 121 (as amended in Public Bill Committee), online: https://publications.parliament.uk/pa/bills/cbill/58-03/0121/220121.pdf.
[304] See visuals by Graham Smith, "Mapping the Online Safety Bill" (March 27, 2022) *Cyberleagle*, online: https://www.cyberleagle.com/2022/03/mapping-online-safety-bill.html.
[305] Caitlin Chin, "The United Kingdom's Online Safety Bill Exposes a Disinformation Divide" (August 11, 2022) *Center for Strategic & International Studies,* online: https://www.csis.org/analysis/united-kingdoms-online-safety-bill-exposes-disinformation-divide.
[306] DSA, *supra* note 189 at Articles 13, 23, 33 OSB, *supra* note 303 at ss 64-65.
[307] Daphne Keller, "Some Humility About Transparency" (March 19, 2021) *The Center for Internet and Society*, online: http://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency.
[308] See online: https://www.esafety.gov.au/.
[309] *Online Safety Act 2021*, No. 76, 2021 (OSB). See discussion of their legislative remit, online: https://www.esafety.gov.au/about-us/who-we-are/our-legislative-functions.
[310] See "Safety by Design", online: https://www.esafety.gov.au/industry/safety-by-design.
[311] The Office of Communications. In earlier research I identified OFCOM as a poor fit for digital rights regulation: Laidlaw, *supra* note 183, chapter 6.

mandates that Members States designate a regulatory body for DSA enforcement.[312] The role and function of a regulator is a point of debate. A useful model is that of privacy commissioners, which have both an educational, research and investigatory role.[313] The powers of the regulator are key, including to issue orders and impose fines.[314] A point of more debate is whether users should have access to a remedial mechanism outside of courts or the intermediary. There seems to be wide agreement that an ombudsperson is important to support users, especially marginalized and racialized groups that are often the targets online. However, the extent to which there should be a tribunal, e-court or social media council is debated.[315]

Finally, human rights are central to protection from online harms. The strength of the DSA is that is that it emphasizes fundamental rights not only in the recitals but in the specific ways that obligations are framed in the substance of the Regulation. This can be similarly observed in the OSB. For example, the OSB provides that when deciding on safety measures, the regulated entity must carry out an impact assessment on freedom of expression and privacy.[316] The expert panel on online harms provided feedback that a duty to act responsibly should entail two separate obligations on online services: protection from online harms and protection of human rights. In this way, for example, a decision about use of automation to manage the risks of online harms would also require assessment of the human rights impact of that approach on rights such as privacy and freedom of expression.[317]

Conclusion

Social media enabled the Convoy to mobilize and network. In many ways, this is precisely what social media was designed to do, and has done, for various movements. The issue is that the attack vector for false information, hatred, violence, extremism, harassment, and other forms of abuse are the same as for posting family photos, promoting your business, sharing cute animal videos, and learning how to repair your iphone screen. To tackle mis-, dis- and mal-information therefore requires the content of manipulative information to be unpacked, and a deeper dive on the actors who spread false information and the techniques they use, the impact on users who consume it, and the design of social media spaces being exploited. There also should be some convergence on what we are talking about when we discuss information manipulation. I suggest simplifying the analysis to disinformation (intentionally shared and

[312] See DSA, *supra* note 189 at Chapter IV.
[313] The eSafety Commissioner plays a crucial role educating and supporting the public and collaborating with industry.
[314] Strengthening the power of privacy commissioners has been a focus of law reform and lessons can be learned in creating an online safety regulator.
[315] I advocated for an e-tribunal for defamation disputes in "Re-Imagining Resolution of Online Defamation Disputes" 56(1) OHLJ 162. Heidi Tworek advocates for a social media council: "Social Media Councils" (October 28, 2019) *CIGI*, https://www.cigionline.org/articles/social-media-councils/. The Commission on Democratic Expression recommends an e-court for disinformation, *supra* note 294.
[316] OSB, *supra* note 309 at s 19.
[317] See *supra* note 121.

knowledge of falsity), misinformation (intentionally shared and believing it is true) and the everything else bucket of harmful content, such as terrorist and violent extremist content and hate speech, that intersect with mis- and dis-information.

The Convoy exposed gaps in Canadian law and policy on social media regulation. Any decisions about how to address Convoy content posted to social media was by the social media companies, based on their community guidelines and using various technical solutions. While each platform is different and these companies can devise creative, human rights sensitive solutions, there is an important discussion to be had about how to incentivize these solutions, create industry standards, and hold companies accountable. Law reform in various countries, including Canada, are taking steps to address online harms, and the readers are encouraged to monitor these developments and advocate for a human rights-based regime in Canada.