After Christchurch:
Hate, harm and the limits of censorship

3. Challenges in regulating online content

David Bromell

Working Paper 21/04

VICTORIA UNIVERSITY OF WELLINGTON TE HERENGA WAKA
Institute for Governance and Policy Studies
A research institute of the School of Government

CAIS CENTER FOR ADVANCED INTERNET STUDIES

INSTITUTE FOR GOVERNANCE AND POLICY STUDIES
WORKING PAPER
21/04

MONTH/YEAR

March 2021

AUTHOR

David Bromell
Senior Associate
Institute for Governance and Policy Studies

INSTITUTE FOR GOVERNANCE AND POLICY STUDIES

School of Government
Victoria University of Wellington
PO Box 600
Wellington 6140
New Zealand

For any queries relating to this working paper, please contact igps@vuw.ac.nz

ACKNOWLEDGEMENT

Research on this series of working papers has been financially supported by a fellowship at the Center for Advanced Internet Studies (CAIS) in Bochum, NRW, Germany (Oct 2020—Mar 2021).

This is paper three in a series of seven working papers, **After Christchurch: Hate, harm and the limits of censorship.**

The series aims to stimulate debate among policy advisors, legislators and the public as New Zealand considers regulatory responses to 'hate speech' and terrorist and violent extremist content online following the terrorist attack on Christchurch mosques in March 2019 and the Royal Commission of Inquiry that reported in November 2020.

The seven working papers in this series are:

Title	Reference
1. The terrorist attack on Christchurch mosques and the Christchurch Call	WP 21/02
2. 'Hate speech': Defining the problem and some key terms	WP 21/03
3. Challenges in regulating online content	WP 21/04
4. Regulating harmful communication: Current legal frameworks	WP 21/05
5. Arguments for and against restricting freedom of expression	WP 21/06
6. Striking a fair balance when regulating harmful communication	WP 21/07
7. Counter-speech and civility as everyone's responsibility	WP 21/08

Dr David Bromell is currently (until March 31, 2021) a research Fellow at the Center for Advanced Internet Studies (CAIS) in Bochum, North Rhine-Westphalia, Germany, which has supported his research on this series of working papers. He is a Senior Associate of the Institute for Governance and Policy Studies in the School of Government at Victoria University of Wellington, and a Senior Adjunct Fellow in the Department of Political Science and International Relations at the University of Canterbury. From 2003 to 2020 he worked in senior policy analysis and advice roles in central and local government.

He has published two monographs in Springer's professional book series:

- *The art and craft of policy advising: A practical guide* (2017)
- *Ethical competencies for public leadership: Pluralist democratic politics in practice* (2019).

Contents

Challenges in regulating online content

Abstract

In light of the terrorist attack on Christchurch mosques in March 2019 and the subsequent Christchurch Call to eliminate terrorist and violent extremist content online (Working paper 21/02), this paper argues that social media and other digital intermediaries are too big and have too much influence not to be subject to government regulation. Harmful digital communication is, however, exceptionally difficult to regulate for reasons that relate to the nature of the internet and digital communications, and to the business models and algorithms used by Big Tech companies.

'Harmful communication' is used in preference to 'hate speech' for reasons discussed in Working paper 21/03, **'Hate speech': Defining the problem and some key terms**.

The paper reviews recent research on whether de-platforming is effective and discusses the de-platforming of then-President Donald Trump in January 2021 by social media companies. It argues that decisions to restrict freedom of expression should be made within a framework of laws defined by democratically elected legislators and be open to review and appeal—not by private companies acting as courts to determine the boundaries of free speech.

Constraining harmful digital communication requires co-ordinated effort by multiple actors with divergent, competing and conflicting interests. What is required is some combination of governmental and inter-governmental regulation, industry self-regulation, industry-wide standards, multi-lateral, multi-stakeholder agreements and initiatives, technology innovation, and market pressure by advertisers, consumers and service users. What constitutes the right mix between government, business, and society in the governance of social media remains an open question.

Internationally aligned anti-trust/competition regulation, tax regimes and enforcement mechanisms will be part of the solution, but as with any exercise of the state's regulatory powers, we need to be mindful of unintended consequences and consider non-regulatory responses as alternatives or complements to censorship.

This suggests a need for integrated, cross-sectoral, strategic policy development—not reactive, piecemeal regulation that, even if it can be enforced, is unlikely to have any significant impact and may unjustifiably restrict the right to freedom of expression.

The remaining four working papers in this series elaborate on current international and national legal frameworks for regulating harmful communication (Working paper 21/05), arguments for and against restricting freedom of expression (Working paper 21/06), the need to strike a fair balance when regulating harmful communication (Working paper 21/07), and counter-speech and civility as everyone's responsibility (Working paper 21/08).

Tags: #ChristchurchCall #hatespeech #socialmedia #deplatforming #freespeech #CapitolRiots

Regulating digital intermediaries is necessary but difficult

Social media companies and other digital intermediaries are too big and have too much influence not to be subject to government regulation. Fielitz and Schwarz (2020, p. 9) comment:

> As social media platforms today connect much more than just users, they assume a vital role in shaping the political culture of democratic societies: they serve as news aggregators, as market places, as spaces for cultural creativity and as stages for political debate … As digital platforms move beyond the bounds of national and democratic controls, the decisions made by their operators require public justification—it is no longer merely a matter of private economic interests (p. 12).

Amitai Etzioni (2019) also argues that elected officials need to be much more involved in setting the regulatory framework specifying what content tech companies cannot remove, and what they must:

> Speech is too important for technocrats to control. Elected officials and the courts should be the main controllers—and they should control only when there is a clear and present danger to our security and to our democratic process.

The internet has not only promoted and enabled democracy. It has also done substantial damage to it (Steinmeier, 2021). It has been used to create digital citizens, the so-called netizens. It has also been used, as in the occupation of the US Capitol, to incite an unbridled mob. This will not change until the internet loses the feature many people love—that there are no rules (Bender, 2021).

Harmful digital communication[1] is, however, difficult to regulate, for reasons that relate to:

- The nature of the internet and digital communications; and
- The business models used by Big Tech companies.

The internet and digital communications

Digital communication is difficult to regulate because of its breadth and reach, the way it blurs boundaries between private and public space, the challenges of content moderation and the persistence of digital content in a global game of Whack-a-Mole.

A world wide web

Former New Zealand Cabinet Minister Steven Joyce (2021) commented following the occupation of the US Capitol on January 6, 2021:

> Society as a whole will have to reckon with social media. The effect of handing everyone the ability to be a broadcaster at the drop of a hat has on one level democratised the media but on another created a cacophony of noise where the first casualty is often the truth.

[1] As noted in Working paper 21/03 (p. 14, fn 11), by 'harmful digital communication', I mean digital public communication that incites discrimination, hostility or violence against members of a social group with a common 'protected characteristic' such as nationality, race or religion. New Zealand's Harmful Digital Communications Act 2015 has a much narrower scope and primarily aims to provide a pathway for individuals subject to direct bullying or harassment online to seek a mediated resolution. A proposed amendment to the Harmful Digital Communications Act will make the posting of intimate images and recordings without consent illegal and punishable by up to three years in prison, and allow courts to issue take-down orders for revenge porn recordings (Wade, 2021).

The internet has enabled communication to spread wide and deep, at low cost and at great speed ('going viral'), bypassing traditional media and enabling interactive, real-time engagement. Samaratunge and Hattotuwa (2014, p. 3) note that:

> This infrastructure has erased traditional geographies—hate and harm against a particular religion, identity group or community in one part of the world or country, can for example within seconds, translate into violent emulation or strident opposition in another part, communicated via online social media and mediated through platforms like Twitter, Facebook and also through instant messaging apps.

Because harmful digital communication occurs on a world wide web, and because social media companies and other digital intermediaries operate globally and to a significant extent beyond the reach of domestic law, we need to be realistic about what domestic regulation can achieve.[2]

Digitisation blurs boundaries between public and private space

Digitisation introduces additional layers of complexity to traditional concepts of public and private space (Gargliadoni, Gal, Alves, & Martinez, 2015, p. 8).

Something is 'public', as opposed to 'private', if it is accessible and 'open to witness' to anyone at all (Bromell, 2017, pp. 59–61; Coleman & Ross, 2010, p. 5). The problem is that digitisation blurs boundaries between what is 'public' and what is 'private'. Content originally intended for a private audience can, for example, be copied to public digital spaces from which it is not easily blocked or removed (Bromell & Shanks, 2021, p. 47).

Because many digital platforms permit anonymity or false identities, users vent private thoughts and emotions in public spaces that they would be less likely to voice in person, face-to-face. A report funded by the New Zealand Law Foundation's Information, Law and Policy Project and Luminate Group (Elliott, Berentson-Shaw, Kuehn & Salter, 2019) recommends reducing 'hate speech' and 'trolling' by using identity verification systems, noting that sites that do not allow anonymisation and force pre-registration solicit quantitatively fewer, but qualitatively better user comments because of the extra effort required to engage in discussion. Anonymity or false identities tends to be associated with greater incivility, with more aggressive and personal comments being made publicly (Kaspar, 2017; Rösner & Krämer, 2016; Halpern & Gibbs, 2013).

On the other hand, for vulnerable minority socio-cultural groups, online anonymity permits safe connection and access to support, and for digital intermediaries there are economic and financial drivers at play:

> As a larger number of platforms have come to rely on targeted advertising for a profitable business model, reliable and granular information about users is crucial. It may in fact be true that Facebook wants users to be identifiable to each other online as a means of limiting harassment, bullying and scams. But it is also true that their business models are predicated on connecting a user to her online activity. In this sense, delegitimizing anonymity benefits the profit strategies of Big Tech (Lingel, 2021, pp. 2536–2537).

[2] See for example analysis of Facebook as a global titan of transnational social media activity and its constant battle to evade jurisdiction controls under EU law in Shleina, Fahey, Klonick, Menéndez González, Murray, & Tzanou, 2020.

There are also valid and significant concerns about privacy and state control and surveillance of private digital communications, as occurs in China and other authoritarian states (Ovide, 2021d). This is a particular issue for monitoring and moderation of messaging apps, especially those that offer end-to-end encryption (Nicas, Isaac & Frenkel, 2021; Chen & Roose, 2021).[3]

Online content is not easy to moderate

There is no universally accepted definition of 'hate speech' or harmful communication that is consistent across platforms, responsive to local context and able to be operationalised to remove harmful content quickly.

Moreover, while internet service providers and large corporations like Alphabet, Facebook and Twitter have developed guidelines around the content they will allow on their platforms, their moderation systems appear to be easily bypassed because they work best in English, and in text-based content rather than videos, images and messaging apps. Samaratunge and Hattotuwa (2014, pp. 4–5) identified significant volumes of harmful communication circulating on Facebook in Sinhalese, for example, despite Facebook's clear guidelines, moderation systems and ability to ban and block users. And moderation of content in languages other than English needs to be sensitive to culture, context and current events, much of which is missed in automated moderation.

A further complicating factor is the tendency by the alt-right,[4] for example, to cloak ideology in generalities and use coded and covert signals, irony, sarcasm, images, memes and in-jokes to shroud borderline content and extreme views in plausible deniability (Binder, Ueberwasser, & Stark, 2020, p. 60; Chen, 2020a, pp. 157–159; Fielitz & Schwarz, 2020, pp. 45–48; Gilbert & Elley, 2020; Miller-Idriss, 2020, pp. 65–66; Schmitt, Harles, & Rieger, 2020; Owen, 2019).[5] This allows them to operate right up to the limits of the rules and guidelines applying to whatever platform they are using. It positions the alt-right as a counterculture and depicts anyone who is offended as 'a triggered, liberal snowflake who doesn't get the joke' (Miller-Idriss, 2020, p. 66). 'Hate speech' is also widely embedded in online games (Breuer, 2017; Groen, 2017), and the Christchurch mosque shooter's manifesto was peppered with in-jokes about video games (Macklin, 2019).

A global game of Whack-a-Mole

Harmful communication persists online in different formats across multiple platforms, which enables anyone wishing to upload or link to it to do so repeatedly, using a different link each time (Gargliadoni, Gal, Alves, & Martinez, 2015, pp. 13–14). As seen following the Christchurch mosque attacks, once content is uploaded to the internet, user interactions and re-postings mean that harmful content can replicate and mutate (with altered digital identifiers) over multiple websites and platforms. GIFCT's Hash Sharing Consortium and Content Incident Protocol (discussed in Working paper 21/02) has gone some way towards addressing this, but it is impossible to eliminate harmful online content, even if the original post has been deleted or redacted. The internet never forgets anything (von Kempis, 2017, p. 121).

[3] Commonly used messaging apps with end-to-end encryption include WhatsApp, Viber, Line, Telegram, Signal, KakaoTalk, Dust, Threema, Wickr, Cyphr, CoverMe, Silence, Pryvat Now, SureSpot and Wire.

[4] The 'alt-right' is also variously termed the 'far right', 'the radical right' and the 'extreme right'. For a simple taxonomy adopted by the Royal Commission of Inquiry (2020, p. 105), see Bjørgo & Ravndal, 2019.

[5] To illustrate this, the Royal Commission of Inquiry (2020, p. 11) cites a style guide for far-right website the Daily Stormer leaked in December 2017.

New Zealand's Chief Censor, David Shanks, has reflected that when the Film, Videos, and Publications Classification Act was passed in 1993, the internet was in its infancy:

> When you live in a world where about 500 hours of content is going up on YouTube every minute you instantly realise having a traditional classification approach of having someone sit and digest that fire hose of content is never remotely going to work. You need to think about new ways, new models to adapt to this digital reality … We're trying to figure out where the harms are and how to mitigate them in ways other than simply banning or restriction (quoted in Somerville, 2021).

Content blocking, moderation and banning by the major social media platforms tends to move harmful communication to lesser known and less well moderated and monitored platforms like Telegram, Twitch, Substack, Signal, Parler, BitChute, DLive, Minds, Gab and the imageboard Meguca. The Halle terrorist attacker, for example, told the court on the first day of his trial that setting up a Twitch live stream was 'the whole point'. He wanted to encourage others, just as he had been encouraged by the attack on Christchurch mosques. He added that he chose Twitch because he had learned from the Christchurch attacker that Facebook would take the stream down too quickly (Knight, 2020).

Following the 2020 US presidential election, 'millions' of conservatives reportedly reacted to fact-checking on Facebook and Twitter by migrating to alternative social media and media sites like Parler, Rumble and Newsmax (Isaac & Browning, 2020). DLive has become a haven for white nationalists, providing them with a platform to earn money from livestreaming (Browning & Lorenz, 2021). And Clubhouse, an audio chat app, has exploded in popularity but is grappling with harassment, misinformation and privacy issues (Griffith & Lorenz, 2021; Roose, 2021).

Cynthia Miller-Idriss (2020, pp. 139–145) discusses the complex tech ecosystem that underpins extremism. She explains that de-platforming and banning processes are more complex than many observers realise, and can in fact fuel extremism by:

> … driving it underground and reinforcing the far right's narrative of suppression, censorship, unfairness, and injustice. As mainstream platforms cracked down on users who violate their terms of use, those users migrated to unregulated, alternative platforms that more deeply incubate a hate-filled, dehumanizing culture; celebrate violence; circulate conspiracy theories; and help translate resentment, shame, or frustration into anger and calls to action (p. 141).

In this sense, the Christchurch Call's objective of eliminating terrorist and violent extremist content online is a global game of Whack-a-Mole—suppress it on one platform, and it will pop up somewhere else.

The business models of Big Tech

The business models that underlie Big Tech companies compound the challenge of regulating harmful digital communication. In a report on digital threats to democracy, Elliott, Berentson-Shaw, Kuehn, Salter, & Brownlie (2019, p. 14) singled out platform monopolies, algorithmic opacity and the 'attention economy'. A further question is whether social media and digital intermediaries should be treated as 'platforms' or as media content publishers and brought under existing regulatory frameworks (Flew, Martin & Suzor, 2019).

Platform monopolies

The regulatory problem is compounded by the global dominance of the so-called FAANG (Facebook, Apple, Amazon, Netflix, Google) or FAMGA (Facebook, Apple, Microsoft, Google, Amazon) digital media and communications platform companies and a tendency towards monopoly or oligopoly in the sector (Flew, Martin & Suzor, 2019, p. 34). A small number of large, powerful corporations control not only our means of communication but also the content that is distributed (Elliott, Berentson-Shaw, Kuehn, Salter, & Brownlie, 2019, p. 14).

This issue is coming into sharp focus through proposed regulatory measures in the EU (Riegert, 2020) and in Britain to pressure the world's biggest tech companies to take down harmful content and open themselves up to more competition. Adam Satariano (2020) reported that:

> One measure, called the Digital Services Act, proposed large fines for internet platforms like Facebook, Twitter and YouTube if they do not restrict the spread of certain illegal content like hate speech. Along with the similar proposal made earlier in the day in Britain, the debate will be closely watched. Large internet companies have faced years of criticism for not doing more to crack down on user-generated content, but governments have been reluctant to impose laws banning specific material for fear of restricting speech and self-expression.[6]

Shoshana Zuboff (2021) reports that in the US, five comprehensive bills, 15 related bills and a draft Data Accountability and Transparency Act were introduced in Congress from 2019 to mid-2020. Each has material significance for what she terms 'surveillance capitalism' (Zuboff, 2019).

Then-Presidential candidate Senator Elizabeth Warren announced on March 8, 2019 a campaign to break up tech monopolies Amazon, Facebook and Google and to restore competition in the sector (Warren, 2019). In June 2019, the US House of Representatives Judiciary Committee initiated what became a 16-month bipartisan investigation into whether Google, Facebook, Amazon and Apple broke the law to squash competition. The investigation, spear-headed by the Sub-committee on Antitrust, Commercial and Administrative Law, issued a staff report in October 2020 that found Google and Facebook, as well as parts of Amazon and Apple, to be monopolies (House Judiciary Committee, 2020). The report recommended restoring competition in the digital economy, strengthening anti-trust laws and reviving anti-trust enforcement.[7] Bipartisan support broke down, however, and the report was signed only by Democrat members. Republicans were divided and issued a separate draft report (Kang & McCabe, 2020; Ovide, 2020).

In October 2020, the Department of Justice, joined by 11 states, initiated a federal anti-trust suit against Google for abuse of its online search monopoly. By December, the Federal Trade Commission had filed a landmark lawsuit against Facebook, which also owns Instagram, for anti-competitive actions, joined by a suit from 48 state attorneys general (Bensinger, 2021a). Soon after, a suit launched by 38 attorneys general challenged Google's core search engine as an anti-competitive means of blocking rivals and privileging its own services (Zuboff, 2021). Private anti-trust lawsuits are now leveraging evidence uncovered in the government cases against Google and Facebook (McCabe, 2021a). Regulators are also investigating Apple and Amazon (Isaac & Kang,

[6] On the European Union's Digital Services Act, see further Working paper 21/05, **Regulating harmful communication: Current legal frameworks.**

[7] Regulation of competition is often referred to in the US as 'anti-trust law'; in the EU, it is referred to as both anti-trust and competition law.

2020). These anti-trust cases will, however, be difficult to prove. The Big Tech companies have deep pockets and are powerful lobbyists (Satariano & Stevis-Gridneff, 2020; Woodhouse & Brody, 2020).

Przemysław Pałka (2021) envisions a 'world of fifty facebooks', where numerous companies would offer interoperable services similar to what Facebook currently provides. Just as we can call and text one another using different telco providers, users of A-Book should be able to find, communicate with and see the content of customers of B-Book, C-Book, etc. He argues that Facebook should be legally obliged to allow potential competitors to become interoperable with its platform, instead of using its monopolistic position to impose excessive costs and unnecessary harms on consumers and on society.

Algorithms and rabbit holes

Social media and other digital intermediaries primarily exist to sell personalised advertising based on their collection of vast amounts of personal data, through clever programming of algorithms. They also happen to provide worldwide communication platforms, for which they do not charge a fee (Schwartmann, 2021). Users benefit from ostensibly 'free' products, like search engines and communication platforms, but in fact pay for them by surrendering their personal data.

Algorithmic engines use this personal data to make ever more precise predictions about what we want to see and hear, and to influence what we think and do. There is little or no transparency about how these algorithms work, or accountability for their impact. The dominant business model plays to the 'attention economy', prioritising and amplifying whatever content is best at grabbing users' attention, while avoiding responsibility for the impact that content has on our collective wellbeing and our democracy (Steinmeier, 2021; Elliott, Berentson-Shaw, Kuehn, Salter, & Brownlie, 2019, p. 14).

Algorithms do not just filter content. Digital intermediaries use them primarily to optimise traffic by recommending other content that viewers may be interested in. Guillaume Chaslot, founder of AlgoTransparency and an advisor at the Center for Humane Technology, worked on YouTube's recommendation algorithm from 2010 to 2011. He explains that YouTube's AI tracks and measures the viewing habits of the user, and users like them, to find and recommend other videos they will engage with. The objective is to increase the amount of time people spend on YouTube (Chaslot, 2019). The stronger the AI becomes—and the more data it has—the more efficient it becomes at recommending user-targeted content. As soon as the AI learns how it engaged one person, it reproduces the same mechanism on other users. The user then delivers what the algorithm rewards (Bender, 2021).

While YouTube's 'up next' feature receives the most attention, other algorithms are just as important, including search result rankings, homepage video recommendations and trending video lists (Matamoros-Fernández & Gray, 2019). Fyers, Kenny & Livingston (2019) explain:

> It's in YouTube's interests to keep users engaged for as long as possible. More watch-time equals more revenue for the platform's parent company, Google. In the same way Amazon predicts consumers' buying habits and Netflix and Lightbox predict the types of shows subscribers want to watch, YouTube's recommendation algorithm suggests videos most likely to keep viewers hooked. The algorithm seems to have worked out that the best way to keep someone watching is to feed them content that reaffirms their existing views, while supplementing with more extreme ones.

Technology sociologist Zeynep Tufekci noticed during the 2016 US presidential election campaign that when she watched YouTube videos of Donald Trump rallies, YouTube started to recommend videos that featured white supremacist content and Holocaust denials. When she created another YouTube account and watched videos of Hillary Clinton and Bernie Sanders, YouTube's recommender algorithm took her to 'leftish' conspiratorial content that also became more and more extreme, including arguments about the existence of secret government agencies and allegations that the US government was behind the attacks of September 11, 2001. When she experimented with non-political topics, the same pattern emerged: 'Videos about vegetarianism led to videos about veganism. Videos about jogging led to videos about running ultramarathons' (Tufecki, 2018).

Tufekci concludes that 'YouTube leads viewers down a rabbit hole of extremism, while Google racks up the ad sales.' The business model of online platforms monetises 'long-tail marketing' that leads inevitably from the relatively mainstream to what is less so (Munger & Phillips, 2019; Friedersdorf, 2018).[8]

Fyers, Kenny & Livingston (2019) report that after searching for news coverage of the Christchurch terrorist attack and watching a couple of videos, YouTube's 'up next' reel suggested content by controversial figures such as psychologist Jordan Peterson, libertarian gun rights activist Kaitlin Bennett, conservative commentator Katie Hopkins and far-right pundit Lauren Southern. One of those videos then led to anti-immigration, anti-feminist and transphobic content, then to a clip of a British anti-Islam activist confronting people at a mosque, referring to them as 'hate preachers'. Comments below were overtly racist and some incited violence.[9]

Ledwich & Zeitsev (2019) in a self-published article found that YouTube's recommendation algorithm does not, in fact, promote inflammatory, radicalising or extremist content as previously claimed, and instead favours mainstream media and cable news content, with a slant towards left leaning or politically neutral channels. A flaw in their methodology, however, is that they did not log in to a YouTube account or accounts (Hale, 2019; Feuer, 2019). A 2019 study out of Cornell University did find evidence of user radicalisation on YouTube (Ribeiro, M., Ottoni, R., West, R., Almeida, V., & Meira, W., 2019).

Research conducted for the Global Research Network on Terrorism and Technology, a partner of the Global Internet Forum to Combat Terrorism (GIFCT),[10] investigated three social media platforms—YouTube, Reddit and Gab—to establish whether the sites' recommender systems promote extremist material (Reed, Whittaker, Votta & Looney, 2019). Their research found that YouTube's recommender system does prioritise extreme right-wing material after interaction with similar content but did not find evidence of this effect on either Reddit or Gab. This is significant because Gab has been identified by many as a haven for the alt-right. The research finding suggests (p. 2) that it is the users, rather than the site architecture, that drives extremist content on Gab.

The real issue with algorithms is not their use but their lack of transparency. We simply do not know what role YouTube's recommendation algorithm plays in nudging people towards more extreme

[8] On the tension between the public and commercial rationales built into social media platforms, see Stockmann (2020).

[9] See also Kevin Roose's report of how Caleb Cain 'fell down the alt-right rabbit hole' (Roose, 2019), and the *New York Time's* 'Rabbit hole' podcast series (Roose, 2020).

[10] On the GIFCT, see Working paper 21/02, **The terrorist attack on Christchurch mosques and the Christchurch Call**, pp. 8–10.

views because we do not understand how it works. Camargo (2020) notes that YouTube is not unusual in this respect:

> A lack of transparency about how algorithms work is usually the case whenever they are used in large systems, whether by private companies or public bodies. As well as deciding what video to show you next, machine learning algorithms are now used to place children in schools, decide on prison sentences, determine credit scores and insurance rates, as well as the fate of immigrants, job candidates and university applicants. And usually we don't understand how these systems make their decisions.

Camargo argues that 'we need to open up these patented technologies, or at least make them transparent enough that we can regulate them.' While introducing counterfactual explanations or auditing algorithms is a difficult, costly process, he argues that the alternative is worse: 'If algorithms go unchecked and unregulated, we could see a gradual creep of conspiracy theorists and extremists into our media, and our attention controlled by whoever can produce the most profitable content' (Camargo, 2020).

'Moral panic' about social media and algorithms, however, overstates the case. The extent of our exposure to diverse points of view is more closely related to our social groupings than to recommendation algorithms:

> At the end of the day, underneath all the algorithms are people. And we influence the algorithms just as much as they may influence us (Mitchell & Bagrow, 2020).

Matamoros-Fernández and Gray (2019) remind us, too, that content creators are not passive participants in algorithmic systems:

> They understand how the algorithms work and are constantly improving their tactics to get their videos recommended. Right-wing content creators also know YouTube's policies well. Their videos are often 'borderline' content: they can be interpreted in different ways by different viewers. YouTube's community guidelines restrict blatantly harmful content such as hate speech and violence. But it's much harder to police content in the grey areas between jokes and bullying, religious doctrine and hate speech, or sarcasm and a call to arms.

Generating income in the attention economy

Those who engage in harmful digital communication appear able to assume a low risk to themselves, while receiving relatively high rewards in terms of social recognition within their digital 'echo chambers' (Kaspar, 2017, p. 68). The 'attention economy' that rewards content with likes, followers and re-posts exerts its own 'gravitational pull', quite apart from those of algorithms and rabbit holes. Ben Smith reflects in a story on Tim Gionet, a former co-worker and white nationalist who figured in the occupation of the US Capitol on January 6, 2021:

> If you haven't had the experience of posting something on social media that goes truly viral, you may not understand its profound emotional attraction. You're suddenly the center of a digital universe, getting more attention from more people than you ever have. The rush of affirmation can be giddy, and addictive. And if you have little else to hold on to, you can lose yourself to it (Smith, 2021a).[11]

[11] Thompson and Warzel (2021) illustrate how the 'attention economy' played out on Facebook, in relation to the riots in Washington DC on January 6, 2021.

Banned from Twitter and YouTube, Gionet made more than US$2000 from livestreaming his participation in the occupation of the Capitol on DLive on January 6, 2021.[12] DLive reportedly has since suspended, forced offline or limited 10 accounts and deleted 100 broadcasts, and is freezing the earnings of streamers who broke into the Capitol:

> But streamers and misinformation researchers said DLive's emergence as a haven for white nationalists was unlikely to change. That's because the site, which was founded in 2017 and is similar to Twitch, the Amazon-owned platform where video gamers livestream their play, helps streamers make tens of thousands of dollars and benefits by taking a cut of that revenue (Browning & Lorenz, 2021).

Because other sites like Parler and Gab do not offer ways to make money, streaming on DLive has become a key strategy for the alt-right. Browning and Lorenz (2021) cite research showing that while most donations are small amounts of money, some donors are giving US$10,000–20,000 a month to streamers on DLive, with top streamers on the platform earning six-figure incomes in 2019.

Publisher or postie?

Should social media and digital intermediaries be treated as media content publishers and be brought under existing regulatory frameworks (Stockmann, 2020; Flew, Martin & Suzor, 2019)? Many digital intermediaries argue that they are not media content publishers and that they should therefore have only limited liability for the content on their platforms (Gargliadoni, Gal, Alves, & Martinez, 2015, p. 30). If they are publishers and not merely platforms, they would have a greatly expanded and very expensive duty to moderate all the content that appears on their sites (Daley, 2021).

There is a difference between content moderation and journalistic or editorial work and publishing (Winseck, 2019), but as New Zealand Prime Minister Jacinda Ardern said days after the Christchurch mosque shootings:

> We cannot simply sit back and accept that these platforms just exist and that what is said on them is not the responsibility of the place where they are published. They are the publisher. Not just the postman. There cannot be a case of all profit, no responsibility (quoted in Small, 2019).

Winseck (2019) argues, however, that content moderation is not the same as journalistic or editorial work or publishing, and that governments should use existing laws to constrain extremist content or enact new laws to fill any identified gaps. He suggests:

- Functionally equivalent data and privacy protection rules across all layers and players in the internet stack;
- Regulated algorithm audits and algorithm transparency;
- Information fiduciary obligations on digital platforms, similar to banks, and perhaps like the international banking system, multinational digital platforms should set up local branches accountable to both domestic and international regulations; and
- Advertising 'white lists'—with a requirement on top advertisers to use regularly updated 'whitelists' of URLs to determine where their advertising dollars go instead of relinquishing control to tech companies' algorithms.

[12] Gionet, known in alt-right circles as 'Baked Alaska', was arrested by the FBI after live-streaming the Capitol riot on January 6, 2021 (Associated Press, 2021).

There is a broader issue about whether digital intermediaries 'publish' or 'share' content. France has legislated a 2019 EU ancillary copyright directive into national law.[13] In January 2021, in a European first and after months of negotiation, Google signed an agreement with a group of French newspapers to enter into individual license agreements with newspapers on digital copyright payments for 'sharing' their content through online searches (Deutsche Welle, 2021a).

News of this complicated a stand-off between Google and the Australian Government over a bill that would require Google and Facebook to pay license fees to Australian media companies for sharing their journalistic content (Cave, 2021a; Keall, 2021b; Klein, 2021). Google threatened to make its search engine unavailable in Australia if the Government enacted the bill in its current form. Christopher Niesche (2021) reported that:

> The internet giant earns A$4.8 billion a year in Australia, and the bulk of that—A$4.3 billion—comes from internet search. The fact that Google says it is willing to forgo this revenue shows what's at stake for Google. If the Australian Government succeeds in making the internet giant pay, then other governments around the world are likely to follow.

In February 2021, Australia's Seven West Media struck a deal with Google, under which the internet giant will pay for journalism in a 'long-term partnership', announced following discussions Australian Government ministers had with media executives, Facebook CEO Mark Zuckerberg, and Sundar Pichai, chief executive of Alphabet Inc. and its subsidiary Google (McGuirk, 2021). Other agreements have followed, including with Rupert Murdoch's News Corp (Cave, 2021b). Facebook took a different line, however, blocking users from reading or sharing news content on the platform from February 17, 2021 (Waters, Murphy & Barker, 2021). As well as news publishers being blocked, pages for Fire and Rescue New South Wales, the Bureau of Meteorology and state police departments were all wiped clean. Even state government pages with public health information about the Covid-19 pandemic were blocked (Isaac, Wakabayashi, Cave & Lee, 2021). By February 22, 2021, however, Facebook had reached an agreement with the Australian Government and agreed to restore news links and articles for Australian users (Easton, 2021; Isaac & Cave, 2021).

Subsequently, Australia's Parliament passed the Treasury Laws Amendment (News Media and Digital Platforms Mandatory Bargaining Code) Act 2020 on February 24, 2021, and Facebook announced it plans to invest $1 billion over three years to 'support the news industry' (Clegg, 2021). A fully nuanced account of the stoush between Facebook and the Australian Government and some of the interests at play is only beginning to emerge (e.g., Barnes, 2021; Murphy, Waters, Barker, & Smyth, 2021; Griffin, 2021).

The Australian dispute has significant implications for internet governance and the 'splintering' of the internet (Swisher, 2021; Ovide, 2021c; Collins, 2021). Canada is likely to adopt the Australian model and is talking with a number of other countries, including France, Germany and Finland (Deutsche Welle, 2021b). New Zealand Broadcasting Minister Kris Faafoi met with Google and Facebook in late February 2021 and encouraged them to have commercial discussions with traditional media as is happening in Australia (Young, 2021).

[13] The German federal government has also begun the process of enacting the EU directive into national law (Budras, 2021).

The complicated business of de-platforming

This section discusses whether de-platforming works, the de-platforming of then-President Donald Trump in January 2021, and how governments, tech companies and civil society might prevent abuse of the internet and maintain a healthy tech ecosystem. It concludes that de-platforming is only a part-solution to complex social problems and processes.

De-platforming here means 'the technical act of disabling social media profiles and rendering their content inaccessible to other users' (Fielitz & Schwarz, 2020, p. 12; Rogers, 2020, pp. 214, 226), rather than the wider application of the term in 'cancel culture', including withdrawing speaking invitations or denying the use of venues.

Does de-platforming work?

Chen (2020b, pp. 172–173) cites an analysis of Reddit's 2015 ban of the subreddits r/fatpeoplehate and r/CoonTown. Working from over 100 million Reddit posts and comments, Chandrasekharan, Pavalanathan, Srinivasan, Glynn, Eisenstein, and Gilbert (2017), found that the ban was effective. More accounts than expected discontinued using the site; those that stayed decreased their hate speech usage by at least 80%. Though many subreddits saw an influx of r/fatpeoplehate and r/CoonTown migrants, those subreddits saw no significant changes in 'hate speech' usage. In other words, other subreddits did not inherit the problem.

A 2019 analysis of Twitter's aggressive account and content takedown in relation to the so-called Islamic State (IS/Daesh) also found that the costs for most pro-IS users of engaging on Twitter (in terms of deflated morale, diffused messages and persistent effort needed to maintain a public presence) now largely outweigh the benefits, and that the IS Twitter community had become almost non-existent (Conway, Khawaja, Lakhani, Reffin, Robertson, & Weir, 2019). A hard core of IS users does, however, persist on Twitter. The authors also caution that IS are not the only jihadists active online and 'a host of other violent jihadists were shown to be subject to much lower levels of disruption by Twitter' (p. 155).

A recent report from the Institut für Demokratie und Zivilgesellschaft (IDZ) in Jena investigates whether, how and to what extent restricting access to social media platforms, particularly the preferred platforms of YouTube, Facebook and Twitter, impacts far-right hate actors in Germany (Fielitz & Schwarz, 2020). This was the first investigation into the effects of de-platforming on the far right in Germany (p. 66). As Fielitz and Schwarz report their findings (Executive Summary, p. 5):

- Deplatforming key far-right actors significantly limits their networked mobilization capacities and denies them a key resource for attaining what they seek most: attention. Based on our data, we may unequivocally state that deplatforming does work in this sense.
- At the same time, hate actors are neither caught by surprise nor unprepared for repressive policies enforced by mainstream platforms. On the contrary: they have developed innovative approaches for responding to such measures, which signalizes their ability to adapt and act. Strategies includes semantic mimicry tactics, the (audio)visualization of propaganda, the creation of fake accounts, the utilization of proxies, recourse to alternative platforms, and the establishment of own digital infrastructures.

- Alternative networks are used for spreading deleted content on large-scale platforms. In this manner, hate actors steer their followers and content across a variety of different platforms.
- The messaging app Telegram has become the most important online platform for hate actors in Germany: 96% of the actors we investigated in our study operate (super)active channels on this platform. Most consider Telegram to be their primary basis for communication. What makes this hybrid platform so special is that 1) it is hardly subject to any forms of moderation by the operator, 2) it displays push notifications directly on phone screens and 3) users can easily switch between its private messaging and public channel functions.
- Resorting to alternative platforms does not completely compensate for deletion from the digital mainstream. Apart from Telegram, no other stable alternative forums have established themselves for the German far right. As such, the ideas they promote are not as effectively circulated on the available alternative platforms because few German-speaking users actually use them.
- In order to avoid the unintended consequences that arise as a result of deplatforming—such as generating even more attention for hate actors and raising their perceived level of importance—measures must be better coordinated, more clearly communicated and implemented in accordance with universal human rights.

Significantly, de-platforming imposes a financial cost on 'hate actors', prompting them to call for donations to offset these costs (p. 42). Of course, to have this effect consistently, de-platforming would also need to target platforms like DLive, Patreon, PayPal, SubscribeStar, Ko-fi, Buy Me a Coffee, CashApp and use of cryptocurrency donations, and not just Big Tech companies like YouTube that allow users to generate income online (Andrews & Pym, 2021).[14]

The de-platforming of President Donald Trump

Following the storming of the US Capitol on January 6, 2021 and the role of President Donald Trump in inciting it as he tried to overturn his defeat in the 2020 presidential election, Twitter permanently banned Trump from using its platform and in the following days said it had removed more than 70,000 accounts that promoted the QAnon conspiracy theory (Twitter Safety, 2021; Conger, 2021).[15] In a statement, Trump said Twitter was trying to silence him and that he was negotiating with other sites and looking at building 'our own platform' (Conger & Isaac, 2021; Browning & Lorenz, 2021).

Facebook and Facebook-owned Instagram indefinitely blocked Trump's ability to post on both Facebook and Instagram on January 7 and subsequently referred its decision to do so to its Independent Oversight Board for review (Dodds, 2021). Snapchat, Pinterest, Reddit, YouTube, Twitch and Shopify also limited Trump's access to their services.

The advocacy coalition Stop Hate for Profit had launched a campaign to pressure the major platforms, including YouTube owner Google, to ban Trump from their services. The organisation, which includes the Anti-Defamation League, the NAACP, the National Hispanic Media Coalition, Free Press and Colour of Change, said it would call for an advertiser boycott if the platforms did not take

[14] Andrews and Pym (2021) note that websites like Amazon, Audible and Teespring/Spring that permit the alt-right to sell books and merchandise would also need to be in scope.

[15] Twitter's clampdown included suspending the accounts of hundreds of New Zealand-based users, many of whom voiced right-wing political opinions (Walls, 2021).

action by January 20, the date of President-elect Joe Biden's inauguration (Stop Hate for Profit, 2021).

Parler, a social network app that markets itself as a 'free speech' alternative to Twitter and Facebook, has become a refuge of conservatives and the alt-right since Twitter started moderating Trump's posts in May 2020 in the lead up to the US presidential election. Sensor Tower, an app data firm, estimates that Parler's app was downloaded more than 10 million times in 2020, with 80 per cent of those downloads in the United States (Nicas & Alba, 2021b).

Following the occupation of the Capitol, Google and Apple removed Parler from their app stores for Parler's failure to enforce its own moderation policies. Amazon said it would no longer host the site on Amazon Web Services (Nicas & Alba, 2021a, 2021b). Parler unsuccessfully sought an injunction in the district court in Seattle. Judge Barbara Rothstein said she was not dismissing Parler's 'substantive underlying claims' against Amazon but said it had fallen short in demonstrating the need for an injunction forcing it back online (O'Brien, 2021). Jack Nicas (2021a) has commented that 'the edicts from Apple and Google were a stark illustration of the power of the largest tech companies to influence what is allowed on the internet, even on sites and apps that are not their own.'[16]

German Chancellor Angela Merkel described Trump's de-platforming by private companies as 'problematic' (Hanfeld, 2021). At one level, a private company is well within its rights to terminate a contract that a user has entered into voluntarily, on terms and conditions specified by the company, but for all the awfulness of Trump's communications it was extraordinary and unprecedented for an incumbent, democratically elected head of state to be blocked from communicating with tens of millions of followers through the world's most popular online services.

James Titcomb (2021) cautions:

> Some 74m people voted for a man who no longer has a place on Facebook or Twitter—and a sizeable proportion of them will now look elsewhere. The rest of us should have concerns, too. The more the online world fragments, the easier it is for radical and dangerous parts of it to spring up, and the harder it is for law enforcement to monitor. Students of the web have for years observed the gradual rise of the 'splinternet', in which the online world fractures along geographical lines, with China, Europe and the US all going in separate directions.

Fringe groups have moved their messaging to Gab, 4chan and encrypted messaging apps like Telegram and Signal, which cannot be as easily monitored as social media platforms (Frenkel, 2021; Nicas, Isaac & Frenkel, 2021; Bleiker, 2021). Several start-ups have promised 'unbiased' and 'free speech' social networks, but the tougher enforcement from Big Tech companies could prevent them from becoming realistic alternatives to more mainstream digital intermediaries. Nicas & Alba (2021a) comment:

> They now face the choice of either stepping up their policing of posts—undercutting their main feature in the process—or losing their ability to reach a wide audience. That may reinforce the primacy of the social-media incumbents, Facebook, Twitter and Instagram. It

[16] Nicas (2021b) subsequently reported that Parler entered into business with DDoS-Guard, a Russian firm, to provide at least temporary web hosting, but users were unable to post. CEO John Matze said in February 2021 that he had been fired because of a difference of opinion with prominent Republican political donor Rebekah Mercer, who supports Parler financially (Browning, 2021). Parler went back online on February 15, 2021, hosted by SkySilk (Nicas, 2021c).

> also gives those companies' decisions more teeth. If they ban a pundit for violating their rules, that person will lack a strong alternative.

This has the effect of both strengthening the anti-competitive actions of Big Tech companies and allocating decisions about freedom of expression and censorship to private companies whose primary imperative is to generate profit.

Deutsche Welle Editor-in-Chief Manuela Kasper-Claridge (2021) argued that pulling the plug on Trump was too little, too late for platform operators. It was their job all along to identify, label or delete 'hate speech' and fake news and, she argues, they need to be regulated democratically and held to account, to ensure that they do so:

> It makes me nervous that a small group of people can decide to slam the door shut on the world's most influential communication platforms … Let's also not forget that these social networks are an important tool for expressing opinions, especially in countries with limited freedom of the press. However, it is not democratic or a sign of plurality when a few company bosses, who are ultimately only responsible to their shareholders, dominate markets and use their power to decide on freedom of expression.

Facebook's Oversight Board, an independent content moderation body established and funded by Facebook, will appoint a panel of five board members, at least one of them American, to decide whether Facebook is to reinstate Trump's account or block him from using the platform indefinitely. The full, 20-person board, members of which reportedly do not feel obligated to Facebook's shareholders, will review the panel's decision (Smith, 2021b).

While the Oversight Board provides a measure of independent accountability for Facebook's decisions (Ovide, 2021b), in effect, this is a new form of transnational, corporate governance acting as a court to determine the boundaries of free speech.[17] Freedom of expression is a qualified right,[18] but decisions to restrict freedom of expression should be made within a framework of laws defined by democratically elected legislators and be open to review and appeal—not by private companies acting as courts to decide for billions of people what is appropriate and 'legal' or 'illegal' expression and behaviour (Economist, 2021; Ovide, 2021a; Stockmann, 2020, p. 256). As Fraser Myers (2021) has commented: 'If the tech monopolies can deny a platform to the leader of the free world, then they can deny a voice to anyone.'

Chris Keall (2021a) reports that the American Civil Liberties Union (ACLU), often in conflict with Trump, has told the *New York Times*: 'We understand the desire to permanently suspend him now, but it should concern everyone when companies like Facebook and Twitter wield the unchecked power to remove people from platforms that have become indispensable for the speech of billions.'

[17] Greg Bensinger (2021b) reports that results from the Oversight Board to date are 'underwhelming'. See also Boland (2021). Jaffer and Bass (2021) argue that the fundamental problem with the Oversight Board is that many of the content moderation decisions the board has been charged with reviewing cannot be separated from platform design decisions and ranking algorithms that Facebook has placed off limits.

[18] See further Working paper 21/05, **Regulating harmful communication: Current legal frameworks**.

UCLA law professor Eugene Volokh (2021) explains:

> On the one hand, deplatforming a user like Mr. Trump is perfectly legal, and the perils of corporate power are often exaggerated.[19] On the other hand, these companies are exercising a sweeping ability to silence all of a politician's speech, not just the dangerous parts. This would be condemned as prior restraint—that is, an action forbidding a wide range of future speech, rather than punishing a specific past statement—if done by the government. Furthermore, these companies are doing this in an environment of limited competition and with little transparency, procedural protection or democratic accountability.

Russian opposition politician Alexei Navalny criticised Twitter's ban on Trump as selective and 'an unacceptable act of censorship', commenting that while Twitter is a private company, there is precedent in Russia and China for such companies becoming the government's best friends and enablers of state censorship (Dambach, 2021). Navalny called for platforms to create a more transparent process, appointing committees whose decisions could be appealed (Goldberg, 2021).

New Zealand's Privacy Commissioner, John Edwards in Twitter posts from his personal account on January 8, 2021, also criticised the ban (Keall, 2021a):

> The Twitter and Facebook bans are arbitrary, cynical, unprincipled and further evidence that regulation of social media platforms is urgently required. Much worse has been allowed, and is still present on both platforms than the precipitating posts. We should not be abdicating responsibility for the tough policy decisions required, and delegating responsibility for our community standards to conflicted corporates.

The roles of governments, tech companies and civil society

Because de-platforming interferes with the right to freedom of expression, decisions about the regulatory framework that governs censorship and de-platforming best sit with democratically elected governments.

Fielitz and Schwartz (2020, p. 63) conclude that 'we must question the power wielded by platforms which make decisions without democratic legitimation and at the boundaries of public discourse'. Nadine Strossen (2018, p 30) similarly cautions against relegating responsibility for censorship to private sector actors that are not directly subject to constitutional constraints:

> This is most important for private-sector entities that are engaged in communications activities, and that in turn affect the communications opportunities of others. Prime examples are private universities and online intermediaries, including internet service providers, search engines, and social media platforms. In general, they should hesitate to bar any expression that government could not bar.

The (relatively) easy part is setting the bottom line—requiring social media companies and other digital intermediaries to block and remove content that is illegal:

> It's not unreasonable to conclude that it's for the Government to regulate to the extent that content or behaviour is illegal regardless of whether it's online or offline. Unless you're an anarchist, that's a reasonable position to take and it's not too big a jump to then expect

[19] Volokh (2021) explains: 'We also shouldn't overstate the danger of corporate power. Facebook and Twitter, unlike the government, can't send us to jail or tax us. But at least governmental speech restrictions are implemented in open court, with appellate review.'

> providers to have a role in preventing the publishing of content that is clearly illegal (Matthews, 2021).

Even this is not straightforward, however, because the definition of illegal content varies across jurisdictions. Co-operation between governments and tech companies can smooth this difficulty, but not entirely remove it.

In the United States, where many tech companies are headquartered, Section 230 of the Communications Decency Act shields them from liability for content their users post online. Various proposals are before Congress to amend Section 230 and tackle disinformation or 'hate speech' online without unduly interfering with free speech rights (McCabe, 2021b).

Tech companies cannot, however, leave all the rule-making and enforcement to governments—and in fact they do not. Rachel Lerman (2020) noted, for example, that even Parler, 'the poster child for free expression online', has since its inception had a long list of community guidelines that outline what it won't allow, including 'obscenity, terrorist content and "fighting words," or calls to incite violence'. She reflected (in July 2020) that Parler is now 'facing the same evolution bigger social media companies have confronted for years—balancing free expression with creating safe and inviting online communities':

> Online conversations are complicated. Facebook has a six-part document outlining its community standards. Twitter has eight separate sections under its set of rules that just oversee safety. Facebook and Google-owned YouTube pay tens of thousands of content moderators to review material on their sites and enforce their policies.

A report by the Election Integrity Partnership on analysis of online misinformation and disinformation that eroded confidence in the 2020 US presidential election includes four recommendations for social media platforms and technology companies:

- Accessibility—tell users about the platform's misinformation policies and provide both rationale and case studies;
- Transparency—share platform research; enable access for external researchers to removed or labelled content, including exhaustive and rapid search capabilities; partner with civil society organisations; and provide greater transparency about why something is removed or censored;
- Support independent cross-platform coalitions and co-ordinate with government officials and civil society to respond to growing narratives; and
- Set a higher bar for people with the most influence and repeat offenders, and publicise the different thresholds of policy offenses (Election Integrity Partnership, 2021, pp. 232–233; Ovide, 2021e).

Governments can encourage, and in some cases may be able to require, digital platforms to set and enforce transparent rules that help maintain a fair balance between the right to freedom of expression and protection from harm. There are risks for freedom and democracy if the state controls the governance of the internet and digital communications. And there are risks if the state does not step up and work with tech companies and civil society to develop, govern and maintain a healthy internet ecosystem.

Some of what governments can achieve is most effective when done at arm's length, for example by encouraging and supporting counter-speech and civil society initiatives as alternatives or

complements to regulation.[20] Options include investment in public education programmes in civics, human rights, conflict resolution and digital literacy; building stronger partnerships with communities, civil society groups, public sector institutions and industry; reducing inequalities and marginalisation on all fronts, with outreach, early intervention and rehabilitation to prevent extremism from taking root; well-funded public broadcasting that provides access to authoritative information and diverse ideas and opinion; and publicly funded, moderated, public-service social media for online civic discussions.[21]

On counter-speech strategies, and civility as everyone's responsibility, see Working paper 21/08.

Technology may provide a part solution

Tim Berners-Lee, inventor of the World Wide Web, thinks too much power and too much personal data now resides with Big Tech companies, enabling them to become surveillance platforms and gatekeepers of innovation (Berners-Lee, 2019; Lohr, 2021). His proposed solution is not more regulation, or even necessarily anti-trust suits to break up monopolies, but a global Contract for the Web that outlines steps to prevent the deliberate misuse of the web and our information, and technology ('PODS' or Personal Online Data Stores) that gives individuals more control over their data and moves towards the web he originally envisaged.[22] Berners-Lee began an open-source software project, Solid, and has founded a company, Inrupt, to kick-start adoption, with pilot projects underway in 2021 for Britain's National Health Service and the government of Flanders (Lohr, 2021; Itega, 2020).

De-platforming is not a silver bullet

To sum up this section, while de-platforming does make some difference, it can only be a part solution to complex social problems and processes. It has unintended consequences and comes with a host of challenges, including under- and over-inclusion and removal of legitimate content, especially by automated moderation processes. Hard questions remain about the right mix of responsibilities between government, business and society in the governance of a healthy internet ecosystem.

De-platforming extremist content does not, in any case, address social problems and processes that occur in complex online and offline interactions and provide the context for extremism and radicalisation. Fielitz and Schwartz, in their report on the impact of de-platforming on far-right hate actors in Germany, conclude:

> While effective, the practice of deplatforming far-right actors is not in itself a sufficient response to their encroaching strategies on the Internet. Deleting a post or an account will neither change people's opinions nor prevent acts of violence or radicalization from taking place, per se. Despite this, the way that social media platforms deal with far-right actors has

[20] The Center for Countering Digital Hate, for example, develops counter-strategies to address identity-based hate that polarises and undermine democracies worldwide. See its report, *#DeplatformIcke* (CCDH, n.d.). Other examples will be provided in Working paper 21/08, **Counter-speech and civility as everyone's responsibility**.

[21] See Appelbaum & Pomerantsev (2021) for a range of ideas on why and how 'the internet does not have to be awful'.

[22] Stockmann (2020, p. 259) reviews proposals to change the ownership structure of the data that forms the core of social media companies' business models.

> an immense influence on their effectiveness in promoting policies and disseminating anti-democratic propaganda (Fielitz & Schwarz, 2020, p. 67).

Cynthia Miller-Idriss (2020, p. 138) cautions against the stereotype that 'radicalization happens primarily to isolated, lone teenagers who stumble into nefarious parts of the dark web from a gloomy corner of their bedroom lit only by the glow from their screen.' She explains: 'The story of how the far right has created, cultivated, and weaponized the internet is much more complex, relying on a strategic combination of online and off-line activities that enables the far right to maximize the circulation, communication, and effectiveness of far-right ideologies' (ibid.).

Miller-Idriss notes that while online spaces offer 'training, advice, how-to guides, ideological materials, and places where violent attacks are livestreamed, downloaded, circulated, and celebrated … online spaces work in tandem with in-person gatherings that also enhance global interconnections, such as transnational music festivals, conferences, MMA [Mixed Martial Arts] tournaments, and festivals associated with or linked to white-supremacist scenes' (p. 21).

She discusses, for example, the expression of far-right ideology in food, fashion, music, publishing and marketing of an 'aesthetic' and associated commercial products. In Germany, attempting to ban extremist messaging in a mainstream aesthetic has led to even more game playing: 'Symbols and messaging have become ever-more coded in order to subvert bans in ways that are playful, humorous, and fun for designers and consumers' (p. 91). She also notes (p. 128) the 'ever-expanding intellectual ecosystem of [far-right] dedicated think tanks, publishing houses, research grants, magazines, conferences, and more.' Consequently, 'significant research is needed to disentangle the complex web of interactions that characterizes the mix of online and off-line influences in radicalization toward violence' (p. 159).

Policy challenges for New Zealand

A report published by the Helen Clark Foundation (Mason & Errington, 2019) identified gaps both in the regulation of social media companies,[23] and in current protections in New Zealand law against 'hate speech' (based on religious beliefs, gender and sexual identity) and hate-motivated crimes. The report recommended that the New Zealand Law Commission review laws governing social media in New Zealand. The authors consider a legislative response necessary 'because ultimately there is a profit motive for social media companies to spread "high engagement" content even when it is offensive, and [because of] a long standing laissez faire culture inside the companies concerned which is resistant to regulation' (p. 6).

Elliott, Berentson-Shaw, Kuehn, Salter & Brownlie (2019) similarly identified a relative neglect of public policy-making on digital media, and the need for a better system for making policy (p. 7). Governments need to look beyond immediate concerns of terrorist and violent extremist content and take into account wider structural issues, including in particular:

- The impact of platform monopolies, in which a handful of people have the power to determine social interactions and access to information of millions of people;

[23] Mason & Errington (2019) note (p. 14) that social media companies in New Zealand are largely left to self-regulate how they monitor and remove harmful content in line with internal policies and guidelines. If the content is objectionable, a patchwork of New Zealand agencies has limited oversight powers: the Privacy Commission, the Ministry of Justice, the Department of Internal Affairs and Netsafe.

- Algorithmic opacity, in which algorithms have ever-increasing influence over what we hear and see without appropriate transparency or accountability; and
- The attention economy, which gives priority to content that grabs attention, without sufficient regard to potential harm (p. 8).

Their report (p. 25) identifies six Rs as urgent areas for change:

- **Restore** a genuinely multi-stakeholder approach to internet governance, including rebalancing power through meaningful mechanisms for collective engagement by citizens/users;
- **Refresh** antitrust and competition regulation, taxation regimes and related enforcement mechanisms to align them across like-minded liberal democracies and restore competitive fairness, with a particular focus on public interest media;
- **Recommit** to publicly funded democratic infrastructure including public interest media and the creation, selection and use of online platforms that afford citizen participation and deliberation;
- **Regulate** for greater transparency and accountability from platforms, including algorithmic transparency and accountability for verifying the sources of political advertising;
- **Revisit** regulation of privacy and data protection to better protect indigenous rights to data sovereignty and redress the failures of a consent-based approach to data management; and
- **Recalibrate** policies and protections to address not only individual rights and privacy but also their collective impacts on wellbeing.

Conclusion: Regulation is necessary but difficult

Regulating social media and other digital intermediaries is necessary but difficult. Constraining harmful digital communication requires co-ordinated effort by multiple actors with divergent, competing and conflicting interests, but as James Lewis from the Center for Strategic and International Studies has said:

> 2021 will be the year of regulation for the tech giants—they are a mature industry now, not shiny young start-ups. We used to say too big to fail for banks, but banks are highly regulated and these guys are moving in this direction too (quoted in Satariano, 2020).

What is required is some combination of governmental and inter-governmental regulation, industry self-regulation, industry-wide standards, multi-lateral, multi-stakeholder agreements and initiatives, technology innovation, and market pressure by advertisers, consumers and service users.

Daniela Stockmann (2020, pp. 256–259) considers the mixed results of self-regulatory initiatives and notes (pp. 258–259) that the core issue to be addressed is the business model of Big Tech itself—the collection, ownership, engineering, application and marketing of user data in the age of 'surveillance capitalism' (Zuboff, 2019, 2021). Stockmann (2020, p. 260) concludes: 'What the "right" mix between government, business, and society is in the governance of social media remains a crucial question that requires much further investigation.'

Internationally aligned anti-trust/competition regulation, tax regimes and enforcement mechanisms will be part of the solution, but as with any exercise of the state's regulatory powers, we need to be

mindful of unintended consequences and consider non-regulatory responses that may, in the end, prove more effective.

This suggests a need for integrated, cross-sectoral, strategic policy development—not reactive, piecemeal regulation that, even if it can be enforced, is unlikely to have any significant impact and may unjustifiably restrict the right to freedom of expression.

The remaining four working papers in this series elaborate on current legal frameworks for regulating harmful communication (Working paper 21/05), arguments for and against restricting freedom of expression (Working paper 21/06), the need to strike a fair balance when regulating harmful communication (Working paper 21/07), counter-speech as an alternative or complement to prohibition and censorship, and civility as everyone's responsibility (Working paper 21/08).

References

Andrews, F., & Pym, A. (2021). The websites sustaining Britain's far-right influencers. *Bellingcat,* February 24, 2021. Accessed February 26, 2021, from https://www.bellingcat.com/news/uk-and-europe/2021/02/24/the-websites-sustaining-britains-far-right-influencers/?s=09

Applebaum, A., & Pomerantsev, P. (2021). How to put out democracy's dumpster fire. *The Atlantic,* March 8, 2021. Accessed March 10, 2021, from https://www.theatlantic.com/magazine/archive/2021/04/the-internet-doesnt-have-to-be-awful/618079/

Associated Press. (2021). From Baked Alaska to a guy with horns: Notable riot arrests. Associated Press, January 17, 2021. Accessed January 18, 2021, from https://apnews.com/article/donald-trump-capitol-siege-alaska-media-riots-709fca95f4e3dce9e40c440d96dcd006

Barnes, C. (2021). Australians deserves [sic.] rebuke for Facebook shakedown. *NZ Herald,* March 1, 2021. Accessed March 2, 2021, from https://www.nzherald.co.nz/business/news/article.cfm?c_id=3&objectid=12424975

Bender, J. (2021). Zukunft der Demokratie: So schadet uns das Internet. *Frankfurter Allgemeine Zeitung,* February 2, 2021. Accessed Februry 3, 2021, from https://www.faz.net/-ikh-a83wq

Bensinger, G. (2021a). Now social media grows a conscience? *New York Times,* January 13, 2021. Accessed January 14, 2021, from https://nyti.ms/3qiGH5M

Bensinger, G. (2021b). People want real change from Facebook. Its 'Supreme Court' isn't delivering. *New York Times,* February 5, 2021. Accessed February 6, 2021, from https://nyti.ms/39Qvx2G

Berners-Lee, T. (2019). I invented the World Wide Web. Here's how we can fix it. *New York Times,* November 24, 2019. Accessed January 15, 2021, from https://nyti.ms/2s9qBSV

Binder, L., Ueberwasser, S., & Stark, E. (2020). Gendered hate speech in Swiss WhatsApp messages. In G. Giusti & G. Iannàccaro (Eds.), *Language gender and hate speech: A multidisciplinary approach* (pp. 59–74). Venice: Edizioni Ca' Foscari. https://doi.org/10.30687/978-88-6969-478-3

Bjørgo, T., & Ravndal, J. *Extreme-right violence and terrorism: Concepts, patterns, and responses*. The Hague: International Centre for Counter-Terrorism Policy Brief. Accessed December 8, 2020, from https://icct.nl/app/uploads/2019/09/Extreme-Right-Violence-and-Terrorism-Concepts-Patterns-and-Responses-4.pdf

Bleiker, C. (2021). Donald Trump supporters flock to niche social media sites. *Deutsche Welle,* January 15, 2021. Accessed January 18, 2021, from https://p.dw.com/p/3nwy9

Boland, H. (2021). Facebook 'can't regulate itself out of the mess it has made'. *The Telegraph,* January 31, 2021. Accessed February 19, 2021, from https://www.telegraph.co.uk/technology/2021/01/31/facebook-cant-regulate-mess-has-made/

Breuer, J. (2017). Hate speech in online games. In K. Kasper, L. Gräßer, & A. Riffi (Eds.), *Online hate speech: Perpektiven auf eine neue Form des Hasses* (pp. 107–112). Düsseldorf: Kopaed. Accessed November 27, 2020, from https://www.grimme-institut.de/fileadmin/Grimme_Nutzer_Dateien/Akademie/Dokumente/SR-DG-NRW_04-Online-Hate-Speech.pdf

Bromell, D. (2017). *The art and craft of policy advising: A practical guide.* Cham, CH: Springer.

Bromell, D., & Shanks, D. (2021). Censored! Developing a framework for making sound decisions fast. *Policy Quarterly*, *17*(1), 42–49. https://doi.org/10.26686/pq.v17i1.6729

Browning, K. (2021). Parler C.E.O. says he was fired. *New York Times,* February 3, 2021. Accessed February 4, 2021, from https://nyti.ms/3tqcUdK

Browning, K., & Lorenz, T. (2021). Pro-Trump mob livestreamed its rampage, and made money doing it. *New York Times,* January 8, 2021. Accessed January 11, 2021, from https://nyti.ms/3pZvdDS

Budras, C. (2021). Regulierung von Youtube & Co: Urheber sollen im Netz besser geschützt werden. *Frankfurter Allgemeine Zeitung,* February 3, 2021. Accessed February 4, 2021, from https://www.faz.net/-gqe-a885s

Camargo, C. (2020). YouTube's algorithms might radicalise people—but the real problem is we've no idea how they work. *The Conversation*, January 21, 2020. Accessed October 9, 2020, from https://theconversation.com/youtubes-algorithms-might-radicalise-people-but-the-real-problem-is-weve-no-idea-how-they-work-129955

Cave, D. (2021a). An Australia with no Google? The bitter fight behind a drastic threat. *New York Times,* January 22, 2021. Accessed January 25, 2021, from https://nyti.ms/2KGLXkq

Clegg, N. (2021). The real story of what happened with news on Facebook in Australia. Facebook, February 24, 2021. Accessed February 26, 2021, from https://about.fb.com/news/2021/02/the-real-story-of-what-happened-with-news-on-facebook-in-australia/

Dave, D. (2021b). Google is suddenly paying for news in Australia. What about everywhere else? *New York Times,* February 17, 2021. Accessed February 18, 2021, from https://www.nytimes.com/2021/02/17/business/media/australia-google-pay-for-news.html

CCDH. (n.d.). *#DeplatformIcke: How Big Tech powers and profits from David Icke's lies and hate, and why it must stop.* Center for Countering Digital Hate. Accessed March 9, 2021, from https://www.counterhate.com/deplatform-icke

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. (2017). You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human–Computer Interaction*, *1*(2), Article 31 (December 2017). Accessed February 2, 2021, from http://comp.social.gatech.edu/papers/cscw18-chand-hate.pdf

Chaslot, G. (2019). The toxic potential of YouTube's feedback loop. *Wired*, July 13, 2019. Accessed October 9, 2020, from https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/

Chen, B., & Roose, K. (2021). Are private messaging apps the next misinformation hot spot? *New York Times,* February 3, 2021. Accessed February 3, 2021, from https://www.nytimes.com/2021/02/03/technology/personaltech/telegram-signal-misinformation.html

Chen, S. (2020a). The spread of online fascism. In A. Chen (Ed.), *Shouting zeros and ones: Digital technology, ethics and policy in New Zealand* (pp. 151–223). Wellington: Bridget Williams Books.

Chen, S. (2020b). A framework for response. In A. Chen (Ed.), *Shouting zeros and ones: Digital technology, ethics and policy in New Zealand* (pp. 169–225). Wellington: Bridget Williams Books.

Coleman, S., & Ross, K. (2010). *The media and the public: 'Them' and 'us' in media discourse.* Chichester: Wiley-Blackwell.

Collins, D. (2021). Britain must hold firm against Facebook's dangerous ban on news. *The Telegraph,* February 18, 2021. Accessed February 19, 2021, from https://www.telegraph.co.uk/technology/2021/02/18/britain-must-hold-firm-against-facebooks-dangerous-ban-news/

Conger, K. (2021). Twitter, in widening crackdown, removes over 70,000 QAnon accounts. *New York Times,* January 11, 2021. Accessed January 12, 2021, from https://nyti.ms/35x3iTW

Conger, K., & Isaac, M. (2021). Twitter permanently bans Trump, capping online revolt. *New York Times,* January 8, 2021. Accessed January 11, 2021, from https://nyti.ms/2Lj0rqN

Conway, M., Khawaja, M., Lakhani, S., Reffin, J., Robertson, A., & Weir, D. (2019). Disrupting Daesh: Measuring takedown of online terrorist material and its impacts. *Studies in Conflict & Terrorism*, 42(1–2), 141–160. https://doi.org/10.1080/1057610X.2018.1513984

Daley, J. (2021). Why are free societies sinking into an anarchic pit of social media hate? *The Telegraph,* January 16, 2021. Accessed January 19, 2021, from https://www.telegraph.co.uk/news/2021/01/16/free-societies-sinking-anarchic-pit-social-media-hate/

Dambach, K. (2021). Russian dissident Alexei Navalny criticizes Trump Twitter ban. *Deutsche Welle,* January 10, 2021. Accessed January 11, 2021, from https://p.dw.com/p/3njq9

Deutsche Welle. (2021a). Google signs deal to provide payments to French publishers, January 21, 2021. Accessed January 21, 2021, from https://p.dw.com/p/3oDpR

Deutsche Welle. (2021b). Australia commits to media law despite Facebook news ban, Canada to follow, February 19, 2021. Accessed February 19, 2021, from https://p.dw.com/p/3pZq9

Dodds, L. (2021). Donald Trump's Facebook ban could be lifted by Oversight Board. *The Telegraph,* January 21, 2021. Accessed January 22, 2021, from https://www.telegraph.co.uk/technology/2021/01/21/donald-trumps-facebook-ban-could-lifted-case-referred-independent/

Easton, W. (2021). Changes to sharing and viewing news on Facebook in Australia. Facebook Australia & New Zealand, update February 22, 2021. Accessed February 23, 2021, from https://about.fb.com/news/2021/02/changes-to-sharing-and-viewing-news-on-facebook-in-australia/

Economist. (2021). Big Tech and censorship. *The Economist*, January 16, 2021. Accessed January 18, 2021, from https://www.economist.com/leaders/2021/01/16/big-tech-and-censorship

Election Integrity Partnership. (2021). *The long fuse: Misinformation and the 2020 election*. Center for an Informed Public, Digital Forensic Research Lab, Graphika, & Stanford Internet Observatory. Stanford Digital Repository: Election Integrity Partnership. v1.2.0. Accessed March 9, 2021, from https://purl.stanford.edu/tr171zs0069

Elliott, M., Berentson-Shaw, J., Kuehn, K., Salter, L., & Brownlie, E. (2019). *Digital threats to democracy: A report from The Workshop*, May 2019. Accessed October 15, 2020, from https://www.digitaldemocracy.nz/

Etzioni, A. (2019). My experience with social media restrictions on free speech. *The National Interest,* December 15, 2019. Accessed October 16, 2020, from https://nationalinterest.org/feature/my-experience-social-media-restrictions-free-speech-105127

Feuer, W. (2019). Critics slam study claiming YouTube's algorithm doesn't lead to radicalization. *CNBC*, December 30, 2019. Accessed October 9, 2020, from https://www.cnbc.com/2019/12/30/critics-slam-youtube-study-showing-no-ties-to-radicalization.html

Fielitz, M., & Schwarz, K. (2020). *Hate not found?! Deplatforming the far right and its consequences*. Institut für Demokratie und Zivilgesellschaft / Amadeu-Antonio-Stiftung. Accessed December 10, 2020, from https://www.idz-jena.de/forschung/hate-not-found-das-deplatforming-der-extremen-rechten/

Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy, 10*(1), 33–50. https://doi.org/10.1386/jdmp.10.1.33_1

Frenkel, S. (2021). Fringe Groups splinter online after Facebook and Twitter bans. *New York Times,* January 11, 2021. Accessed January 12, 2021, from https://nyti.ms/38AIDQN

Friedersdorf, C. (2018). YouTube extremism and the long tail. *The Atlantic*, March 12, 2018. Accessed October 9, 2020, from https://www.theatlantic.com/politics/archive/2018/03/youtube-extremism-and-the-long-tail/555350/

Fyers, A., Kenny, K., & Livingston, T. (2019). How YouTube spreads extremist ideas that inspire events like the Christchurch mosque shootings. *Stuff*, March 29, 2019. Accessed October 9, 2020, from https://www.stuff.co.nz/national/christchurch-shooting/111605530/how-youtube-spreads-extremist-ideas-that-inspire-events-like-the-christchurch-mosque-shootings

Gargliadoni, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech.* UNESCO Series on Internet Freedom. Paris: UNESCO. Accessed November 25, 2020, from https://unesdoc.unesco.org/ark:/48223/pf0000233231

Gilbert, J. & Elley, B. (2020). Shaved heads and Sonnenrads: Comparing white supremacist skinheads and the alt-right in New Zealand. *Kōtuitui: New Zealand Journal of Social Sciences Online, 15*(2), 280-294. https://doi.org/10.1080/1177083X.2020.1730415

Goldberg, M. (2021). The scary power of the companies that finally shut Trump up. *New York Times*, January 11, 2021. Accessed January 12, 2021, from https://nyti.ms/38ysP0Y

Griffith, E., & Lorenz, T. (2021). Clubhouse, a tiny audio chat app, breaks through. *New York Times,* February 15, 2021. Accessed February 16, 2021, from https://nyti.ms/3rVg6MH

Griffin, P. (2021). Fickle friend: How Facebook showed Australia who's boss. *New Zealand Herald,* March 11, 2021 (reprinted from *NZ Listener*). Accessed March 11, 2021, from https://www.nzherald.co.nz/business/news/article.cfm?c_id=3&objectid=12427825

Groen, M. (2017). 'Gogo let's rape them': Sexistischer Sprachgebrauch in Online Gaming Communities. In K. Kasper, L. Gräßer, & A. Riffi (Eds.), *Online hate speech: Perpektiven auf eine neue Form des Hasses* (pp. 113–119). Düsseldorf: Kopaed. Accessed November 27, 2020, from https://www.grimme-institut.de/fileadmin/Grimme_Nutzer_Dateien/Akademie/Dokumente/SR-DG-NRW_04-Online-Hate-Speech.pdf

Hale, J. (2019). There's a fatal flaw in the new study claiming YouTube's recommendation algorithm doesn't radicalize viewers. *Tubefilter*, December 30, 2019. Accessed October 9, 2020, from https://www.tubefilter.com/2019/12/30/youtube-radicalization-study-extremist-content-wormhole-rabbit-hole/

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior, 29*(3), 1159–1168. https://doi.org/10.1016/j.chb.2012.10.008

Hanfeld, M. (2021). Trumps Social-Media-Bann: Stumm geschaltet. *Frankfurter Allgemeine Zeitung,* January 11, 2021. Accessed January 12, 2021, from https://www.faz.net/-gsb-a7dyu

House Judiciary Committee. (2020). *Investigation of competition in digital markets.* Majority staff report and recommendations, Sub-committee on Antitrust, Commercial and Administrative Law, October 6, 2020. Accessed October 13, 2020, from https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf

Isaac, M., & Browning, K. (2020). Fact-checked on Facebook and Twitter, conservatives switch their apps. *New York Times,* November 11, 2020. Accessed January 5, 2021, from https://nyti.ms/36uP1a2

Isaac, M., & Cave, D. (2021). Facebook strikes deal to restore news sharing in Australia. *New York Times,* February 22, 2021, updated February 23, 2021. Accessed February 23, 2021, from https://nyti.ms/3uhVwbk

Isaac, M., & Kang, C. (2020). 'It's Hard to Prove': Why antitrust suits against Facebook face hurdles. *New York Times*, December 10, 2020. Accessed December 16, 2020, from https://nyti.ms/37PDWBh

Isaac, M., Wakabayashi, D., Cave, D., & Lee, E. (2021). Facebook blocks news in Australia, diverging with Google on proposed law. *New York Times,* February 17, 2021. Accessed February 18, 2021, from https://www.nytimes.com/2021/02/17/technology/facebook-google-australia-news.html

Itega. (2020). Privacy beat: Berners-Lee steps up to plate on data privacy with 'Inrupt'. February 21, 2020. Accessed January 15, 2021, from https://itega.org/2020/02/21/privacy-beat-berners-lee-steps-up-to-plate-on-data-privacy-with-inrupt/

Jaffer, J., & Bass, K. (2021). Facebook's 'Supreme Court' faces its first major test. *New York Times,* February 17, 2021. Accessed February 22, 2021, from https://nyti.ms/3u9bh4a

Joyce, S. (2021). Democracy will endure its trashing by Donald Trump. *NZ Herald,* January 8, 2021. Accessed January 11, 2021, from https://www.nzherald.co.nz/business/steven-joyce-democracy-will-endure-its-trashing-by-donald-trump

Kang, C., & McCabe, D. (2020). Big tech was their enemy, until partisanship fractured the battle plans. *New York Times,* October 6, 2020. Accessed October 13, 2020, from https://nyti.ms/3d3NJpn

Kaspar, K. (2017). Hassreden im Internet: Ein besonderes Phänomen computervermittelter Kommunikation? In K. Kasper, L. Gräßer, & A. Riffi (Eds.), *Online hate speech: Perpektiven auf eine neue Form des Hasses* (pp. 63–70). Düsseldorf: Kopaed. Accessed November 27, 2020, from https://www.grimme-institut.de/fileadmin/Grimme_Nutzer_Dateien/Akademie/Dokumente/SR-DG-NRW_04-Online-Hate-Speech.pdf

Kasper-Claridge, M. (2021). Social media simply pulling the plug won't work. *Deutsche Welle,* January 11, 2021. Accessed January 11, 2021, from https://p.dw.com/p/3nlYk

Keall, C. (2021a). NZ Privacy Commissioner critical of Facebook and Twitter's decision to ban Trump. *NZ Herald,* January 9, 2021. Accessed January 11, 2021, from https://www.nzherald.co.nz/business/nz-privacy-commissioner-critical-of-facebook-and-twitters-decision-to-ban-trump/E6XPAKU3DFTI6FHGLPR637456M/

Keall, C. (2021b). Google threatens to pull out of Australia, pays up in France. *NZ Herald*, January 22, 2021. Accessed January 25, 2021, from https://www.nzherald.co.nz/business/google-threatens-to-pull-out-of-australia-pays-up-in-france/BZNORE22SMOZ2CYKNUPYAZJ3X4/

Klein, R. (2021). Google vs. Australia: 5 questions and answers. *Deutsche Welle,* January 25, 2021. Accessed January 26, 2021, from https://p.dw.com/p/3oOmv

Knight, B. (2020). Neo-Nazi attack survivors create tool to track racist extremists. *Deutsche Welle,* December 16, 2020. Accessed December 16, 2020, from https://p.dw.com/p/3mnkT

Ledwich, M., & Zaitsev, A. (2019). Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *First Monday*, *25*(3). https://doi.org/10.5210/fm.v25i3.10419

Lerman, R. (2020). The conservative alternative to Twitter wants to be a place for free speech for all. It turns out, rules still apply. *Washington Post,* July 15, 2020. Accessed March 8, 2021, from https://www.washingtonpost.com/technology/2020/07/15/parler-conservative-twitter-alternative/

Lingel, J. (2021). A queer and feminist defense of being anonymous online. *Proceedings of the 54th Hawaii International Conference on System Sciences, 2021*, pp. 2534–2543. Accessed January 5, 2021, from http://hdl.handle.net/10125/70925

Lohr, S. (2021). He created the Web. Now he's out to remake the digital world. *New York Times,* January 10, 2021. Accessed January 11, 2021, from https://nyti.ms/35rp0c2

McCabe, D. (2021a). Big Tech's next big problem could come from people like 'Mr. Sweepy'. *New York Times*, February 16, 2021. Accessed February 16, 2021, from https://nyti.ms/2NpYKsp

McCabe, D. (2021b). Tech's legal shield appears likely to survive as Congress focuses on details. *New York Times,* March 9, 2021. Accessed March 9, 2021, from https://nyti.ms/3ruy7lh

McGuirk, R. (2021). Major Australian media company strikes Google news pay deal. *Associated Press*, February 15, 2021. Accessed February 15, 2021, from https://apnews.com/article/technology-australia-media-journalism-sundar-pichai-9aa2105cf512177b276bdf6ae9e045a7

Macklin, G. (2019). The Christchurch attacks: Livestream terror in the viral video age. Combating Terrorism Centre, 12(6), July 2019. Accessed December 9, 2020, from https://ctc.usma.edu/christchurch-attacks-livestream-terror-viral-video-age/

Mason, C., & Errington, K. (2019). *Anti-social media: Reducing the spread of harmful content on social media networks*. Auckland: Helen Clark Foundation. Accessed October 15, 2020, from https://helenclark.foundation/reports/anti-social-media/

Matamoros-Fernández, A., & Gray, J. (2019). Don't just blame YouTube's algorithms for 'radicalisation'. Humans also play a part. *The Conversation*, October 30, 2019. Accessed October 12, 2020, from https://theconversation.com/dont-just-blame-youtubes-algorithms-for-radicalisation-humans-also-play-a-part-125494

Matthews, P. (2021). Harmful content vs free speech online: Who should decide? *NZ Herald,* January 23, 2021. Accessed January 25, 2021, from https://www.nzherald.co.nz/business/paul-matthews-harmful-content-vs-free-speech-online-who-should-decide/5DZ7JXVZ4QA2W4SDWCV4T3USJI/

Miller-Idriss, C. (2020). *Hate in the homeland: The new global far right.* Princeton, NJ: Princeton University Press.

Mitchell, L., & Bagrow, J. (2020). Do social media algorithms erode our ability to make decisions freely? The jury is out. *The Conversation,* October 11, 2020. Accessed October 12, 2020, from https://theconversation.com/do-social-media-algorithms-erode-our-ability-to-make-decisions-freely-the-jury-is-out-140729

Munger, K., & Phillips, J. (2019). A supply and demand framework for YouTube politics. Pre-print, 2019. Accessed January 5, 2021, from https://osf.io/73jys/download

Murphy, H., Waters, R., Barker, A., & Smyth, J. (2021). Media blackout: Why Facebook pulled the plug on news in Australia. *NZ Herald*, February 28, 2021. Accessed March 2, 2021, from https://www.nzherald.co.nz/business/news/article.cfm?c_id=3&objectid=12424956

Myers, F. (2021). Like him or not, this censorship of Donald Trump has set a terrifying precedent. *The Telegraph,* January 11, 2021. Accessed January 12, 2021, from https://www.telegraph.co.uk/news/2021/01/11/like-not-censorship-donald-trump-has-set-terrifying-precedent/

Nicas, J. (2021a). Parler pitched itself as Twitter without rules. Not anymore, Apple and Google said. *New York Times*, January 9, 2021. Accessed January 11, 2021, from https://nyti.ms/35jgNXp

Nicas, J. (2021b). Parler tries to survive with help from Russian company. *New York Times,* January 19, 2021. Accessed January 20, 2021, from https://nyti.ms/2KsUjfc

Nicas, J. (2021c). Parler, a social network that attracted Trump fans, returns online. *New York Times,* February 15, 2021. Accessed February 16, 2021, from https://nyti.ms/2OHxBlt

Nicas, J., & Alba, D. (2021a). Amazon, Apple and Google cut off Parler, an app that drew Trump supporters. *New York Times,* January 9, 2021. Accessed January 11, 2021, from https://nyti.ms/3nzdD86

Nicas, J., & Alba, D. (2021b). How Parler, a chosen app of Trump fans, became a test of free speech. *New York Times,* January 10, 2021. Accessed January 11, 2021, from https://nyti.ms/39lUD7Y

Nicas, J., Isaac, M., & Frenkel, S. (2021). Millions flock to Telegram and Signal as fears grow over Big Tech. *New York Times,* January 13, 2021. Accessed January 14, 2021, from https://nyti.ms/38D4PK9

Niesche, C. (2021). Google's threat to turn Australia's search function off. *NZ Herald,* January 24, 2021. Accessed January 25, 2021, from https://www.nzherald.co.nz/business/christopher-niesche-googles-threat-to-turn-australias-search-function-off/O26BRNTCPOGFV56Y5ND3JCYUNQ/

O'Brien, M. (2021). Judge says Amazon won't have to restore Parler web service. *Associated Press*, January 22, 2021. Accessed January 25, 2021, from https://apnews.com/article/donald-trump-media-social-media-web-services-courts-bf948fa0599ade9c8036b65534a00ea1

Ovide, S. (2020). Congress agrees: Big tech is broken. *New York Times*, October 7, 2020. Accessed October 13, 2020, from https://nyti.ms/2GxKeM0

Ovide, S. (2021a). Who should make the online rules? *New York Times*, January 11, 2021. Accessed January 14, 2021, from https://nyti.ms/2LK5Fvu

Ovide, S. (2021b). Facebook invokes its 'Supreme Court'. *New York Times,* January 22, 2021. Accessed January 25, 2021, from https://nyti.ms/398zYWd

Ovide, S. (2021c). The internet is splintering. *New York Times*, February 17, 2021. Accessed February 19, 2021, from https://nyti.ms/3at3ZAE

Ovide, S. (2021d). Copying China's online blockade. *New York Times*, March 1, 2021. Accessed March 2, 2021, from https://nyti.ms/3sENkAl

Ovide, S. (2021e). Fixing what the internet broke. *New York Times*, March 4, 2021. Accessed March 9, 2021, from https://nyti.ms/3rh90Cn

Owen, T. (2019). Decoding the racist memes the alleged New Zealand shooter used to communicate. *Vice*, March 15, 2019. Accessed December 14, 2020, from https://www.vice.com/en/article/vbwn9a/decoding-the-racist-memes-the-new-zealand-shooter-used-to-communicate

Palka, P. (2021). The world of fifty (interoperable) Facebooks. *Seton Hall Law Review*, *51*(4), forthcoming. Accessed March 10, 2021, from http://dx.doi.org/10.2139/ssrn.3539792

Reed, A., Whittaker, J., Votta, F., & Looney, S. (2019). *Radical filter bubbles: Social media personalisation algorithms and extremist content.* Global Research Network on Terrorism and Technology, July 26, 2019. Accessed October 16, 2020, from https://rusi.org/publication/other-publications/radical-filter-bubbles-social-media-personalisation-algorithms-and

Ribeiro, M., Ottoni, R., West, R., Almeida, V., & Meira, W. (2019). Auditing radicalization pathways on YouTube. arXiv.org, December 4, 2019. Accessed October 9, 2020, from https://arxiv.org/abs/1908.08313

Riegert, B. (2020). EU takes on tech giants. *Deutsche Welle,* December 16, 2020. Accessed December 16, 2020, from https://p.dw.com/p/3mpHn

Rogers, R. (2020). Deplatforming: Following extreme internet celebrities to Telegram and alternative social media. *European Journal of Communication, 35*(3), 213–229. https://doi.org/10.1177/0267323120922066

Roose, K. (2019). The making of a YouTube radical. *New York Times*, June 8, 2019. Accessed October 9, 2020, from https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

Roose, K. (2020). 'Rabbit hole', a narrative audio series. *New York Times*, April–June 2020. Accessed October 12, 2020, from https://nyti.ms/3833dYC

Roose, K. (2021). Can Clubhouse move fast without breaking things? *New York Times,* February 25, 2021. Accessed February 26, 2021, from https://www.nytimes.com/2021/02/25/technology/clubhouse-audio-app-experience.html

Rösner, L., & Krämer, N. (2016). Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments. *Social Media + Society, 2*(3), 1–13. https://doi.org/10.1177/2056305116664220

Royal Commission of Inquiry. (2020). *Ko tō tātou kāinga tēnei.[This is our home.] Report: Royal Commission of Inquiry into the terrorist attack on Christchurch masjidain on 15 March 2019*. November 26, 2020. Accessed December 8, 2020, from https://christchurchattack.royalcommission.nz/

Samaratunge, S., & Hattotuwa, S. (2014). *Liking violence: A study of hate speech on Facebook in Sri Lanka.* Colombo: Centre for Policy Alternatives Sri Lanka, September 2014. Accessed December 31, 2020, from https://www.cpalanka.org/liking-violence-a-study-of-hate-speech-on-facebook-in-sri-lanka/

Satariano, A. (2020). Big fines and strict rules unveiled against 'big tech' in Europe. *New York Times*, December 15, 2020. Accessed December 16, 2020, from https://nyti.ms/3afsp0O

Satariano, A., & Stevis-Gridneff, M. (2020). Big tech turns its lobbyists loose on Europe, alarming regulators. *New York Times,* December 14, 2020. Accessed December 16, 2020, from https://nyti.ms/3833dYC

Shleina, V., Fahey, E., Klonick, K., Menéndez González, N., Murray, A., & Tzanou, M. (2020). The law of Facebook: Borders, regulation and global social media. City Law School Research Paper 2020/01. Accessed March 10, 2021, from https://openaccess.city.ac.uk/id/eprint/24375

Schmitt, J., Harles, D., & Rieger, D. (2020). Themen, Motive und Mainstreaming in rechtsexttremen Online-Memes. *Medien & Kommunikationswissenschaft, 68*(1–2), 73–93. https://doi.org/10.5771/1615-634X-2020-1-2-73

Schwartmann, R. von (2021). Der Social-Media-Kodex. *Frankfurter Allgemeine Zeitung,* January 13, 2021. Accessed January 18, 2021, from https://www.faz.net/-ikh-a7g5p

Small, Z. (2019). 'Global alliance': Jacinda Ardern joins world leaders calling for tech giant accountability. *Newshub,* March 20, 2019. Accessed October 12, 2020, from https://www.newshub.co.nz/home/politics/2019/03/global-alliance-jacinda-ardern-joins-world-leaders-calling-for-tech-giant-accountability.html

Smith, B. (2021a). We worked together on the internet. Last week, he stormed the Capitol. *New York Times,* January 10, 2021. Accessed January 11, 2021, from https://nyti.ms/39hvETn

Smith, B. (2021b). Trump wants back on Facebook. This star-studded jury might let him. *New York Times,* January 24, 2021. Accessed January 25, 2021, from https://nyti.ms/3cfJSqK

Somerville, T. (2021). Censor sensibility: The books you can't read in New Zealand, and why. *Stuff*, January 10, 2021. Accessed January 12, 2021, from https://www.stuff.co.nz/entertainment/books/300181768/censor-sensibility-the-books-you-cant-read-in-new-zealand-and-why

Steinmeier, F.-W. (2021). Federal President Frank-Walter Steinmeier on the occasion of the 11th Forum Bellevue on the future of democracy: 'Democracy and the digital public sphere—A transatlantic challenge' at Schloss Bellevue, March 1, 2021. Accessed March 1, 2021, from https://www.bundespraesident.de/SharedDocs/Downloads/DE/Reden/2021/03/210301-Forum-Bellevue-Englisch.pdf

Stockmann, D. (2020). Media or corporations? Social media governance between public and commercial rationales. In H. Anheier & T. Baums (Eds), *Advances in corporate governance: Comparative perspectives* (pp. 249–268). Oxford University Press. https://doi.org/10.1093/oso/9780198866367.003.0011

Stop Hate for Profit. (2021). Stop Hate for Profit coalition calls on Twitter and all social media platforms to #BanTrumpSaveDemocracy after Donald Trump incites violent attack on U.S. Capitol. Press release, January 8, 2021. Accessed January 11, 2021, from https://apnews.com/press-release/pr-newswire/donald-trump-business-race-and-ethnicity-war-and-unrest-products-and-services-58c13698a5e2f33bb006b4558c1a3cae

Strossen, N. (2018). *Hate: Why we should resist it with free speech, not censorship*. New York: Oxford University Press.

Swisher, K. (2021). A new-media showdown in Australia. *New York Times,* February 18, 2021. Accessed February 19, 2021, from https://www.nytimes.com/2021/02/18/opinion/Australia-Facebook-Google-news.html

Thompson, S., & Warzel, C. (2021). How Facebook incubated the insurrection. *New York Times,* January 14, 2021. Accessed January 14, 2021, from https://nyti.ms/38FfjsC

Titcomb, J. (2021). The consequences of Trump's Twitter ban go beyond the President. *The Telegraph*, January 10, 2021. Accessed January 11, 2021, from https://www.telegraph.co.uk/technology/2021/01/10/consequences-trumps-twitter-ban-go-beyond-president/

Tufekci, Z. (2018). YouTube, the great radicalizer. *New York Times*, March 10, 2018. Accessed October 9, 2020, from https://nyti.ms/2GeTMa6

Twitter Safety. (2021). An update following the riots in Washington, DC. January 12, 2021. Accessed January 13, 2021, from https://blog.twitter.com/en_us/topics/company/2021/protecting--the-conversation-following-the-riots-in-washington--.html

Volokh, E. (2021). Trump was kicked off Twitter. Who's next? *New York Times,* January 11, 2021. Accessed January 12, 2021, from https://nyti.ms/2KbOlzq

von Kempis, F. (2017). Contenance. Interview mit Lars Gräßer (Grimme-Institut). In K. Kasper, L. Gräßer, & A. Riffi (Eds.), *Online hate speech: Perpektiven auf eine neue Form des Hasses* (pp. 121–124). Düsseldorf: Kopaed. Accessed November 27, 2020, from https://www.grimme-institut.de/fileadmin/Grimme_Nutzer_Dateien/Akademie/Dokumente/SR-DG-NRW_04-Online-Hate-Speech.pdf

Wade, A. (2021). 'Revenge porn' law change to criminalise posting intimate recordings without consent backed by all parties. *NZ Herald,* March 9, 2021. Accessed March 9, 2021, from https://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&objectid=12427137

Walls, J. (2021). Twitter crackdown reaches NZ: Hundreds of NZ right-wing users kicked off Twitter. *NZ Herald*, January 13, 2021. Accessed January 13, 2021, from https://www.nzherald.co.nz/nz/twitter-crackdown-reaches-nz-hundreds-of-nz-right-wing-users-kicked-off-twitter/2TFJVW6NT4WT3VM5KDYYRGNIBY/

Warren, E. (2019). It's time to break up Amazon, Google, and Facebook. *Team Warren*, March 8, 2019. Accessed October 13, 2020, from https://medium.com/@teamwarren/heres-how-we-can-break-up-big-tech-9ad9e0da324c

Waters, R., Murphy, H., & Barker, A. (2021). Big Tech versus journalism: Publishers watch Australia fight with bated breath. *NZ Herald,* February 19, 2021. Accessed February 19, 2021, from https://www.nzherald.co.nz/business/big-tech-versus-journalism-publishers-watch-australia-fight-with-bated-breath/FPABRFWHZW5TDHKKI4DPJCBYHI/

Winseck, D. (2019). Ready, fire, aim: Why digital platforms are not media companies and what to do about them. Presentation to the *2019 IAMCR Conference, Communication, Technology and Human Dignity: Disputed Rights, Contested Truths*, July 7–11, 2019, Madrid. Accessed October 12, 2020, from https://core.ac.uk/download/pdf/223233053.pdf

Woodhouse, S., & Brody, B. (2020). Amazon sets new lobbying record as tech antitrust scrutiny grows. *Bloomberg,* July 21, 2020. Accessed December 16, 2020, from https://www.bloomberg.com/news/articles/2020-07-21/amazon-sets-new-lobbying-record-as-tech-antitrust-scrutiny-grows

Young, A. (2021). Kris Faafoi hopeful meaningful talk will occur between Facebook, Google and local media. *NZ Herald*, March 3, 2021. Accessed March 4, 2021, from https://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&objectid=12425805

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for the future at the new frontier of power.* London: Profile Books.

Zuboff, S. (2021). The coup we are not talking about. *New York Times,* January 29, 2021. Accessed February 1, 2021, from https://nyti.ms/3iZZdx1