After Christchurch:
Hate, harm and the limits of censorship

7. Counter-speech and civility as everyone's responsibility

David Bromell

Working Paper 21/08

VICTORIA UNIVERSITY OF WELLINGTON
TE HERENGA WAKA

Institute for Governance and Policy Studies
A research institute of the School of Government

CENTER FOR ADVANCED INTERNET STUDIES

INSTITUTE FOR GOVERNANCE AND
POLICY STUDIES
WORKING PAPER
21/08

MONTH/YEAR

April 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).

DISCLAIMER

The views, opinions, findings, and conclusions or recommendations expressed in this paper are strictly those of the author. They do not necessarily reflect the views of the Institute for Governance and Policy Studies, the School of Government, Victoria University of Wellington, or the Center for Advanced Internet Studies (CAIS).

This is the final paper 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

From October 2020 to March 2021, Dr David Bromell was a research Fellow at the Center for Advanced Internet Studies (CAIS) in Bochum, North Rhine-Westphalia, Germany, which 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

Counter-speech and civility as everyone's responsibility

Abstract

This is the final working paper in the series, *After Christchurch: Hate, harm and the limits of censorship*.

Passing laws does not solve complex social problems, and as discussed in Working paper 21/07, **Striking a fair balance when regulating harmful communication**, the state can deploy its expressive and not only its coercive powers, and encourage and support counter-speech as an alternative and/or complement to prohibition and censorship.

This paper summarises the preceding six papers, then discusses counter-speech and online civic interventions, introduces some counter-speech strategies, and identifies opportunities for government investment. In the private sector, advertisers, shareholders and customers have roles to play, as does the Fourth Estate (the press and news media in their role as watchdogs of a free, open and democratic society). Counter-speech and civility are everyone's responsibility. In the World Wide Web, the party's over, and we all need to help clean up.

Tags: #ChristchurchCall #censorship #hatespeech #freespeech #freedomofexpression

Introduction: Hate, harm and the limits of censorship

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

Working paper 21/02 provided background and context—the terrorist attack on Christchurch mosques on March 15, 2019, and the Christchurch Call summit in Paris on May 15. The paper concluded that while the Christchurch Call is aspirational and well-intentioned, the goal of eliminating terrorist and violent extremist content online is impossible—unless governments employ draconian measures of censorship that would severely restrict freedom of expression and commitment to a free, open and secure internet. Impossibility is not, however, an excuse for policy makers to do nothing.

Working paper 21/03 summarised survey findings in New Zealand, Australia, Europe and Germany on the extent of 'hate crimes' and exposure to 'hate speech'. A difficulty is that these surveys use broad and subjective definitions of 'hate speech' that, if carried over into legislation, would undermine freedom of expression. The paper offered definitions of 'hate crime' and 'hate speech' informed by international law, scholarly debate and existing regulation in the United Kingdom, Denmark, Canada, Germany and New Zealand. It argued that regulation should provide protection not from the *emotions* of 'hate' or offence, but from the *effect* of harm. For this reason, it is preferable to refer to 'harmful communication' rather than 'hate speech' when considering regulatory and non-regulatory options to address it.

Working paper 21/04 argued that social media and other digital intermediaries are too big and have too much influence not to be subject to government regulation, but harmful digital communication is exceptionally difficult to regulate for reasons that relate both to the nature of the internet and digital communications, and to the business models and algorithms used by Big Tech companies. The paper reviewed recent research on whether de-platforming is effective and discussed the de-platforming of then-President Donald Trump by social media companies in January 2021. It concluded that while constraining harmful digital communication requires co-ordinated effort by multiple actors, 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.

Working paper 21/05 summarised provisions in international human rights standards, and in New Zealand law and in a number of other jurisdictions, that protect and restrict the right to freedom of expression. It noted relevant recommendations of the Royal Commission of Inquiry into the terrorist attack on Christchurch mosques. A free, open and democratic society protects everyone's right to freedom of opinion and expression but may justifiably qualify this freedom to prevent harm to others, if it does so in ways that conform to strict tests of legality, proportionality and necessity. There is an established consensus in international human rights standards that it may be justifiable to restrict public communication that incites discrimination, hostility or violence against a social group with a common 'protected characteristic' such as nationality, race or religion. Regulation to protect social groups from criticism, offence or lack of respect is not, however, a justifiable restriction of freedom of expression.

Working paper 21/06 presented arguments both for and against the state restricting the right to freedom of expression. It examined arguments for restricting freedom of expression in relation to the respective interests of 'protagonists', 'antagonists' and 'the audience'. The political and social objective is to balance rights and responsibilities in ways that create and maintain a civil, well-ordered society. Arguments against restricting freedom of expression involve considerations of individual autonomy, human agency and legal responsibility, reason and 'the marketplace of ideas', political legitimacy and representative democracy, restraining the state, a dilemma regulation can create for human rights law, and legal efficacy.

Working paper 21/07 discussed some further considerations to ensure that the state strikes a fair balance when regulating harmful communication. Important distinctions are between public and private communication, harm and offence, and persons and groups. Any proposal to regulate private communication, or to protect social groups and their beliefs, values and practices from criticism and offence, exacts too high a price because of the extent to which this restricts the right to freedom of opinion and expression. Policy makers need to refrain from using the coercive power of the state to enforce a 'heckler's veto'—or a 'mourner's veto.' Passing laws does not in any case solve complex social problems, and in a free, open and democratic society, agonistic respect and toleration are preferable to 'calling out', 'cancel culture' and de-platforming. This does not necessarily imply state neutrality. Governments can deploy expressive rather than coercive powers to address 'lawful hate speech', upholding the ideal of free and equal citizenship and learning from what does and does not work to maintain public order and civility.

This final working paper (21/08) discusses counter-speech and online civic interventions, reviews what works and what does not, introduces some counter-speech strategies and identifies opportunities for government investment. In the private sector, advertisers, shareholders and customers have roles to play, as does the Fourth Estate (the press and news media in their role as watchdogs of a free, open and democratic society). Counter-speech and civility are everyone's responsibility.

‘Hate speech’ and counter-speech

The European Commission against Racism and Intolerance (ECRI), in its lengthy preamble to General Policy Recommendation No. 15 on combatting ‘hate speech’ (ECRI, 2016), recommended criminal prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence and meets the six-point threshold test in the Rabat Plan of Action.[1]

ECRI also acknowledged, however, that ‘criminal prohibitions are not in themselves sufficient to eradicate the use of hate speech and are not always appropriate’, and that ‘an important means of tackling hate speech is through confronting and condemning it directly by counter-speech that clearly shows its destructive and unacceptable character’ (ECRI, 2016, p. 4).

Theodore Shaw (in Molnar, 2012, pp. 406–407) says:

> I think that the deal ought to be that if we allow hate speech, if we tolerate hate speech, then the social compact ought to be that when people hear hate speech that they respond to it and condemn it; no matter to whom it’s directed, they have to respond to it. That’s the price I think we ought to pay for tolerating hate speech. If we are not going to condemn it, whenever we see it or hear it, then we need to rethink whether we should tolerate it, because of the impact it has.

Jeffrey Howard (2018, p. 21) acknowledges that criminalisation of ‘hate speech’ is often an ineffective strategy. He asks how we might combat dangerous ideas effectively through counter-speech, both in the streets and online, and argues that the effectiveness of counter-speech strategies and initiatives also needs to be evaluated and assessed. In subsequent articles (2019a, 2019b) he discusses this further, concluding that:

> What matters is the overall assessment of counterspeech’s prospects for success, compared with that of coercion. To insist on counter-speech over coercion even in cases when coercion has a far greater prospect of success unacceptably endangers the prospective victims of the incited harm (Howard, 2019a, p. 252).[2]

Online civic intervention

Social media and other digital intermediaries have critical roles to play, given how quickly and how far harmful communication can spread online (Working paper 21/04, **Challenges in regulating online content**). Online platforms are responding to harmful user-generated content not only by employing professional content moderators and applying artificial intelligence, but also by relying on users to intervene when they are exposed to abusive language.

Facebook’s community guidelines, for example, explicitly address ‘hate speech’ as ‘objectionable content’ (Facebook, 2020). Because people post disagreeable or disturbing content that does not technically breach the community guidelines, however, Facebook considers counter-speech and the

[1] The Rabat Plan of Action is discussed in Working paper 21/05, **Regulating harmful communication: Current legal frameworks**, pp. 10–11.

[2] For a 2019 literature review on counter-speech, see Buerger & Wright, 2019. See further the section below on What works, and what does not?

tools its platform provides to help promote it as potentially a more effective way to tackle the problem.[3]

UN Special Rapporteur for minority rights, Rita Izsák, noted in 2015 that:

> Civil society organizations have developed innovative approaches and actions to combat hate in the media, including the Internet. Those actions include identifying hate trends, tracking and monitoring hate speech websites, notifying potentially affected or targeted communities about hate activities, working closely with Internet providers and governmental agencies to report hateful content and providing online educational materials and training programmes (UN General Assembly, 2015, para. 94).

Examples she cites include:

- Studio Ijambo, an initiative of the organisation Search for Common Ground to address inter-ethnic violence in Burundi;
- The Umati project in Kenya;
- The Panzagar Movement ('flower speech campaign') in Myanmar;
- Norikoe Net in Japan; and
- The International Network Against CyberHate (INACH).

Kunst, Porten-Cheé, Emmer & Eilders (2021, p. 2) define 'online civic intervention' (OCI) as 'action taken by ordinary users to fight disruptive online behavior with the aim of restoring civil and rational public discourse.' OCI can take various forms. Low-threshold OCI uses media platform tools, including reporting harmful communication to online platform providers and rating user comments using social buttons (for example, thumbs down). High-threshold OCI engages directly in polite counter-speech, urging those who post 'hate speech' or disparaging messages to be respectful toward others. An example of high-threshold OCI is a hashtag activist group known as *#jagärhär* in Sweden and *#ichbinhier* in Germany [*#Iamhere*] (Kunst, Porten-Cheé, Emmer & Eilders, 2021, p. 2; Ley, 2018).

Another tactic employed by a group of mostly young women and girls is Instagram 'tag cleaning'. As volunteer moderators, they drown out gore and harassment by flooding a user's tagged photos with pleasant images, which pushes graphic material and disrespectful memes to the bottom of an account until formal moderation processes can remove them (Farokhmanesh, 2019).

[3] Facebook's COO Sheryl Sandberg said in a January 2016 speech at the World Economic Forum in Davos, Switzerland: 'The best antidote to bad speech is good speech. The best antidote to hate is tolerance. Amplifying … counter-speech to the speech that's perpetrating hate is, we think, by far the best answer' (cited in Wagner, 2016). This needs to be understood, however, in the context of Facebook's historical tendency to invest in content regulation only when required to do so. As noted in Working paper 21/05, **Regulating harmful communication: Current legal frameworks** (p. 24), an impact of Germany's 2017 Network Enforcement Act was that already, by January 2018, an estimated 16 per cent of Facebook's content moderators worldwide were located in Germany, which accounts for only 1.5 per cent of global Facebook users (Turner, 2018).

Three examples of organisations leading online civic intervention to reduce digital harm are Moonshot, the Center for Countering Digital Hate, and the Institute for Strategic Dialogue.

- Moonshot CVE (http://moonshotcve.com/) works globally to disrupt violent extremism (Ellis, 2021). They connect vulnerable individuals with counsellors, positive content, opportunities, mentors and jobs, and publish articles, infographics, videos and reports.
- The Center for Countering Digital Hate (https://www.counterhate.com/) is an international NGO with offices in London and Washington DC that seeks to disrupt the architecture of online hate and misinformation.
- The Institute for Strategic Dialogue (ISD) (https://www.isdglobal.org/) combines research with an advanced digital analysis capability to track, measure and analyse extremism online, provides policy and advisory support and training to governments and cities, develops education resources to address online extremism, and collaborates with networks of community influencers, city and government officials and the tech sector. ISD has partnered with Google to deliver a £1m Innovation Fund across the UK to stimulate creative responses from organisations that have delivered impact in areas such as technology, sports and the arts. And its campaign training, delivered through initiatives such as YouthCAN, have measurably shifted attitudes away from polarisation on issues such as racism, Islamophobia, anti-Semitism and migration.

Other examples include Hope not Hate (https://www.hopenothate.org.uk/), the web video campaign Begriffswelten Islam (Bundeszentrale für politische Bildung, 2015), the #NotInMyName initiative of the Active Change Foundation in November 2015 after the Paris terrorist attacks, the online campaign The Redirect Method from Alphabet's Jigsaw (Jigsaw, n.d.), and the Network Code / Netzkodex initiative in NRW, Germany (Schmid & Appelhoff, 2017).

Organisations like the USA-based Anti-Defamation League, Women, Action and the Media (WAM!), the Australian-based Online Hate Prevention Institute, the Canada-based Sentinel Groups for Genocide Prevention, and the British-based Tell MAMA (Measuring Anti-Muslim Attacks) combat hate speech online by targeting digital intermediaries, lobbying them to take greater responsibility in moderating content and raising awareness among users, and by collecting complaints from users about specific types of content (Gargliadoni, Gal, Alves, & Martinez, 2015, pp. 40–41).

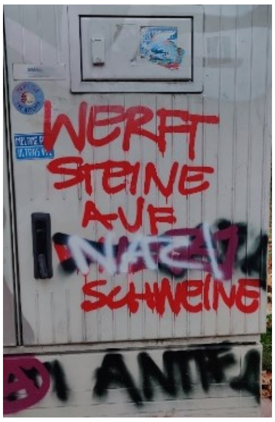
In New Zealand, Community Research announced on its Facebook page that it is producing a video on how to create memes to disrupt online racism (Community Research, 2020). And in Germany, a number of initiatives address the risk of radicalisation through online gaming, including:

- Hier spielt Vielfalt #TeamDiversity, and the Joint Declaration of the German Games Industry for More Diversity (German Games Industry Association, n.d.);
- The Keinen Pixel den Faschisten (2020) initiative of websites, media professionals, research collectives and developer studios from computer game culture who want to work for an inclusive climate in their communities through anti-fascist work;
- The 'Good Gaming—Well Played Democracy' project of the Amadeu-Antonio-Stiftung (n.d.); and
- 'Game over hate' (n.d.) for inclusive gaming communities.[4]

[4] As noted in Working paper 21/04, **Challenges in regulating online content**, p. 8, 'hate speech' is 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).

Through its #DataforGood programme, Twitter has partnered with the University of Otago's National Centre for Peace and Conflict Studies (NCPACS) in New Zealand to use Twitter data to study the ways online conversations can be used to promote tolerance and inclusion instead of division and exclusion (Hinesley, 2020). The initiative followed the Christchurch terror attacks on March 15, 2019 and other examples of 'digitally amplified polarisation' around the world. Lead researcher Sanjana Hattotuwa analysed data generated from tens of thousands of public tweets in the first 48 hours after the mosque attacks and discovered a local and global outpouring of support for victims, solidarity with the citizens of New Zealand, the affirmation of democratic ideals, pushback against terrorism and unequivocal condemnation of the perpetrator. The research partnership will support Twitter's focus on how it can tackle extremism while promoting dialogue and inter-community understanding.

On my walk home



On my walk into the city centre from CAIS, I used to pass a telecommunications junction box with red, spray-painted graffiti: 'Werft Steine auf *** Schweine' ('Throw stones at *** pigs').

I don't know who the original intended target was, because others, including individuals who appear to associate with Antifa, have blacked out the identifying word, then others(?) have written 'Nazi' over the black paint in purple, then over the purple again in white, so the graffiti now reads 'Werft Steine auf Nazi Schweine'. This is an example of counter-speech, but of the sort that heaps hatred on hatred, incitement to violence on incitement to violence.

The 'blacking out' can, however, be an important first step in counter-speech—identifying what, under the rule of law, is prohibited as harmful communication in public space, outside of which all opinions are allowed (Raue, 2020). This forbidden core needs to be kept to the absolute minimum—public communication that incites discrimination, hostility or violence against people based on their actual or supposed belonging to a social group with a common 'protected characteristic' such as nationality, race or religion.

What works, and what does not?

In 2015, Facebook asked Demos to prepare a series of research reports on the extent to which different types of counter-speech are produced and shared on Facebook (Bartlett & Krasadomski-Jones, 2015, 2016). The 2015 study found that counter-speech posts that were funny or satirical attracted the most interactions, and that 'constructive' counter-speech generated more 'likes' than 'non-constructive' counter-speech (sending abuse or aggressive threats). Photos and videos are the most effective types of content to post to reach a broader audience.

The 2016 Demos report stated:

> At Demos we believe it is important that the principle of internet freedom should be maintained; and that the internet should be a place where people feel they can speak their minds openly and freely. We therefore believe that debate, disagreement and challenge is nearly always preferable to censorship and removal of content, including when dealing with extreme or radical content, whatever its origin. However, we also believe that this can and

> should be put on an empirical basis to help us better understand the phenomenon and how to respond (Bartlett & Krasadomski-Jones, 2016, p. 5).

Demos particularly examined how speech that challenges extreme Islamist narratives is produced and shared in France, India, Indonesia, Morocco, Tunisia and the UK. Their overall conclusions were that:

> - To be effective, counter-speech must be reactive and responsive to news and current affairs;
> - Counter-speech content operates differently as the context in which it is produced and shared changes;
> - The volume of counter-speech also varies according to event-specific contexts; and
> - There is no all-encompassing approach that covers the whole of Facebook, but rather a series of country-specific approaches for which Facebook can provide an important platform to spread messages confronting IS narratives and ideology (Bartlett & Krasadomski-Jones, 2016, pp. 35–38).

A 2016 report on a field study of counter-speech on Twitter (Benesch, Ruths, Dillon, Saleem & Wright, 2016a) identified four 'vectors' of speech and counter-speech (one-to-one exchanges, many-to-one, one-to-many, and many-to-many), and eight counter-speech strategies:
1) presentation of facts to correct misstatements or misperceptions; 2) pointing out hypocrisy or contradictions; 3) warning of possible offline and online consequences of speech; 4) identification with the original speaker or target group; 5) denouncing speech as hateful or dangerous; 6) use of visual media; 7) use of humour; and 8) use of a particular tone, for example an empathetic tone (p. 18).

A related study (Benesch, Ruths, Dillon, Saleem & Wright, 2016b) identified two measures of effective counter-speech:

- A positive impact on the original ('hateful') Twitter user, shifting his or her discourse if not also his or her beliefs—as indicated by an apology or recanting, or the deletion of the original tweet or account; and
- A positive impact on the discourse norms of the 'audience' of a counter-speech conversation: all of the other Twitter users or 'cyber-bystanders' who read one or more of the relevant exchange of tweets.

Counter-speech strategies that appeared to have a positive impact on 'hateful' Twitter users involved warning of consequences, shaming and labelling, empathy and affiliation, and humour and images.[5]

Counter-speech that sets out to expose, correct or ridicule risks automatic rejection (Hemmingsen & Castro, 2017). Other strategies that appear to be ineffective are adopting a hostile or aggressive tone and insulting the user, fact-checking,[6] harassment and silencing. Aggressive wording tends to provoke more aggressive wording by other commenters, and this effect is even stronger when platforms enable anonymity (Halpern & Gibbs, 2013; Rösner & Krämer, 2016).

[5] For an example of counter-speech that effectively uses visual media and humour, see German public broadcaster ZDF's Neo Magazine Royale video clip with Jan Böhmermann, *Be Deutsch! [Achtung! Germans on the rise!]* (ZDF, 2016).

[6] Kelly (2021) cautions: 'Fact-checking can be a powerful tool in the fight against online falsehood. But it can be used as a means of censorship if not only facts but also opinions and narratives are checked.'

Schmitt, Rieger, Ernst and Roth (2018, p. 3) note that counter-speech effectiveness seems to increase with greater narrativity—telling a story (Morten, Frischlich, Rieger, & Bente, 2017), and when counter-messages are created by former extremists or 'exiters' (Frischlich, Rieger, Morten, & Bente, 2017). Schmitt, Caspari, Wolf, Bloch and Rieger (forthcoming) found that 'one-sided' narratives that offered alternative perspectives were less effective in changing attitudes than 'two-sided' narratives that engaged with opposing arguments, because they reduce the risk of people reacting against and resisting the counter-message and appear to increase the extent to which people are willing to listen to or read and evaluate other positions. Engaging well demands calmness, composure and civility (von Kempis, 2017, p. 122), declining to fuel others' anger or to 'feed the algorithm' (Metz, 2021).

A challenge, however, is that counter-speech messages are placed in the same online environment as harmful communication and are often tagged with similar keywords, so can be linked by social bots and automated algorithms directly to extremist content in personalised 'recommendations' (Schmitt, Rieger, Rutkowski, & Ernst, 2018; Frischlich, Boberg, & Quandt, 2017). This is a risk particularly for younger users, who may not be aware of algorithms and how they work, so it reinforces the importance of investment in digital literacy education (Wagner, 2020). Of course, social bots and algorithms can also spread tolerance and inclusion, and not only hatred (Frischlich, Boberg, & Quandt, 2017, p. 74). A key issue is, therefore, algorithmic transparency.

Five counter-speech strategies

Five counter-speech strategies for 'good netizens' are: focus on persons, not groups; choose when to keep silent, and when to speak up; re-frame; re-claim; and engage in calling in, not calling out.

Focus on persons, not groups

Carolin Emcke (2019, p. xii) reflects:

> Hate is fuzzy. It is difficult to hate with precision. Precision would bring delicate nuance, attentive looking and listening; precision would bring that discernment that perceives individual persons, with all their diverse, contradictory qualities and propensities, as human beings.

From within our socially constructed categories of 'us' and 'them', it can be easy to hate a category or stereotype.[7] It is harder to hate an individual person in all their multi-faceted particularity.

Writing about basic human equality, Jeremy Waldron (2017, pp.154–160) uses the term 'scintillation' to refer to the way we move our attention back and forth between some relevant 'range property' and the particularity of its manifestation in an individual person over a complete life. Waldron suggests, for example, that 'scintillation' enables a society to behave towards a terrorist both as a fellow human being who is entitled to protection from cruel or inhuman treatment and a fair trial, and as someone who, because of what she or he is alleged to have done, ought to be apprehended, tried according to the law and, if found guilty, punished (cf. Vlastos, 1984, p. 55).

[7] On pluralism and civility, and our human tendency to cluster into tribes of 'us' and 'them', see Bromell, 2019, Chap. 2.

Counter-speech can be effective when it initiates a 'scintillation' between the social stereotype of a disliked 'other' (e.g. gays, Indians, skinheads) and knowledge and experience of individual lives: 'our teacher (who's gay)', 'my doctor (who's Indian)', 'the boys down the street (who are skinheads)'.

It becomes harder (if still not impossible) to hate a 'type' when you have got to know an individual and built some shared experiences together. This is why, in the history of gay liberation, 'coming out' to family, friends and colleagues was a potent driver of social change.[8]

Focusing on persons rather than groups aligns with Waldron's insistence that regulation of harmful communication is to protect the dignity of the *individual* 'when defamatory imputations are associated with shared characteristics such as race, ethnicity, religion, gender, sexuality, and national origin', not the dignity of *a culture, religion or other social group* (Waldron, 2012, p. 60).

Choose when to keep silent, and when to speak up

When confronted as a protagonist by harmful communication, one strategy is to keep silent or 'walk off the stage', declining to react or escalate conflict, especially when one's emotions are highly aroused or there is a risk of violence. Choosing for ourselves the time and place to engage with antagonists affirms our own dignity and agency.

Nadine Strossen (2018, p. 161) comments:

> Paradoxically, in some circumstances the most effective form of counterspeech can be silence. By deliberately choosing to ignore provocative, hateful speakers, silence can powerfully convey implicit messages of disdain, while at the same time denying hateful speakers the attention they seek and often get from sparking controversy. Those engaged in counterspeech should be careful not to act in ways that are ultimately counterproductive, including efforts to silence hateful speakers through aggressive counter-demonstrations. Although such tactics might seem morally justified, they almost always backfire.

When to keep silent and when to speak up is also a choice for the audience. If protagonists need to develop 'thicker skins', members of the audience need to develop 'thinner skins' (Strossen, 2018, pp. 171–172) as 'good citizens' (Kunst, Porten-Cheé, Emmer & Eilders, 2021; von Kempis, 2017). The Royal Commission of Inquiry (2020, pp. 16–17, 99, 388–90) noted in its report the lack of a public-facing counter-terrorism strategy in New Zealand of the 'see something, say something' sort.

Across Austria and Germany, *Omas gegen Rechts*—Grannies Against the Right (Opas/grandpas join in too)—organise, attend vigils when neo-Nazis organise marches, speak out against anti-Semitism, racism, misogyny and fascism, and advocate for a free and open democratic society organised under the rule of law (Omas gegen Rechts, n.d.). Anna Ohnweiler, who introduced the initiative to Germany in January 2018, lives by her motto: 'Those who fall asleep in a democracy will wake in a dictatorship' (Hensen, 2020).

[8] Cf. Gordon Allport's (1954) intergroup contact theory developed following World War II, with its hypothesis that under certain conditions interpersonal contact can effectively reduce prejudice between majority and minority group members. Enabling conditions are, however, highly contextual and include equal status, common (or subordinate) goals, inter-group co-operation and the support of authorities, law or customs. For a meta-analysis of intergroup contact theory, see Pettigrew & Tropp, 2006. See also Paluck, Green & Green, 2018, who evaluate the state of contact hypothesis research from a policy perspective.

Paternalistic rescuing and protectionism that take over and control the discourse do not, however, empower members of minority social groups. While at one level it is unfair to expect the targeted protagonists to bear responsibility for counter-speech, and while they have no moral duty to do so, and while 'the audience' does have a moral duty to stand beside and speak up for them (Howard, 2019b), things go better in the long run when protagonists speak for themselves—and the audience *listens*.

In his 2016 commencement address to graduating students at Howard University, President Barack Obama said:

> There will be times when you shouldn't compromise your core values, your integrity, and you will have the responsibility to speak up in the face of injustice. But listen. Engage. If the other side has a point, learn from them. If they're wrong, rebut them. Teach them. Beat them on the battlefield of ideas. And you might as well start practicing now, because one thing I can guarantee you—you will have to deal with ignorance, hatred, racism, foolishness, trifling folks. I promise you, you will have to deal with all that at every stage of your life. That may not seem fair, but life has never been completely fair. Nobody promised you a crystal stair. And if you want to make life fair, then you've got to start with the world as it is (Obama, 2016).

Re-frame

In the early 1990s, while working as an ordained minister who happens to be gay, I received a number of speaking invitations, some of which I felt were framed, however kindly, as requests to address the 'problem' of homosexuality and Christian faith. I choose to re-frame the 'problem' as one of patriarchy and homophobia, so I gave talks on, for example, 'the dis-ease of heterosexuality'.

Re-framing was a common counter-speech tactic in gay liberation. If asked, for example, 'When did you decide to be gay?' a suggested response was, 'How old were you when you decided to be straight?'

Re-framing is unlikely to prove effective with emotionally invested antagonists, but it can be effective with an undecided audience in creating an environment that is less receptive to harmful communication that stirs up discrimination, hostility and violence.

Strossen (2018, pp. 162–163) discusses an example of 'humorous subversion' in 2014, when an annual neo-Nazi march in Wunsiedel in Bavaria was re-framed by *Rechts gegen Rechts* (Right against Right):

> Without the marchers' knowledge, local residents and businesses sponsored the 250 participants of the march on 15 November in what was dubbed Germany's 'most involuntary walkathon'. For every metre they walked, €10 went to a programme called EXIT Deutschland, which helps people escape extremist groups. Campaigners hung humorous posters to make the march look more like a sporting event, with slogans such as 'If only the Führer knew!' and 'Mein Mampf' (my munch) next to a table laden with bananas. They even hung a sign at the end, thanking the marchers for their 'donations' (Cresci, 2014).

The approach spread to several other German towns and one in Sweden, where it was billed as *Nazis Against Nazis* (Velasquez-Manoff, 2017). Velasquez-Manoff commented:

> Humor is a particularly powerful tool—to avoid escalation, to highlight the absurdity of absurd positions and to deflate the puffery that, to the weak-minded at any rate, might resemble heroic purpose.

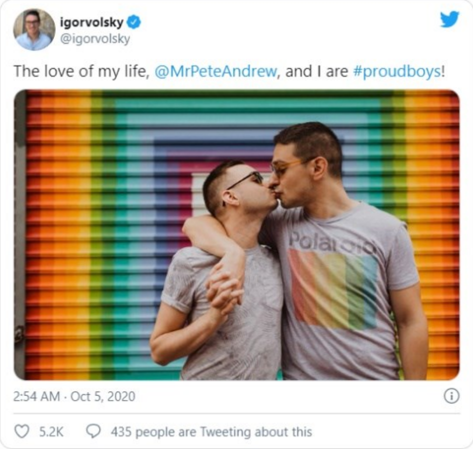
Reclaim

Reclaiming terms of abuse is another counter-speech strategy. The word 'queer', for example, was reclaimed from a derogatory term used from the end of the nineteenth century through to the 1990s, when activist groups like Queer Nation started to use it to self-identify.

A delightfully subversive act of reclaiming occurred in October 2020 in relation to Proud Boys, an alt-right, anti-immigrant, male-only political organisation that promotes and engages in political violence in the United States and Canada.[9] During the first US presidential debate on September 29, 2020, former President Donald Trump notoriously said:

> Proud Boys—stand back and stand by. But I'll tell you what ... somebody's got to do something about antifa [anti-fascist activists] and the left because this is not a right-wing problem (BBC News, 2020).

In a campaign to drown out content from Proud Boys, gay men took to using the #ProudBoys hashtag on Twitter, Reddit and TikTok to post images of queer love to 'reclaim our pride' from right-wing extremists (Rosenblatt, 2020).



igorvolsky
@igorvolsky

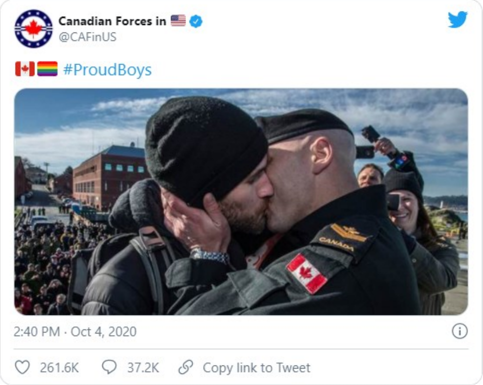
The love of my life, @MrPeteAndrew, and I are #proudboys!

2:54 AM · Oct 5, 2020

♡ 5.2K 435 people are Tweeting about this

Figure 1: Twitter, @igorvolsky, October 5, 2020

The official Twitter account of the Canadian Armed Forces in the United States joined the fun, sharing an image of a serviceman kissing his partner, captioned with emojis of the Canada flag, the rainbow pride flag and the hashtag #ProudBoys.



Canadian Forces in 🇺🇸
@CAFinUS

🇨🇦🏳️‍🌈 #ProudBoys

2:40 PM · Oct 4, 2020

♡ 261.6K 37.2K Copy link to Tweet

Figure 2: Twitter, @CAFinUS, October 4, 2020

[9] Proud Boys played a key role in the insurrection at the US Capitol on January 6, 2021. Members of Proud Boys have subsequently derided Trump as 'extraordinarily weak' and 'a total failure' after he disavowed the storming of the Capitol on January 6 and did nothing to help those who face legal action for their role in the riot (Frenkel & Feuer, 2021). Canada designated Proud Boys as a terrorist entity in February 2021 (Gillies, 2021).

As a tactic, this act of reclaiming was smart, because it involved humour, joy, love and inclusive pride as counter-speech to discrimination, hostility and violence.

Carolin Emcke (2019, p. 128) writes:

> To me, civil resistance against hatred also includes taking back the spaces of imagination. The dissenting strategies against resentment and contempt also include—and this may seem surprising after everything said up to now—the stories of happiness. In view of all the different instruments and structures of power that marginalize and disenfranchise people, resistance against hatred and contempt must include taking back the various ways of living a happy and truly free life.

Engage in calling in, not calling out

At Smith College, visiting professor Loretta Ross is combatting 'cancel culture' (or 'calling out') with an alternative strategy of 'calling in':

> Calling in is like calling out, but done privately and with respect. That may mean simply sending someone a private message, or even ringing them on the telephone(!) to discuss the matter, or simply taking a breath before commenting, screen-shotting or demanding one 'do better' without explaining how. Calling out assumes the worst. Calling in involves conversation, compassion and context. It doesn't mean a person should ignore harm, slight or damage, but nor should she, he or they exaggerate it (Bennett, 2020).[10]

Cancel culture, according to Ross, includes the following characteristics: presumption of guilt (without facts or nuance getting in the way), essentialism (when criticism of bad behaviour becomes criticism of a bad person), pseudo-intellectualism (proclaiming one's own moral high ground), unforgiveability (no apology is good enough), and contamination (guilt by association). Cancel culture is a tyranny of 'woke', groupthink and 'virtue signalling' (moral grandstanding) (Booker, 2020; Grubbs, 2020). Michael Hanfeld (2021) has noted how this is playing out among journalists and the threat it represents to freedom of the press.

'Calling in' instead of 'calling out' means toning down the anger (Waldron, 2018; Nussbaum, 2016) and the nastiness (Quinn, 2020), declining to engage in 'doom scrolling',[11] consistently demonstrating respect and maintaining open doors, rather than self-righteously and punitively de-platforming those with whom we disagree.[12] Ross says:

> Some people you can work with and some people you can work around. But the thing that I want to emphasize is that the calling-in practice means you always keep a seat at the table for them if they come back (Bennett, 2020).

'Calling in' means resisting the tribalism of playing to an audience of your own 'sort' and posting sectarian and nasty comments online that you would not speak directly to someone you were arguing with face to face (Hehir, 2020). More intolerance is not the answer to intolerance (Battersby,

[10] The context in which Loretta Ross is doing this is revealed in Michael Powell's (2021) investigation of a battle over race, class and power at Smith College and commentary by Bret Stephens (2021).

[11] 'Doom scrolling' or 'schaden-surfing' (cf. Schadenfreude) is the practice of continuing to scroll through negative or bad news, even when it makes you feel bad. It feeds on negativity and taking a perverse delight in how bad things are.

[12] Dan Levin (2020) reports a nasty game of 'Gotcha' that played out at Heritage High School in Leesburg, Virginia. It reflects poorly on all involved—the protagonist, the antagonist and the audience.

2020), and an 'encroachment of the unsayable' (Stephens, 2020a) does not lead to a more free and equal society.

In July 2020, 150 public intellectuals, including Loretta Ross, published an open letter in *Harper's Magazine*, condemning attacks on free expression and voicing concern about 'an intolerance of opposing views, a vogue for public shaming and ostracism, and the tendency to dissolve complex policy issues in a blinding moral certainty' (Harper's Magazine, 2020). The open letter argued that 'the way to defeat bad ideas is by exposure, argument, and persuasion, not by trying to silence or wish them away'.

An editorial in Australian newspaper *The Age* commented:

> Many who partake in the cancel culture believe they have the right to determine what limitations should be enforced on free speech. When someone crosses the line, punishment is quick to follow. That is a problem. That is mob justice with no official adjudicator. We need to get back to debating ideas rather [than] deleting ideas. Everyone has the right to criticise the work and opinion of others – even harshly – as we all have the freedom to accept or reject feedback. But we need to pull back from this punitive purge of everything we don't like. It's malice to seek to have people sacked, to have their work banned, to demand history be erased, especially when those under attack are acting in good faith. It only leads to a greater fracturing of civil society into warring cliques, when there is much that people who differ can work on productively (The Age, 2020).

President Joe Biden, in his Inauguration Address on January 20, 2021, called upon a nation of citizens to join forces, treat each other with dignity and respect, 'stop the shouting, and lower the temperature' (Biden, 2021).

Government needs to invest, not just regulate

Merely passing and enforcing laws does not solve complex social problems. Governments need to encourage and support counter-speech strategies, as alternatives or complements to regulation.

Corey Brettschneider (2012, Chap. 4) argues that 'the state should pursue democratic persuasion, not only through expression, but also in its capacity as spender or subsidizer' (p. 109), and that 'state subsidies, combined with expression, can effectively counter the spread of hateful viewpoints while being non-coercive and compatible with the right of free speech' (p. 110). This might take the form, for example, of:

- Cultural policies that amplify capacity in minorities to speak for themselves (Malik, 2009, p. 106);
- Funding for public broadcasting that provides access to diverse information and ideas (Rowbottom, 2009);[13]
- Funding for public education programmes in civics, human rights and digital literacy, including development of critical thinking skills and ethically reflective use of social media

[13] In February 2021, the New Zealand Government announced a new, contestable fund with NZ$55m to support 'public interest journalism' (Faafoi, 2021). This needs to be allocated to public interest journalism across the political spectrum. See Sowman-Lund (2021) for a summary and commentary on details of the fund as announced in April 2021.

that teach people to slow down, fact check and evaluate evidence, and assess the reliability of sources (Chen, 2020);

- Outreach campaigns like the Council of Europe's 'No hate speech movement' (Council of Europe, 2020; Gerstmann, Güse, & Hempel, 2017) or the EU-funded project BRICkS / Building Respect on the Internet by Combating Hate Speech (von Lobenstein & Schneider, 2017; Wenzel, 2017);
- Grants to organisations that promote the values of free and equal citizenship; and
- Withholding non-profit status and tax privileges from groups that fail to respect the ideal of freedom and equality for all citizens.

Noting that the networked structure of the internet facilitates the dissemination of extremist messages and often makes removal of harmful digital communication impossible, Schmitt, Rieger, Ernst & Roth (2018) argue that 'equipping media users with critical (preventive) skills appears a more promising strategy than trying to block any exposure to extremist messages' (p. 1).[14] Digital literacy programmes also raise fewer privacy concerns than state-operated web filters. In Finland, for example, media literacy education starts in primary school, and Taiwan uses a 'humour not rumour' framework to debunk misinformation and disinformation online (Henk, 2021). In New Zealand, the Ministry of Education's decision to drop media studies from the curriculum for NCEA Level 1 (NZ Ministry of Education, 2020) is a step in the wrong direction (Hope, 2021).

Cynthia Miller-Idriss (2020, p. 92) notes that while progress has been made in educating young people (and older people) about their internet footprint and issues of online safety, privacy and cyber-bullying, much less is being done to help people 'recognize extremist content when they see it or understand how media manipulation and disinformation work, or how advertising algorithms may be shaping their consumer preferences or behavior.' She also notes that significant investment in research capacity and expertise is needed to understand the *where* and *when* of radicalisation, extremism and terrorism, and not only the *how* and *why* (Miller-Idriss, 2020, Conclusion).

Adequate, stable and ongoing funding of community and voluntary sector organisations is also required to build social cohesion, and for outreach programmes to minority ethno-cultural groups, particularly new migrants and former refugees, to enable integration, language learning, credentials recognition, job mentoring and employment support, and interaction with other social groups.

Hemmingsen and Castro (2017, p. 7) recommend that:

> In the broader population, the focus should be on bolstering general resources through capacity-building and inclusion; embracing diversity, openness, and freedoms to avoid feelings of marginalisation and the polarisation of society; strengthening critical thinking and knowledge about how propaganda works; and addressing real social or individual issues that

[14] See, for example, the resource produced by the Amadeu-Antonio-Stiftung with 33 social media tips for civil society (Darmstadt, Prinz, & Saal, 2020). The SIFT method for teaching digital literacy has been picked up by some US universities and Canadian high schools: (1) Stop; (2) Investigate the source; (3) Find better coverage, and (4) Trace claims, quotes and media to the original context (Warzel, 2021). A 2020 European Commission Science for Policy report by the Joint Research Centre notes, however, that business models prevalent in the online 'attention economy' constrain solutions that are achievable without regulatory intervention and that 'calls for greater "media literacy" or "critical thinking" are, by themselves, therefore likely to be insufficient to counteract any adverse effects on democracy from political online behaviour' (Lewandowsky et al., 2020, p. 15).

.

> may in the long run leave individuals vulnerable to any risks, including but not limited to involvement in extremist milieus.

The UK Government's CONTEST counter-terrorism strategy (UK Home Office, 2018) has four strands: Prevention, Pursuit, Protection and Preparedness. The Prevention Strand aims to stop people from becoming terrorists or supporting terrorism:

- Focus our activity and resources in those locations where the threat from terrorism and radicalisation is highest.
- Expand our Desistance and Disengagement Programme[15] with an immediate aim over the next 12 months to more than double the number of individuals receiving rehabilitative interventions.
- Develop a series of multi-agency pilots to trial methods to improve our understanding of those at risk of involvement in terrorism and enable earlier intervention.
- Focus our online activity on preventing the dissemination of terrorist material and building strong counter-terrorist narratives in order to ensure there are no safe places for terrorists online.
- Build stronger partnerships with communities, civil society groups, public sector institutions and industry to improve Prevent delivery.
- Re-enforce safeguarding at the heart of Prevent to ensure our communities and families are not exploited or groomed into following a path of violent extremism.

The UK Government is spending more than £2 billion per year on implementing its CONTEST strategy. In New Zealand, the Royal Commission of Inquiry (2020, p. 92) has recommended deep engagement of the national security system with communities, civil society, local government and the private sector. This will, of course, require significant public investment and sustained effort over many years.

In Germany, some 60 pro-democracy civic groups wrote to the federal government in November 2020 urging adequate and continuing funding, rather than time-limited, project-by-project funding. They warned that their work to uphold a liberal, open, democratic culture is at risk because of the funding model and the growth of the far right (Deutsche Welle, 2020).

An example of what can be done is *Respect!*, a programme of the Baden-Württemberg Democracy Centre in Germany (Demokratiezentrum Baden-Württemberg, n.d.). The Center invites citizen reporting of online incitement of hatred, anti-Semitic and anti-democratic incidents (including insults, marches, leaflet distribution, stickers and graffiti), provides a platform for reporting harmful digital communication and brokers confidential counselling for people affected by right-wing violence. The Centre is supported by state and federal funding.

In general, however, governments are advised to steer clear of direct involvement in counter-speech initiatives, because the identity and credibility of the counter-speech messenger is critical for getting through to key audiences. Governments can, however, play an indirect, facilitative role, supporting civil society efforts, and in some cases, it is appropriate for government to fund counter-narrative

[15] The UK's Desistance and Disengagement Programme focuses on rehabilitating individuals who have been involved in terrorism or terrorism-related activity and reducing the risk they pose to the UK (UK Home Office, 2019).

activities where this does not impact on the credibility of the product, campaign or message (Briggs & Feve, 2013, p. 26; cf. Gelber, 2012).

The private sector has a role to play

Money talks when companies are made aware when their advertisements appear on social media pages that disseminate harmful content. For example, in 2013, Women, Action and the Media (WAM!) and the Everyday Sexism Project in the UK launched a shared campaign that resulted in Nissan and the insurance company Nationwide pulling their advertisements from Facebook. The campaign was then extended and backed by online supporters and activists. As a result, 15 major companies decided to remove their advertisements from Facebook (Gargliadoni, Gal, Alves, & Martinez, 2015, p. 44).

More recently, a civil rights group-led Stop Hate for Profit campaign urged advertisers to put financial pressure on Facebook and Instagram to implement stricter policies against racism, violence, hate speech and election disinformation (Stop Hate for Profit, 2020). The initiative is led by the National Association for the Advancement of Coloured People, Colour of Change, the National Hispanic Media Coalition and the Anti-defamation League (Furness, 2020). By mid-June 2020, a temporary advertising boycott had grown to include more than 100 civil rights groups and more than 1100 companies, including Verizon, Hershey, Unilever, Volkswagen, Ford, Adidas, Microsoft, LEGO, Pepsi, Dunkin' Donuts, Target, Starbucks, North Face, Icebreaker and Stuff, New Zealand's largest news group (Anthony, 2020; Bell, 2020; Vance, 2020). Facebook's stock price fell by more than eight per cent on June 26, 2020.

In June 2020, Mark Zuckerberg announced that Facebook will remove posts—even from political leaders—that incite violence or attempt to suppress voting, and that the company will affix labels on posts that violate its other policies as well. He met with Stop Hate for Profit coalition leaders on July 7, 2020. Lerman and Timberg (2020) reported that 'the moves amount to major reversals amid rising public pressure, employee unrest and a burgeoning advertiser boycott over Facebook's long-standing refusal to more aggressively address hate speech and other platform violations from politicians such as [former] President Trump'. Civil rights leaders welcomed these 'modest concessions' but consider Facebook still is not doing enough to set effective policies and enforce them.

In support for the Christchurch Call, the New Zealand Super Fund issued a call to other big global investors to join its Social Media Investor Initiative, to pressure the boards of Twitter, Alphabet (Google) and Facebook to accept accountability for eradicating the live streaming of objectionable content on their social media platforms (Stock, 2020). By June 30, 2020 the initiative had 102 participating global investors representing NZD13.5 trillion in assets under management (NZ Super Fund, 2020, pp. 66–67). The report notes (p. 67):

> Facebook, Twitter and Alphabet have all moved to strengthen controls to prevent live-streaming and dissemination of objectionable content. However, as the additional mass shootings across the world have shown, the platforms are still open to abuse. Through the collective engagement, we have sought to represent the investor voice in the debate, ensuring that the companies know we expect them to fully manage this issue by investing in technical solutions and collaborating with other key industry players. These platforms are global; therefore, global solutions are needed. With a persistent narrative, we have kept the issue at the forefront of their minds.

The report records, however, 'no progress' on achieving board-level accountability and strengthened governance by Facebook, Twitter and Alphabet: 'Our experience in trying to engage these companies in an impactful way has lead us to draw a specific focus on proceeding with another round of (private) engagement where a smaller subgroup of investors seeks further meetings with the companies to push for strengthened governance at the board subcommittee level with regards to content' (p. 67).

The Fourth Estate can help restore civility to the public square

The Fourth Estate (the press and news media in their role as watchdogs of a free, open and democratic society) has a role to play in addressing negative or stereotyped portrayal and under-representation of minority social groups (UN General Assembly, 2015, Section B). UN Special Rapporteur Rita Izsák has commented:

> Poor reporting by the media on features such as ethnicity and religion involves, inter alia, labelling, selected use of data, generalizing incidents, negative stereotyping, giving one side of a story, use of derogatory words, mixing facts and views, absence of fact checking, and mismatching of the content of the text and headlines, images and sound. Lack of knowledge about ethnic and religious issues by media reporters, absence of in-house training, poor financial situation of media outlets, heavy workload and scarce time to prepare reports were highlighted as obstacles to good, unprejudiced reporting (UN General Assembly, 2015, para. 64).

Izsák comments that 'media outlets can actively engage in fighting incitement to hatred and violence in the media by adopting principles and guidelines of ethical and responsible journalism to improve the quality of information and reporting to avoid bias, prejudice and manipulation, as well as by promoting diversity among media workers and investing in adequate training for media professionals' (para. 87), consistent with the 2014 Brussels Declaration of the International Federation of Journalists (International Federation of Journalists, 2014). Her comments were reinforced in 2017 by Special Rapporteur Mutuma Ruteere (UN General Assembly, 2017, para. 21), specifically in relation to media bias in a counter-terrorism context.

The North-Rhine-Westphalian Media Authority (Landesanhalt für Medien NRW) has been pursuing awareness-raising and approaches to combating online hate crimes (Landesanhalt für Medien NRW, n.d.). The Media Authority notes that freedom of speech is often invoked as a defence for online hate and harassment, but argues that hate speech is a danger to society when it infringes on others' freedom of speech:

> Poisoned discourse can lead to people no longer having the courage to express their opinions due to a fear of hateful reactions. Some newsrooms avoid entire topics or close their comment sections because they don't feel confident moderating uncivilized disputes.

The NRW Media Authority monitors hate speech, collecting data from an annual Forsa survey, and in 2017 initiated its *Verfolgen statt nur Löschen* [Pursue instead of just deleting] campaign. The idea is that merely deleting or blocking illegal content is not enough—and can result in content that is criminally relevant no longer being available to support criminal investigation and prosecution. Neither is there any feedback to whoever posted the objectionable content. *Verfolgen statt nur Löschen* brings together representatives from media oversight agencies, law enforcement authorities and media companies to identify and sanction illegal hate speech on the internet and facilitate criminal prosecution. Notifications informed action by police in nine federal states in

November 2019. To date, almost 900 cases have been reported and more than 550 preliminary proceedings initiated. In 16 cases, a final conviction has been handed down. By February 2021, the initiative had grown to include more than 20 partners (Landesanstalt für Medien NRW, 2021).

The NRW Media Authority has also called on newsrooms to crack down firmly on hateful commenting and re-civilise the debate. It has developed resources to support this, including a 10-point plan to combat 'hate speech':

- Decisive moderation;
- A direct approach—not just deleting or blocking problematic content;
- Encourage and reward counter-speech creators;
- Devise journalistic programmes, formats and events that address the root causes of hate;
- Don't allow a loud minority to dominate the discourse;
- Regularly publish constructive, solution-oriented content;
- Use AI and algorithms to pre-filter content, so human moderators can concentrate on positive user posts;
- Create a moderation zone free of sarcasm and cynicism;
- Provide resources and enable a focus on content-related aspects of debate; and
- Earn respect by consistently maintaining boundaries with repeat offenders.

Kramp and Weichert (2018) provide examples of how newsrooms might implement each of the 10 points in a summary on pp. 6–9.

Media regulations that guide rather than proscribe may support effective performance of democratic functions by public service media (Rowbottom, 2009). This assumes, however, that media do actually function to 'refine and enlarge' our political passions (Garston, 2020) rather than being captured by an intellectual monoculture of 'censoriousness, groupthink and intellectual-risk aversion' that cares more about eliminating opposing points of view than engaging with them (Stephens, 2020b; cf. du Fresne, 2021).

Conclusion: The party's over and we all need to help clean up

Constraining harmful digital communication requires co-ordinated effort by multiple actors through 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. Internationally aligned anti-trust and 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 prohibition and censorship.

Jeffrey Howard argues that whether or not governments opt to criminalise harmful communication (he calls it 'dangerous expression'), all citizens have a moral duty to engage in counter-speech, depending on the gravity of the harms that need to be prevented and the risk that those harms will eventuate:

> Brandeis's canonical 'more speech, not enforced silence' formulation implies that we must choose between criminalization *or* counter-speech. But this is a false dichotomy. It is perfectly coherent to advocate banning dangerous expression, such as hate speech or

> terrorist advocacy, while nevertheless enjoining counter-speech as a vital supplement (Howard, 2019b, p. 2).

He adds that it is not enough to cheerlead the abstract importance of counter-speech and then sit back and hope for the best: ‘Someone needs to do the work, and do it well’ (Howard, 2019b, p. 1). And this, he argues, is a moral duty, a ‘samaritan’ duty, that falls on all citizens.

Carolin Emcke (2019, pp. xvii, 99) similarly reflects that:

> Certain forms of hatred fall under the responsibility of police and prosecutors. But the forms of exclusion and inclusion, the nasty little techniques of discrimination in gestures and habits, practices and beliefs—these are the responsibility of everyone in a society … When it comes to drying up the pools in which hatred grows (and not just terrorism and organized violence), when it comes to identifying as early as possible the mechanics of exclusion, the processes of an increasingly radical thinking, everyone everywhere is called upon to join in the efforts to prevent fanaticism: social milieus, neighbourhoods, peer groups, families, online communities ... Defending an open, plural society in which religious and political and sexual diversity can prosper is everyone’s duty.

In a free, open and democratic society, preserving ‘the floor of decency’ (Brooks, 2020) and restoring the lost art of civility (Bromell, 2019, pp. 39–41; Waldron, 2018) is everyone’s responsibility. Justus Bender (2021) reflects that the early years of the internet were a great party—wild and dangerous. But now the party is over, and all the guests have to help clean up.

And I mean *all* the guests. We don’t ‘clean up’ by silencing, imprisoning or willing into non-existence those with whom we disagree—even those with extremist views we judge to be abhorrent. In a liberal democracy, we cannot simply wish people away or make them non-persons for political purposes.

They are us. And there is no one who is not us. Singly and together, we exemplify the terrible ambiguity of the human condition and of the societies and cultures that shape us and that we in turn shape for those who come after us. As Terence (Publius Terentius Afer, the Roman African playwright) put it in the second century BCE: *Homo sum, humani nihil a me alienum puto*—I am a human being, and nothing human is alien to me.

In the historic district of Kaiserswerth, Düsseldorf, I saw the memorial on St Suitbertus to Friedrich Spee, a German Jesuit who fiercely opposed torture and the witch trials of the early seventeenth century. From the long history of heresy trials and the Inquisition, wars of religion and witch trials, we learn that attempts to shun, drive out or exterminate ideas and people from the virtuous, right-thinking and right-living society never have come to any good end.

Brenton Tarrant, the Christchurch mosque shooter, was convicted in a court of law of 51 charges of murder, 40 of attempted murder and a terrorist act. He was sentenced to life imprisonment without parole. But this does not change his status as a human being ‘born free and equal in dignity and rights’ (Universal Declaration of Human Rights, Article 1). Tarrant committed an atrocity, but he is nevertheless *one of us*, a fellow human being, and he is entitled to be cared for, treated justly and protected from cruel or inhuman treatment.

In a free, open and democratic society, we do not all need to like, approve of or agree with each other. We just need to respect our fundamental democratic principles of freedom and equality, and find ways to resolve our inevitable conflicts politically, without recourse to domination, humiliation, cruelty or violence.

If we do resort to violence, or incite others to violence, instead of attempting to persuade by reasoned argument in the public square and at the ballot box, then we can expect that the liberal state will apply the full force of the law to protect those we seek to harm. But for the rest, we have to learn to live together—and to take each other seriously as fellow citizens and human persons of equal dignity and worth.

This demands prudent and wise political leadership, and a pluralist, democratic politics characterised more by *agonism* than antagonism (Working paper 21/07, p. 13). It suggests a politics that steers away from a sectarian tribalism of 'us' and 'them'—a democratic politics of difference without exclusion (*they are us*), of community without identity or belonging, a 'togetherness of strangers' (Young, 1990, p. 237). Because we're all in this together, and civility is everyone's responsibility.

References

Allport, G. (1954). *The nature of prejudice.* Cambridge, MA: Addison-Wesley Pub. Co.

Amadeu-Antonio-Stiftung. (n.d.). Good gaming—well played democracy. Accessed December 15, 2020, from https://www.amadeu-antonio-stiftung.de/projekte/good-gaming-well-played-democracy/

Anthony, J. (2020). Icebreaker joins Facebook boycott, pledges to enhance inclusion and diversity. *Stuff,* July 15, 2020. Accessed October 20, 2020, from https://www.stuff.co.nz/business/industries/122124061/icebreaker-joins-facebook-boycott-pledges-to-enhance-inclusion-and-diversity

Bartlett, J., & Krasadomski-Jones, A. (2015). *Counter-speech: Examining content that challenges extremism online*. London: Demos. Accessed November 25, 2020, from https://demosuk.wpengine.com/wp-content/uploads/2015/10/Counter-speech-1.pdf

Bartlett, J., & Krasadomski-Jones, A. (2016). *Counter-speech on Facebook*. London: Demos. Accessed November 25, 2020, from https://demosuk.wpengine.com/wp-content/uploads/2016/09/Counter-speech-on-facebook-report.pdf

Battersby, J. (2020). More intolerance is not the answer to intolerance. *Stuff,* March 12, 2020. Accessed December 7, 2020, from https://www.stuff.co.nz/national/christchurch-shooting/120177092/more-intolerance-is-not-the-answer-to-intolerance

BBC News. (2020). Trump condemns all white supremacists after Proud Boys row. BBC News, October 2, 2020. Accessed November 16, 2020, from https://www.bbc.com/news/election-us-2020-54381500

Bell, E. (2020). For Facebook, a boycott and a long, drawn-out reckoning. *Colombia Journalism Review*, July 9, 2020. Accessed October 20, 2020, from https://www.cjr.org/tow_center/for-facebook-a-boycott-and-a-long-drawn-out-reckoning.php

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

Benesch, S., Ruths, D., Dillon, K., Saleem, H., & Wright, L. (2016a). *Counterspeech on Twitter: A field study.* A report for Public Safety Canada under the Kanishka Project. Accessed November 25, 2020, from https://www.scribd.com/document/327586365/Counterspeech-on-Twitter-A-Field-Study#from_embed

Benesch, S., Ruths, D., Dillon, K., Saleem, H., & Wright, L. (2016b). *Considerations for successful counterspeech.* A report for Public Safety Canada under the Kanishka Project. Accessed November 25, 2020, from https://dangerousspeech.org/considerations-for-successful-counterspeech/

Bennett, J. (2020). What if instead of calling people out, we called them in? *New York Times*, November 19, 2020. Accessed November 24, 2020, from https://nyti.ms/32UOqNN

Biden, J. (2021). Inaugural Address by President Joseph R. Biden, Jr. The White House, January 20, 2021. Accessed January 21, 2021, from https://www.whitehouse.gov/briefing-room/speeches-remarks/2021/01/20/inaugural-address-by-president-joseph-r-biden-jr/

Booker, C. (2020). The tyranny of woke. *Mail Online*, March 1, 2020. Accessed December 7, 2020, from https://www.dailymail.co.uk/debate/article-8060751/CHRISTOPHER-BOOKER-tyranny-woke.html

Brettschneider, C. (2012). *When the state speaks, what should it say? How democracies can protect expression and promote equality.* Princeton, NJ: Princeton University Press.

Briggs, R., & Feve, S. (2013). *Review of programs to counter narratives of violent extremism: What works and what are the implications for government?* London: Institute for Strategic Dialogue. Commissioned by Public Safety Canada. Accessed November 27, 2020, from https://apo.org.au/sites/default/files/resource-files/2013-12/apo-nid37101.pdf

Bromell, D. (2019). *Ethical competencies for public leadership: Pluralist democratic politics in practice.* Cham, CH: Springer.

Brooks, D. (2020). Trump's Presidency smashed the 'decency floor'. *New York Times,* October 28, 2020. Accessed November 7, 2020, from https://nyti.ms/3mvzduc

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

Buerger, C., and Wright, L. (2019). Counter-speech: A literature review. *Dangerous Speech.* Accessed November 25, 2020, from https://dangerousspeech.org/wp-content/uploads/2019/11/Counterspeech-lit-review_complete-11.20.19-2.pdf

Bundeszentrale für politische Bildung. (2015). Begriffswelten Islam. Accessed November 27, 2020, from https://www.bpb.de/lernen/digitale-bildung/medienpaedagogik/213243/webvideos-begriffswelten-islam

Chen, B. (2020). How to deal with a crisis of misinformation. *New York Times*, October 14, 2020. Accessed December 7, 2020, from https://nyti.ms/3nPMrDk

Community Research. (2020). #memeuptoracism, December 10, 2020. Accessed December 10, 2020, from https://www.facebook.com/hashtag/memeuptoracism

Council of Europe. (2020). No hate speech youth campaign. Accessed November 26, 2020, from https://www.coe.int/en/web/no-hate-campaign

Cresci, E. (2014). German town tricks neo-Nazis into raising thousands of euros for anti-extremist charity. *The Guardian*, November 18, 2014. Accessed November 20, 2020, from https://www.theguardian.com/world/2014/nov/18/neo-nazis-tricked-into-raising-10000-for-charity

Darmstadt, A., Prinz, M., & Saal, O. (2020). *Menschenwürde online verteidigen: 33 Social Media-Tipps für die Zivilgesellschaft.* Berlin: Amadeu-Antonio-Stiftung. Accessed December 16, 2020, from https://www.amadeu-antonio-stiftung.de/wp-content/uploads/2020/03/Broschu%CC%88re-CIVIC-Internet.pdf

Demokratiezentrum Baden-Württemberg. (n.d.). respect! Die Meldestelle für Hetze im Netz. Accessed November 27, 2020, from https://demokratiezentrum-bw.de/demokratiezentrum/vorfall-melden/#respect

Deutsche Welle. (2020). German democracy under 'open attack,' says SPD chief. *Deutsche Welle,* November 23, 2020. Accessed November 25, 2020, from https://p.dw.com/p/3lghA

du Fresne, K. (2021). New Zealand is being transformed, but not in a good way. *Spectator Australia,* February 13, 2021. Accessed February 17, 2021, from https://www.spectator.com.au/2021/02/new-zealand-is-being-transformed-but-not-in-a-good-way/

ECRI. (2016). ECRI general policy recommendation No. 15 on combating hate speech adopted on 8 December 2015. Strasbourg: Council of Europe, March 21, 2016. Accessed November 19, 2020, from https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01

Ellis, N. (2021). A Moonshot against extremism. *The Hill,* February 23, 2021. Accessed February 24, 2021, from https://thehill.com/policy/national-security/539976-a-moonshot-against-extremism

Emcke, C. (2019). *Against hate.* Cambridge: Polity Press. First published in German as *Gegen den Hass*, Frankfurt am Main: Fischer, 2016.

Faafoi, K. (2021). Government backs sustainable public interest journalism. Media release, February 12, 2021. Accessed February 15, 2021, from https://www.beehive.govt.nz/release/government-backs-sustainable-public-interest-journalism

Facebook. (2020). *Community guidelines, Hate speech.* Accessed November 25, 2020, from https://www.facebook.com/communitystandards/hate_speech

Farokhmanesh, M. (2019). Instagram 'tag cleaners' are fighting against digital vandalism. *The Verge*, July 19, 2019. Accessed January 29, 2021, from https://www.theverge.com/2019/7/19/20698192/instagram-moderation-tag-cleaners-digital-vandalism-gore-harassment-images-bianca-devins

Frenkel, S., & Feuer, A. (2021). 'A total failure': The Proud Boys now mock Trump. *New York Times,* January 20, 2021. Accessed January 21, 2021, from https://nyti.ms/2XZbh8g

Frischlich, L., Boberg, S., & Quandt, T. (2017). Unmenschlicher Hass: Die Rolle von Empfehlungsalgorithmen und Social Bots für die Verbreitung von Cyberhate. In K. Kasper, L. Gräßer, & A. Riffi (Eds.), *Online hate speech: Perpektiven auf eine neue Form des Hasses* (pp. 71–79). 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

Frischlich, L., Rieger, D., Morten, A., & Bente, G. (2017). Wirkung. In L. Frischlich, D. Rieger, A. Morten, & G. Bente (Eds.). *Videos gegen Extremismus? Counter-Narrative auf dem Prüfstand* (pp. 81–40). In co-operation with Forschungsstelle Terrorismus/Extremismus (FTE). Wiesbaden: Bundeskriminalamt.

Furness, H. (2020). Prince Harry, Meghan Markle back Facebook boycott in campaign against hate speech. *Stuff,* June 28, 2020. Accessed October 19, 2020, from https://www.stuff.co.nz/technology/social-networking/300044073/prince-harry-meghan-markle-back-facebook-boycott-in-campaign-against-hate-speech

Game over hate. (n.d.) Game over hate for inclusive gaming communities. Accessed December 16, 2020, from https://gameoverhate.eu/

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

Garsten, B. (2020). How to protect America from the next Donald Trump. *New York Times,* November 9, 2020. Accessed December 7, 2020, from https://nyti.ms/3ncwnux

Gelber, K. (2012). Reconceptualizing counterspeech in hate speech policy (with a focus on Australia). In M. Herz & P. Molnar (Eds), *The content and context of hate speech: Rethinking regulation and responses* (pp. 198–216). Cambridge: Cambridge University Press.

German Games Industry Association. (n.d.). *Hier spielt Vielfalt #TeamDiversity.* Accessed December 15, 2020, from https://hier-spielt-vielfalt.de/en/

Gerstmann, M., Güse, L., & Hempel, L. (2017). Wir müssen die rechte Gehirnhälfte erreichen. In K. Kasper, L. Gräßer, & A. Riffi (Eds.), *Online hate speech: Perpektiven auf eine neue Form des Hasses* (pp. 141–147). 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

Gillies, R. (2021). Canada designates the Proud Boys as a terrorist entity. *Associated Press*, February 4, 2021. Accessed February 4, 2021, from https://apnews.com/article/canada-proud-boys-terrorist-group-510b8cd8286f1207a726904f61e63e4d

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

Grubbs, J. (2020). Think twice before shouting your virtues online—moral grandstanding is toxic. *The Conversation,* January 14, 2020. Accessed December 7, 2020, from https://theconversation.com/think-twice-before-shouting-your-virtues-online-moral-grandstanding-is-toxic-128493

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). Cancel-Culture im Netz: Was sind das für Antirassisten? *Frankfurter Allgemeine Zeitung,* February 3, 2021. Accessed February 3, 2021, from https://www.faz.net/-gqz-a86nw

Harper's Magazine. (2020). A letter on justice and open debate, July 7, 2020. Accessed December 7, 2020, from https://harpers.org/a-letter-on-justice-and-open-debate/

Hehir, L. (2020). Online abuse and the standard you walk by. The Democracy Project, Victoria University of Wellington, October 20, 2020. Accessed December 7, 2020, from https://democracyproject.nz/2020/10/20/liam-hehir-online-abuse-and-the-standard-you-walk-by/

Hemmingsen, A.-S., & Castro, K. (2017). *The trouble with counter-narratives.* DIIS Report 2017:1. Copenhagen: Danish Institute for International Studies. Accessed November 27, 2020, from https://pure.diis.dk/ws/files/784884/DIIS_RP_2017_1.pdf

Henk, M. (2021). Facebook and the disinformation wars. *NZ Herald,* February 22, 2021. Accessed February 22, 2021, from https://www.nzherald.co.nz/business/mandy-henk-facebook-and-the-disinformation-wars/BD4UIWGWXDBWNNO2GPTJ44CXPM/

Hensen, L. (2020). Neo-Nazis face an old enemy in Grannies Against the Right. *Deutsche Welle*, November 14, 2020. Accessed November 25, 2020, from https://p.dw.com/p/3lDf2

Hinesley, K. (2020). Our #DataforGood partnership with New Zealand's NCPACS. Twitter blog, March 11, 2020. Accessed December 7, 2020, from https://blog.twitter.com/en_us/topics/company/2020/christchurch-otago-nspacs.html

Hope, W. (2021). In an age of digital disinformation, dropping level 1 media studies in NZ high schools is a big mistake. *The Conversation,* February 3, 2021. Accessed February 16, 2021, from https://theconversation.com/in-an-age-of-digital-disinformation-dropping-level-1-media-studies-in-nz-high-schools-is-a-big-mistake-151475

Howard, J. (2018). Should we ban dangerous speech? *British Academy Review*, *32*, 19–21. Accessed November 2, 2020, from https://www.thebritishacademy.ac.uk/publishing/review/32/should-we-ban-dangerous-speech/

Howard, J. (2019a). Dangerous speech. *Philosophy & Public Affairs, 47*(2), 208–254. https://doi.org/10.1111/papa.12145

Howard, J. (2019b). Terror, hate and the demands of counter-speech. *British Journal of Political Science,* 1–16. https://doi.org/10.1017/S000712341900053X

Institute for Strategic Dialogue. (n.d.). Institute for Strategic Dialogue: Powering solutions to extremism, hate and disinformation. Accessed December 17, 2020, from https://www.isdglobal.org/

International Federation of Journalists. (2014). IFJ conference agrees declaration to stand up against hate speech, April 25, 2014. Accessed November 26, 2020, from https://www.ifj.org/media-centre/news/detail/category/press-releases/article/ifj-conference-agrees-declaration-to-stand-up-against-hate-speech.html

Jigsaw. (n.d.). *The Redirect Method: A blueprint for bypassing extremism*. Accessed November 27, 2020, from https://redirectmethod.org/

Keinen Pixel den Faschisten. (2020). Was ist 'Keinen Pixel den Faschisten!'?, April 22, 2020. Accessed December 15, 2020, from https://keinenpixeldenfaschisten.de/2020/04/22/pressemitteilung/

Kelly, J. (2021). How 'fact-checking' can be used as censorship. *NZ Herald,* February 19, 2021. Accessed February 19, 2021, from https://www.nzherald.co.nz/business/how-fact-checking-can-be-used-as-censorship/BKI6WCRJ4H5VVSCGUMFXMGSAN4/

Kramp, L., & Weichert, S. (2018). *Hateful commenting online: Control strategies for newsrooms*. Düsseldorf: Landesanstalt für Medien NRW. Accessed April 20, 2021, from https://www.medienanstalt-nrw.de/fileadmin/user_upload/lfm-nrw/Foerderung/Forschung/Dateien_Forschung/Hateful_Commenting_Online_Control_Strategies_for_Newsrooms.pdf

Kunst, M., Porten-Cheé, P., Emmer, M., & Eilders, C. (2021). Do 'good citizens' fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Information Technology & Politics.* https://doi.org/10.1080/19331681.2020.1871149

Landesanhalt für Medien NRW. (n.d.). *Hate speech on the internet*. Accessed October 16, 2020, from https://www.medienanstalt-nrw.de/about-us/topics/hate-speech-on-the-internet.html

Landesanstalt für Medien NRW. (2021). 'Verfolgen statt nur löschen' zieht Zwischenbilanz. Media release, February 23, 2021. Accessed February 25, 2021, from https://www.medienanstalt-nrw.de/presse/pressemitteilungen-2021/2021/februar/verfolgen-statt-nur-loeschen-zieht-zwischenbilanz.html

Lerman, R., & Timberg, C. (2020). Bowing to pressure, Facebook will start labeling violating posts from politicians. But critics say it's not enough. *Washington Post*, June 27, 2020. Accessed October 13, 2020, from https://www.washingtonpost.com/technology/2020/06/26/facebook-hate-speech-policies/

Levin, D. (2020). A racial slur, a viral video, and a reckoning. *New York Times,* December 26, 2020. Accessed February 15, 2021, from https://www.nytimes.com/2020/12/26/us/mimi-groves-jimmy-galligan-racial-slurs.html

Lewandowsky, S., Smillie, L., Garcia, D., Hertwig, R., Weatherall, J., Egidy, S., Robertson, R.E., O'connor, C., Kozyreva, A., Lorenz-Spreen, P., Blaschke, Y., & Leiser, M. (2020). *Technology and democracy: Understanding the influence of online technologies on political behaviour and decision-making***.** EUR 30422 EN, Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/709177

Ley, H. (2018). *#ICHBINHIER: Zusammen gegen Fake News und Hass im Netz*. Cologne: Dumont.

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/

Malik, M. (2009). Extreme speech and liberalism. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 96–120). Oxford: Oxford University Press.

Metz, C. (2021). Feeding hate with video: A former alt-right YouTuber explains his methods. *New York Times,* April 15, 2021. Accessed April 20, 2021, from https://nyti.ms/3aeoUXE

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

Molnar, P. (2012). Interview with Theodore Shaw. In M. Herz & P. Molnar (Eds), *The content and context of hate speech: Rethinking regulation and responses* (pp. 399–413). Cambridge: Cambridge University Press.

Morten, A., Frischlich, L., Rieger, D., & Bente, G. (2017). Wirksamkeit. In L. Frischlich, D. Rieger, A. Morten, & G. Bente (Eds.). *Videos gegen Extremismus? Counter-Narrative auf dem Prüfstand* (pp. 161–224). In co-operation with Forschungsstelle Terrorismus/Extremismus (FTE). Wiesbaden: Bundeskriminalamt.

Nussbaum, M. (2016). *Anger and forgiveness: Resentment, generosity, justice.* New York: Oxford University Press.

NZ Ministry of Education. (2020). NCEA Level 1 subject changes to give students a broader foundational education. Media release, December 3, 2020. Accessed February 16, 2021, from https://www.education.govt.nz/news/ncea-level-1-subject-changes/

NZ Super Fund. (2020). *The test of time | Tā te wā whakamātau*. Guardians of New Zealand Superannuation, annual report 2020, October 19, 2020. Accessed October 20, 2020, from https://www.nzsuperfund.nz/news-and-media/guardians-of-new-zealand-superannuation-annual-report-2020/

Obama, B. (2016). Commencement address, Howard University, May 7, 2016. Accessed November 19, 2020, from https://www.politico.com/story/2016/05/obamas-howard-commencement-transcript-222931

Omas gegen Rechts. (2018). Grundsatz, September 23, 2018. Accessed November 25, 2020, from http://www.omasgegenrechts.de/grundsatztext/

Paluck, E., Green, S., & Green, D. (2018). The contact hypothesis re-evaluated*. Behavioural Public Policy, 3*(2), 129–158. https://doi.org/10.1017/bpp.2018.25

Pettigrew, T., & Tropp, L. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology, 90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Powell, M. (2021). Inside a battle over race, class and power at Smith College. *New York Times,* February 24, 2021. Accessed February 26, 2021, from https://nyti.ms/3upi1LL

Quinn, P. (2020). Don't let the nasties take over our political discourse. *Stuff,* January 6, 2020. Accessed December 8, 2020, from https://www.stuff.co.nz/national/politics/opinion/118571161/dont-let-the-nasties-take-over-our-political-discourse

Raue, S. (2020). Rettet die Meinung! *Frankfurter Allgemeine Zeitung,* October 22, 2020. Accessed November 16, 2020, from https://www.faz.net/aktuell/feuilleton/medien/zum-zustand-der-meinungsfreiheit-17013130.html

Rosenblatt, K. (2020). Gay Twitter users flood #ProudBoys hashtag with LGBTQ pride images. NBC News, October 5, 2020. Accessed November 16, 2020, from https://www.nbcnews.com/feature/nbc-out/twitter-users-flood-proudboys-hashtag-gay-pride-images-n1242101

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 and Society, 2*(3), 1–13. https://doi.org/10.1177/2056305116664220

Rowbottom, J. (2009). Extreme speech and the democratic functions of the mass media. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 608–630). Oxford: Oxford University Press.

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/

Schmid, T., & Appelhoff, M. (2017). Die Initiative „Netzkodex“. In K. Kasper, L. Gräßer, & A. Riffi (Eds.), *Online hate speech: Perpektiven auf eine neue Form des Hasses* (pp. 157–160). 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

Schmitt, J., Caspari, C., Wulf, T., Bloch, C., & Rieger, D. (forthcoming). Two sides of the same coin? The persuasiveness of one-sided vs. two-sided narratives in the context of radicalization prevention. *Studies in Communication and Media.*

Schmitt, J., Rieger, D., Ernst, J., & Roth, H.-J. (2018). Critical media literacy and Islamist online propaganda: Feasibility, applicability and impact of three learning arrangements. *International Journal of Conflict and Violence*, *12*, 1–19. https://doi.org/10.4119/UNIBI/ijcv.642

Schmitt, J., Rieger, D., Rutkowski, O., & Ernst, J. (2018). Counter-messages as prevention or promotion of extremism?! The potential role of YouTube recommendation algorithms. *Journal of Communication*, *68*(4), 780–808. https://doi.org/10.1093/joc/jqy029

Sowman-Lund, S. (2021). ‘Three pillar’ approach for new $55 million public interest journalism fund revealed. *The Spinoff*, April 8, 2021. Accessed April 10, 2021, from https://thespinoff.co.nz/media/08-04-2021/three-pillar-approach-to-new-55-million-public-interest-journalism-fund-revealed/

Stephens, B. (2020a). The encroachment of the unsayable: Our compromised liberalism has left a generation of writers weighing their words in fear. *New York Times*, Opinion, October 19, 2020. Accessed October 20, 2020, from https://nyti.ms/31j3BQj

Stephens, B. (2020b). Groupthink has left the Left blind. *New York Times,* November 16, 2020. Accessed December 7, 2020, from https://nyti.ms/2UvGEFN

Stephens, B. (2021). Smith College and the failing liberal bargain. *New York Times,* March 1, 2021. Accessed March 2, 2021, from https://nyti.ms/2Mlo5hm

Stock, R. (2020). Twitter, Facebook, Google boards ignore NZ Super Fund's call for accountability. *Stuff,* October 20, 2020. Accessed October 20, 2020, from https://www.stuff.co.nz/business/123134827/twitter-facebook-google-boards-ignore-nz-super-funds-call-for-accountability

Stop Hate for Profit. (2020). Statement from Stop Hate for Profit on July 2020 ad pause success and #StopHateForProfit Campaign, July 30, 2020. Accessed October 19, 2020, from https://www.stophateforprofit.org/

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

The Age. (2020). Don’t just delete ideas, debate them. Editorial, July 13, 2020. Accessed December 7, 2020, from https://www.theage.com.au/national/don-t-just-delete-ideas-debate-them-20200713-p55bih.html

Turner, Z. (2018). Facebook, Google have a tough new job in Germany: Content cop. *Wall Street Journal*, January 10, 2018. Accessed October 13, 2020, from https://www.wsj.com/articles/facebook-google-have-a-tough-new-job-in-germany-content-cop-1515605207

UK Home Office. (2018). CONTEST: The United Kingdom’s strategy for countering terrorism, June 2018. Accessed December 7, 2020, from https://www.gov.uk/government/publications/counter-terrorism-strategy-contest-2018

UK Home Office. (2019). Fact sheet: Desistance and disengagement programme. Home Office news team, November 5, 2019. Accessed April 20, 2021, from https://homeofficemedia.blog.gov.uk/2019/11/05/fact-sheet-desistance-and-disengagement-programme/

UN General Assembly. (2015). Report of the Special Rapporteur on minority issues, Rita Izsák, January 5, 2015. UN Docs A/HRC/28/64. Accessed November 25, 2020, from https://documents-dds-ny.un.org/doc/UNDOC/GEN/G15/000/32/PDF/G1500032.pdf?OpenElement

UN General Assembly. (2017). Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance, Mutuma Ruteere, August 4, 2017. UN Docs A/72/287. Accessed December 9, 2020, from https://digitallibrary.un.org/record/1304009?ln=en

Vance, A. (2020). Andrea Vance: Why Jacinda Ardern won't delete her Facebook account. *Stuff,* August 2, 2020. Accessed October 20, 2020, from https://www.stuff.co.nz/national/politics/opinion/122308433/andrea-vance-why-jacinda-ardern-wont-delete-her-facebook-account

Velasquez-Manoff, M. (2017). How to make fun of Nazis. *New York Times*, August 17, 2017. Accessed November 20, 2020, from https://nyti.ms/2vHcTpH

Vlastos, G. (1984). Justice and equality. In J. Waldron, (Ed.), *Theories of rights* (pp. 41–76). Oxford: Oxford University Press. Originally published in R. Brandt (Ed.), *Social justice* (pp. 31–72). Englewood Cliffs, NJ: Prentice-Hall, 1962.

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

von Lobenstein, M.-J., & Schneider, A. (2017). Das Project BRICkS: Auswertung von Hate Speech-Fallbeispielen. In K. Kasper, L. Gräßer, & A. Riffi (Eds.), *Online hate speech: Perspektiven auf eine neue Form des Hasses* (pp. 127–133). 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

Wagner, J. (2020). Are memes the digital gateway to social media manipulation? *Deutsche Welle,* November 16, 2020. Accessed December 16, 2020, from https://www.dw.com/en/are-memes-the-digital-gateway-to-social-media-manipulation/a-55622104

Wagner, K. (2016). Want to combat hate speech on Facebook? Try a 'Like attack,' says COO Sheryl Sandberg. *Recode,* January 21, 2016. Accessed November 25, 2020, from https://www.vox.com/2016/1/21/11588986/want-to-combat-hate-speech-on-facebook-try-a-like-attack-says-coo

Waldron, J. (2012). *The harm in hate speech.* Cambridge, MA: Harvard University Press.

Waldron, J. (2017). *One another's equals: The basis of human dignity.* Cambridge, MA: Belknap Press.

Waldron, J. (2018). *Polls apart: Reclaiming respect in a time of polarised politics*. Sir John Graham Lecture 2017. Auckland, NZ: Maxim Institute. Accessed December 8, 2020, from https://www.maxim.org.nz/sjgl2017/

Warzel, C. (2021). Don't go down the rabbit hole. *New York Times,* February 18, 2021. Accessed February 18, 2021, from https://www.nytimes.com/2021/02/18/opinion/fake-news-media-attention.html

Wenzel, J. (2017). Gemeinsam gegen Hate Speech: Workshops mit Jugendlichen. In K. Kasper, L. Gräßer, & A. Riffi (Eds.), *Online hate speech: Perpektiven auf eine neue Form des Hasses* (pp. 135–140). 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

Young, I. (1990). *Justice and the politics of difference.* Princeton, NJ: Princeton University Press.

ZDF. (2016). Be Deutsch! [Achtung! Germans on the rise]. ZDF Magazin Royale with Jan Böhmermann, March 31, 2016. Accessed November 30, 2020, from https://www.youtube.com/watch?v=HMQkV5cTuoY&list=WL&index=4