After Christchurch:
Hate, harm and the limits of censorship

2. ‘Hate speech’: Defining the problem and some key terms

David Bromell

Working Paper 21/03

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

CAIS CENTER FOR ADVANCED INTERNET STUDIES

INSTITUTE FOR GOVERNANCE AND POLICY STUDIES
WORKING PAPER
21/03

MONTH/YEAR

March 2021

AUTHOR

David Bromell
Senior Associate
Institute for Governance and Policy Studies

INSTITUTE FOR GOVERNANCE AND POLICY STUDIES

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

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

ACKNOWLEDGEMENT

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

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

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

The seven working papers in this series are:

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

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

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

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

Contents

‘Hate speech’: Defining the problem and some key terms

Abstract

Following the terrorist attack on Christchurch mosques in March 2019, the Government pledged to review New Zealand’s regulation of ‘hate speech’ and ‘hate crime’. The Royal Commission of Inquiry that reported in November 2020 made four recommendations on ‘hate speech’ and ‘hate crime’, to which the Government has agreed in principle.

This paper summarises 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 the right to freedom of expression.

The paper offers two definitions informed by international human rights law, scholarly debate and existing regulation in the United Kingdom, Denmark, Canada, Germany and New Zealand:

- A **‘hate crime’** involves the commission of a criminal offence, for example assault and injury to another person, or damage to property, associated with a motivation and/or demonstration of hostility to the victim as a member of a social group with a common ‘protected characteristic’ such as nationality, race or religion.
- **‘Hate speech’** is public communication that incites discrimination, hostility or violence against members of a social group with a common ‘protected characteristic’ such as nationality, race or religion.

A democratic state can justifiably use its coercive powers to protect its citizens from harmful public communication that incites discrimination, hostility or violence against them.

A democratic state cannot justifiably restrict freedom of opinion and expression by criminalising criticism, satire, disapproval, dislike, ‘hurtful’ remarks—or even hatred. 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.

The remaining five working papers in this series develop this argument further and elaborate on challenges in regulating online content (Working paper 21/04), current legal frameworks for regulating harmful communication (Working paper 21/05), arguments for and against restricting freedom of expression (Working paper 21/06), striking a fair balance when regulating harmful communication (Working paper 21/07), and counter-speech and civility as everyone’s responsibility (Working paper 21/08).

Tags: #ChristchurchAttack #ChristchurchCall #hatespeech #hatecrime #censorship #freespeech

Introduction: ‘Hate speech’ and ‘hate crime’

Since the terrorist attack on Christchurch mosques on March 15, 2019 the Islamic Women’s Council, Gamal Fouda (Imam of Masjid Al Noor, the site of the first attack), the Federation of Islamic Associations of New Zealand (FIANZ), and others have called for specific recognition of ‘hate crimes’ and ‘hate speech’ in New Zealand law, a safe system (with a single process) to report ‘hate speech’ and ‘hate crime’, and for that system to be linked to security agencies’ databases (FIANZ, 2020, pp. 124–128; RNZ, 2020).[1]

Anjum Rahman from the Islamic Women’s Council was reported in January 2020 as arguing that New Zealand’s current laws are not fit for purpose:

> It is currently difficult to address expressions of extreme hate and bullying, abusive language which are not directed to a specific person but are made against whole communities or groups of people. There are gaps between state agencies as to who is responsible in relation to this kind of expression. Our concerns are especially around the kind of online and offline hate directed at groups of people, that encourages violence and harassment towards them (quoted in Kenny, 2020a).

Two weeks after the Christchurch mosque attacks on March 15, 2019 Justice Minister Andrew Little initiated a review of New Zealand's existing ‘hate speech’ legislation (Duff, 2019). In an opinion piece, he commented:

> … in the immediate wake of the March 15 mosque attacks, many citizens from minority ethnic and religious communities told of how opinions and statements they routinely see on social media and other public platforms make them feel threatened, unwelcome and alienated. Others have said these types of statements allow a climate to develop that is tolerant of harmful discriminatory expression. A responsible government must consider these claims, and on a principled basis (Little, 2019).

Noting that the current legal framework imposes sanctions on incitement of disharmony on racial grounds but not, for example, on grounds of religious faith, the Minister asked the Ministry of Justice to work with the Human Rights Commission to examine whether New Zealand law properly balances the issues of freedom of speech and ‘hate speech’. He stated: ‘The process should not be rushed, and I expect a report for public comment towards the end of the year’ [i.e., 2019]. Given that ‘drawing the line is not simple’, he noted the need for ‘a robust public discussion from all quarters’ (Little, 2019).

In March 2020, almost a year after his initial announcement, the Minister advised that options were ‘working their way’ through the Cabinet process and he expected there would be an announcement in a matter of weeks (Devlin, 2020a). The Ministry of Justice had consulted with ‘affected communities’, and the Human Rights Commission had ‘facilitated a series of community conversations with groups of people who may have experienced, or been at risk of experiencing, harmful speech’ (Devlin, 2020a). The promised ‘robust public discussion from all quarters’ had not occurred.[2]

[1] See also the FIANZ (2021) submission to Hon. Andrew Little in February 2021, following an engagement process with Muslim communities following the release of the report of the Royal Commission of Inquiry in December 2020.

[2] This pre-dates the Covid-19 lockdown that began with closing New Zealand’s border on March 19, 2020 and a Level 4 nationwide lockdown beginning on March 25.

In June 2020, the Minister said Labour was still in talks with its support parties and that legislation was not likely to go to Cabinet until after the general election (Devlin, 2020b). Due to Covid-19, the election was in turn postponed from September 19 to October 17, 2020. In the seventh Labour Government's Ministerial List announced on November 2, 2020, Andrew Little was replaced by Kris Faafoi as Minister of Justice.

On December 8, 2020, the New Zealand Government released the report of the Royal Commission of Inquiry into the Terrorist Attack on Christchurch Masjidain on 15 March 2019 (Royal Commission of Inquiry, 2020a), with a companion report on 'hate speech'- and 'hate crime'-related legislation (Royal Commission of Inquiry, 2020b). Prime Minister Jacinda Ardern announced the appointment of Andrew Little as co-ordinating Minister for the Government's implementation of the 44 recommendations in the report, to which the Government has agreed in principle (Ardern, 2020).

The Royal Commission of Inquiry reported that it had considered human rights principles throughout its inquiry, noting that these affect 'the balance between freedom of expression and the expression of views that are hateful toward members of New Zealand's ethnic and religious communities' (Royal Commission of Inquiry, 2020a, p. 86). Four of its 44 recommendations concern 'hate speech' and 'hate crime':

> **39. Amend legislation** to create hate-motivated offences in:
> a) the Summary Offences Act 1981 that correspond with the existing offences of offensive behaviour or language, assault, wilful damage and intimidation; and
> b) the Crimes Act 1961 that correspond with the existing offences of assaults, arson and intentional damage.
>
> **40. Repeal** section 131 of the Human Rights Act 1993 and insert a provision in the Crimes Act 1961 for an offence of inciting racial or religious disharmony, based on an intent to stir up, maintain or normalise hatred, through threatening, abusive or insulting communications with protected characteristics that include religious affiliation.
>
> **41. Amend** the definition of 'objectionable' in section 3 of the Films, Videos, and Publications Classification Act 1993 to include racial superiority, racial hatred and racial discrimination.
>
> **42. Direct** New Zealand Police to revise the ways in which they record complaints of criminal conduct to capture systematically hate-motivations for offending and train frontline staff in:
> a) identifying bias indicators so that they can identify potential hate crimes when they perceive that an offence is hate-motivated;
> b) exploring perceptions of victims and witnesses so that they are in a position to record where an offence is perceived to be hate-motivated; and
> c) recording such hate-motivations in a way which facilitates the later use of section 9(1)(h) of the Sentencing Act 2002 (Royal Commission of Inquiry, 2020a).

In relation to hate speech and hate crime, the Prime Minister's press statement included the following commitments:

> We will establish the New Zealand Police programme Te Raranga, The Weave, to make improvements in Police's frontline practice to identify, record, and manage hate crime, and deliver a service that is more responsive to victims.
>
> We will also increase the capacity of the Human Rights Commission by increasing the funding so they can develop a team of highly skilled individuals who can provide mediation, facilitate conversations or be more proactive in exercising the Commission's inquiry function.
>
> We also propose to continue our work to update our current hate speech legislation. We are conscious there are a range of views on this issue. We will be undertaking consultation with

> community groups and parties from right across Parliament to test these proposals before bringing forward legislative change. I do want to emphasise though, these are issues that are longstanding, they predate March 15, and they effect many members of the community, including our LGBTIQ community, and different and diverse religions. We will take the time to get it right (Ardern, 2020).

In a media statement, Minister of Justice Chris Faafoi reiterated the Government's intention to strengthen laws related to hate-motivated activity and inciting hatred against an individual or group:

> Speech which is abusive or threatening and incites hostility towards a group or person can cause significant harm. In line with the Royal Commission of Inquiry's recommendations, Cabinet has agreed to a number of measures to improve provisions in the Human Rights Act (1993) relating to incitement. The Government intends to establish an engagement process with community groups to discuss these changes (Faafoi, 2020).

In Question Time in Parliament on December 8, 2020, Ardern acknowledged that implementing the Royal Commission of Inquiry's recommendations will not be straightforward:

> The Government accepts the findings of its [the Royal Commission of Inquiry's] report and agrees, in principle, to all recommendations. Implementing some of the recommendations will require further consideration (Hansard, 2020).

Asked specifically about the Government's intentions to 'follow the royal commission's advice and implement British-style hate speech laws without the exemptions for free and open debate present in that country's laws,' Ardern replied:

> What we will be seeking to do is to work across Parliament. We do want consensus where we can build it, because, of course, that will stop this debate becoming divisive and potentially leading to the targeting of certain communities (Hansard, 2020; cf. Adams, 2021).

The Royal Commission of Inquiry's recommendations are discussed further in Working paper 21/05, **Regulating harmful communication: Current legal frameworks**.

How much of a problem is 'hate speech' and 'hate crime'?

Because 'hate crime' is not currently an offence in and of itself in New Zealand, incidents are not recorded by New Zealand Police as a specific offence type, limiting the available data (Ensor, 2020).

New Zealand's Human Rights Commission has, however, published a summary of media reports of racially and religiously motivated crime in New Zealand between 2004 and 2012, noting that 'the absence of systematically collected data and information on racially and religiously motivated crime in New Zealand makes it very difficult to have an informed discussion about their prevalence and design effective measures to counter them' (Human Rights Commission, 2019a, p. 1).

In the 2019 New Zealand Crime and Victims Survey, respondents were asked if they thought incidents they had experienced were motivated by discrimination—that is, by the offender's attitude towards the victim's race, sex, gender identity, sexual orientation, age, religion or disability. The survey found that:

- Twenty-five per cent of all incidents and 32 per cent of all personal offences were seen by the victim as motivated by discriminatory attitudes;

- Sexual assault (82 per cent), threats and damages (34 per cent) and physical offences (assault and robbery) (34 per cent) were the most common offence types to be considered by the victim as having been driven by discrimination; and
- Twenty-three per cent of victims of Asian ethnicity felt that the incidents that happened to them were driven by discrimination towards their race, ethnicity or nationality, compared to seven per cent of victims overall (NZ Ministry of Justice, 2020).

Netsafe and the Islamic Women's Council have called for better recording of 'hate crimes' in light of Netsafe's 2019 survey of online 'hate speech' (Kenny, 2020a; Netsafe, 2019). New Zealand Police said in a statement in February 2020 that they were 'working actively to create tracking (monitoring) resources that will allow [officers] to flag reported "hate crimes" and/or incidents within [information technology systems] and to allow timely access to data for these types of offences/incidents' (Kenny, 2020b). These are the improvements to NZ Police's Te Raranga|The Weave programme flagged by the Prime Minister (Ardern, 2020).

Research by Australia's eSafety Commissioner with New Zealand's Netsafe and the UK Safer Internet Centre (eSafety Commissioner, 2020) found that in Australia, around 14% of the adult population was estimated to have been the target of online 'hate speech' in the 12 months to August 2019. In New Zealand, this was around 15% in the 12 months to June 2019. In both countries, younger adults were more likely to have experienced 'hate speech'.

This finding is similar to the 18% of young people in the European SELMA project who reported experiencing 'hate speech' over a period of three months. While religion, political views, race and gender were the most common reasons cited in both Australia and New Zealand for experiencing 'hate speech', young people interviewed as part of the SELMA project were instead more likely to be targeted because of their appearance and their sexuality. 'Hate speech' was found to spread through several popular online channels, and a similar proportion (5–6 per cent) of participants in the Australian and New Zealand studies acknowledged intentionally visiting sites that target others or promote 'hate speech' (eSafety Commissioner, 2020, pp. 6, 16).[3]

In 2020, in the annual survey on online hate speech awareness conducted by Forsa for the State Media Authority of North Rhine-Westphalia, Germany, 94 per cent of respondents aged 14 to 24 stated that they had observed hate speech on the internet. Within the last five years, the proportion of those who stated that they have reported hate speech has almost doubled from 34 to 67 per cent. Thirty-eight per cent of respondents considered that public postings on the internet are more often hate comments than objective expressions of opinion, rising to as much as 50 per cent among younger respondents (Landesanstalt für Medien NRW, n.d.)

In the largest survey to date on 'hate speech' in Germany, conducted in 2019 by YouGov and evaluated by the Institute for Democracy and Civil Society, 40 per cent had observed 'hate speech' online and eight per cent of respondents said they had been personally affected, with 18-to-24-year-olds (17 per cent) and people from immigrant families (14 per cent) reporting significantly higher values (Geschke, Klaßen, Quent, & Richter, 2019). Sixteen per cent of German internet users stated they had left social networks on account of hate-related content (ibid., p. 28).

[3] Extremist ideologies circulate in online forums, blogs, media sites and videos, so the odds of stumbling across extremist material is high (Ernst, Schmitt, Rieger, Beier, Vorderer, Bente, & Roth, 2017, p. 2). See further Working paper 21/04, **Challenges in regulating online content**.

Survey findings like this reflect, however, broad and imprecise definitions of 'hate speech'. Summarising Australian adults' understanding of 'hate speech', the eSafety Commissioner (2020, p. 7) explained that:

> Hate speech was frequently noted as anything negative that was directed at another person. Therefore, it was seen as going beyond the incitement or spreading of hate to communication that is hurtful, or which simply causes offence.[4]

If our definitions fail to distinguish clearly between 'hate speech' and 'hate crimes', or between criticism, causing offence, and inciting others to discrimination, hostility and/or violence, we increase the risk of people talking past each other when discussing regulatory and non-regulatory responses to harmful communication.

There is a boundary between free speech and harmful extremism, even if identifying that boundary is difficult (Royal Commission of Inquiry, 2020a, p. 534). Based on the protection of freedom of expression in section 14 of the New Zealand Bill of Rights Act 1990, the Royal Commission of Inquiry (2020b, p. 11) expressed 'considerable reservations whether a police policy of investigating and recording non-crime hate incidents would withstand legal scrutiny in New Zealand.'

'Hate crime' and 'hate speech' are different but related concepts

'*Hate crimes'* involve the commission of an offence, for example assault and injury to another person, or damage to property, associated with a motivation and/or demonstration of hostility to the victim as a member of a group with a common 'protected characteristic', such as nationality, race or religion[5] (Royal Commission of Inquiry, 2020a, p. 700; Human Rights Commission, 2019b, p. 7).

Jeremy Waldron (2012, p. 35) explains that in 'hate crime' legislation, hatred as motivation is treated as a distinct element of the crime, or as an aggravating factor as in New Zealand law, where hate motivation can be taken into consideration at sentencing (Sentencing Act 2002).[6]

[4] The definition of 'hate speech' that informed the survey conducted in Germany by YouGov was 'aggressive or generally pejorative statements against persons who are assigned to certain groups' (Geschke, Klaßen, Quent, & Richter, 2019, p. 15).

[5] Protected characteristics (prohibited grounds of discrimination) in New Zealand's Human Rights Act 1993 s21 are sex, marital status, religious belief, ethical belief, colour, race, ethnic or national origins, disability, age, political opinion, employment status, family status and sexual orientation. Protected characteristics in the Harmful Digital Communications Act 2015 s6(1) are colour, race, ethnic or national origins, religion, gender, sexual orientation and disability.

[6] In New Zealand's Sentencing Act 2002, an 'aggravating factor' that the court must consider to the extent that it is applicable is:

> That the offender committed the offence partly or wholly because of hostility towards a group of persons who have an enduring common characteristic such as race, colour, nationality, religion, gender identity, sexual orientation, age, or disability; and
> (i) the hostility is because of the common characteristic; and
> (ii) the offender believed that the victim has that characteristic (s9(1)(h).

'*Hate speech*' legislation, on the other hand, focuses not on hatred as *motivation* but on hatred as a possible *effect* of certain forms of speech:

> Many statutory definitions of what we call hate speech make the element of 'hatred' relevant as an aim or purpose, something that people are trying to bring about or incite. For example, the Canadian formulation … refers to the actions of a person 'who, by communicating statements in any public place, *incites* hatred against any identifiable group.' Or it is a matter of foreseeable effect, whether intended or not: the British formulation refers to speech that, in all the circumstances, is '*likely to stir up* hatred' (Waldron, 2012, p. 35, emphasis his).

The Royal Commission of Inquiry (2020b, p. 4) noted that:

> Unlike hate crime (such as a hate-motivated assault), conduct criminalised by a hate speech offence—in this case, what has been said—is not usually independently illegal. The difference between legitimately criminalised hate speech and a vigorous exercise of the right to express opinions is not easy to capture—at least with any precision—in legislative language. As well, the more far reaching a law creating hate speech offences, the greater the potential for inconsistency with the right to freedom of expression.

UK-based NGO Article 19 notes that there is no universally accepted definition of the term 'hate speech' in international human rights law (Article 19, 2012, p. 5). In its policy brief, Article 19 focuses on the kind of 'hate speech' described in article 20(2) of the International Covenant on Civil and Political Rights (ICCPR), where it is defined as 'advocacy of hatred on prohibited grounds that constitutes incitement to discrimination, hostility or violence'.[7]

A report published in the UNESCO Series on Internet Freedom notes that:

> In national and international legislation, hate speech refers to expressions that advocate incitement to harm (particularly, discrimination, hostility or violence) based upon the target's being identified with a certain social or demographic group. It may include, but is not limited to, speech that advocates, threatens, or encourages violent acts. For some, however, the

[7] Building on the Camden Principles on Freedom of Expression and Equality, Article 19 recommends the following definitions of key terms in ICCPR article 20(2) and the International Covenant on the Elimination of all Forms of Racial Discrimination article 4(a).

- 'Hatred' is a state of mind characterised as 'intense and irrational emotions of opprobrium, enmity and detestation towards the target group.'
- 'Discrimination' shall be understood as any distinction, exclusion, restriction or preference based on race, gender, ethnicity, religion or belief, disability, age, sexual orientation, language[,] political or other opinion, national or social origin, nationality, property, birth or other status, [or] colour which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights and fundamental freedoms in the political, economic, social, cultural or any other field of public life.
- 'Violence' shall be understood as the intentional use of physical force or power against another person, or against a group or community that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation.
- 'Hostility' shall be understood as a manifested action of an extreme state of mind. Although the term implies a state of mind, an action is required. Hence, hostility can be defined as the manifestation of hatred—that is the manifestation of 'intense and irrational emotions of opprobrium enmity and detestation towards the target group' (Article 19, 2012, p. 19).

On international human rights law and regulation of 'hate speech' and terrorist and violent extremist content, see further Working paper 21/05, **Regulating harmful communication: Current legal frameworks**.

> concept extends also to expressions that foster a climate of prejudice and intolerance on the assumption that this may fuel targeted discrimination, hostility and violent attacks (Gargliadoni, Gal, Alves, & Martinez, 2015).

That is, there is broad agreement that 'hate speech' is public communication that incites discrimination, hostility or violence against members of a social group with a common 'protected characteristic' such as nationality, race or religion (Erjavec & Kovačič, 2012, p. 900).

The target is *a social group*,[8] or an individual based on their actual or supposed membership of a social group, rather than an individual per se. As Herz & Molnar (2012, p. 3) explain, 'telling an ex-lover "I hate you" might be an expression of hate, but it is not "hate speech".' We therefore need to exercise caution in labelling personal criticism or insult as 'hate speech'.[9]

While 'hate crime' and 'hate speech' are distinct concepts, there is, however, some evidence of a link between them (Williams, Burnap, Javed, Liu, & Ozalp, 2020; Berentson-Shaw & Elliott, 2019; Mills, Freilich, & Chermak, 2017). UN Special Rapporteur on minority issues, Rita Izsák, has reported that: 'Although not all hateful messages result in actual hate crimes, hate crimes rarely occur without prior stigmatization and dehumanization of targeted groups and incitement to hate incidents fuelled by religious or racial bias' (UN General Assembly, 2015, para. 26).

The term 'hate speech' is both imprecise and misleading

'H*ate speech'* is 'a less precise term' than 'hate crime' (Royal Commission of Inquiry, 2020a, p. 700). Both the word 'hate' and the word 'speech' can mislead when we are considering regulatory and non-regulatory responses to harmful communication.

It's about harm, not hate

First, calling it '*hate* speech' implies that the problem is the subjective emotion of hatred, as if governments can and should legislate against people feeling certain emotions. To feel animosity towards an individual or social group is not and should not ever be a crime. In a democracy, government should neither prescribe nor proscribe what citizens feel, think, believe or value. And as Robert Post (2009, p. 124) points out, while hatred is an extreme and troublesome human emotion, it can also serve constructive social purposes.

Further, as Bhikhu Parekh explains:

> Hate speech is often expressed in offensive, angry, abusive, and insulting language, and its impact generally depends on that, but it is not necessary that it should be so expressed. Hate speech can also be subtle, moderate, nonemotive, even bland; its message conveyed through ambiguous jokes, innuendoes, and images (Parekh, 2012, p. 41; cf. Kunst, Porten-Cheé, Emmer & Eilders, 2021, p. 3).

[8] A 'social group' comprises a number of people who interact with one another, share some common interests and a common identity, and display some degree of social cohesion.

[9] On the distinction between public and private communication, see further Working paper 21/07, **Striking a fair balance when regulating harmful communication**.

The issue at stake when considering regulatory and non-regulatory responses to harmful communication is not the *emotion* (hate) but the *effect* (harm)—public expression that stirs up and incites discrimination, hostility or violence.

Waldron (2012, p. 8) notes that this is the regulatory approach currently pursued, for example, in Canada, the UK, Denmark, Germany—and New Zealand.

> **Canada**
> Every one who, by communicating statements in any public place, incites hatred against any identifiable group where such incitement is likely to lead to a breach of the peace is guilty of
> (a) an indictable offence and is liable to imprisonment for a term not exceeding two years; or
> (b) an offence punishable on summary conviction (Criminal Code 1985, s319(1)).
>
> **United Kingdom**
> A person who uses threatening, abusive or insulting words or behaviour, or displays any written material which is threatening, abusive or insulting, is guilty of an offence if—
> (a) he intends thereby to stir up racial hatred, or
> (b) having regard to all the circumstances racial hatred is likely to be stirred up thereby (Public Order Act 1986, s18(1)).
>
> **Denmark**
> Any person who, publicly or with the intention of wider dissemination, makes a statement or imparts other information by which a group of people are threatened, insulted or degraded on account of their race, colour, national or ethnic origin, religion, or sexual inclination shall be liable to a fine or to imprisonment for any term not exceeding two years (Criminal Code s266(b)(1)).
>
> **Germany**
> Whoever, in a manner which is suitable for causing a disturbance of the public peace, 1. incites hatred against a national, racial, religious group or a group defined by their ethnic origin, against sections of the population or individuals on account of their belonging to one of the aforementioned groups or sections of the population, or calls for violent or arbitrary measures against them or 2. violates the human dignity of others by insulting, maliciously maligning or defaming one of the aforementioned groups, sections of the population or individuals on account of their belonging to one of the aforementioned groups or sections of the population incurs a penalty of imprisonment for a term of between three months and five years (Criminal Code, s130(1)).

Current New Zealand legislation similarly focuses on 'matter or words likely to excite hostility against or bring into contempt any group of persons in or who may be coming to New Zealand on the ground of the colour, race, or ethnic or national origins of that group of persons' (Human Rights Act 1993, s61(1)), and 'intent to excite hostility or ill-will against, or bring into contempt or ridicule, any group of persons in New Zealand on the ground of the colour, race, or ethnic or national origins of that group of persons' (s131).[10]

And it's not only about speech

Secondly, the term 'hate speech' is misleading because 'harmful communication' involves more than *speech*. Any form of public communication can 'stir up' and incite discrimination, hostility and

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

violence, whether spoken (speech), written, mimed, memed, graffitied, cartooned or tweeted. Waldron draws attention to:

> … expressions that become a permanent or semipermanent part of the visible environment in which our lives, and the lives of members of vulnerable minorities, have to be lived. No doubt a speech can resonate long after the spoken word has died away … but to my mind, it is the enduring presence of the published word or the posted image that is particularly worrying in this connection; and this is where the debate about 'hate speech' regulation should be focused (Waldron, 2012, pp. 37–38).

For these reasons, like Nadine Strossen (2018) and many other commentators, I put quotation marks around 'hate speech' (and 'hate crime'). Some preferred terms are 'harmful speech', 'dangerous speech/expression', 'extreme speech', 'fear speech' or, as I propose, 'harmful communication'.[11]

Another reason to avoid the term 'hate speech' is because it is increasingly being used in popular discourse to demonise a wide array of disfavoured views:

> Many people have hurled the epithet 'hate speech' against a diverse range of messages that they reject, including messages about many important public policy issues. Moreover, too much rhetoric equates 'hate speech' with violent criminal conduct. On many campuses, for example, students complain that they have been 'assaulted' when they are exposed to ideas that offend them, or even if they learn that a provocative speaker has been invited to campus. This false equation between controversial ideas and physical violence fuels unwarranted calls for outlawing and punishing ideas, along with violence (Strossen, 2018, pp. 1–2).

Christopher Ferguson (2020) analyses what he calls 'the Mourner's Veto' and provides some useful suggestions for responding to it. By 'the Mourner's Veto'[12] he means emotional attempts to suppress controversial or unpopular speech:

> Individuals will say that a speaker or a piece of writing has caused them to become distressed or sad or angry or frightened, and they will support these claims with allegations of 'harm' or even threats to their 'right to exist.' Reasonable debate and discussion then becomes impossible as activists make unfalsifiable but furiously emotive claims about alleged threats to their safety and wellbeing amid much weeping and claims of exhaustion and mental fragility. It is not healthy for the limits of permissible speech to be dictated by the most sensitive person in the room, nor to allow emotional appeals to supplant robust argument as the most effective strategy in a debate (Ferguson, 2020).

[11] In New Zealand, referring to 'harmful communication' rather than 'hate speech' does, however, risk confusion with 'harmful digital communication' terminology commonly used in relation to the Harmful Digital Communications Act 2015. That Act defines 'harm' as 'serious emotional distress' and is primarily aimed at providing a pathway for individuals subject to direct bullying or harassment to seek a mediated resolution. These papers are concerned with a much broader set of 'harmful communications' or 'harmful digital communications' than those envisaged by the Harmful Digital Communications Act, and this should be borne in mind by New Zealand-based readers. See further Working paper 21/05, **Regulating harmful communication: Current legal frameworks**.

[12] See also the section on the 'heckler's' or 'thug's veto' in Working paper 21/07, **Striking a fair balance when regulating harmful communication**.

When defining 'hate speech' (harmful communication), it is critical, therefore, to distinguish, in a manner that can be regulated and enforced, communication that:

- Expresses or advocates views but does not call for action;
- Is abusive or insulting but not threatening;
- Expresses dislike of a group but does not incite discrimination, hostility or violence against them;
- Is subtle and not obviously abusive or insulting; and
- Takes a demeaning or denigrating view of a group but wishes it no harm (Parekh, 2012, p. 40).

Parekh (2012, p. 53) does argue for regulation of 'hate speech', but stresses that the relevant concepts must be 'defined with great care and distinguished from such vague expressions as offensive, hurtful, and distressing remarks'.[13]

The harm principle and the presumption of liberty

In both international human rights law and in New Zealand domestic law, the presumption is freedom of opinion and expression. Any limitation of this freedom by the state should be subject 'only to such reasonable limits prescribed by law as can be demonstrably justified in a free and democratic society' (New Zealand Bill of Rights Act 1990, s5).[14]

The 'harm principle', and other liberty-limiting principles that extend it in various ways, can inform deliberation on whether regulation to restrict freedom of expression is reasonable and 'demonstrably justifiable' (Bromell & Shanks, 2021; Bromell, 2019, pp. 76–84; Feinberg, 1973, 1980).

The harm principle holds that *restricting freedom may be justifiable if (and only if) the intervention prevents harm to specified others (private harm) or unspecified others (public harm)*.

The *private harm principle* may justify a state enacting laws, for example, that prohibit and punish burglary, assault, child sexual abuse, rape, manslaughter and homicide.

The *public harm principle* may justify restricting a person's freedom to prevent public harms, which are of two main sorts:

- behaviours that risk significant harm to unspecified others; for example, driving while under the influence of drugs and/or alcohol, discharging a weapon in a public place, or selling a product known to be unsafe; and
- behaviours that risk significant harm to public institutions and practices; for example, tax evasion, welfare benefit fraud, refusing to perform jury service, counterfeiting currency, or smuggling.

Committing a criminal offence associated with a motivation and/or demonstration of hostility to an individual as a member of a social group with a common 'protected characteristic' (a 'hate crime') may constitute both a private harm (damage to the person and/or property of specified others) and

[13] See further Working paper 21/07, **Striking a fair balance when regulating harmful communication**.

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

a public harm (the risk that the behaviour will become general and impact on numerous unspecified others or otherwise be 'injurious to the public good'[15]).

Public communication that incites discrimination, hostility and violence against a social group with a common 'protected characteristic' ('hate speech'/harmful communication) constitutes a public harm to unspecified others and may occasion private harm to specified others if acted upon—harm being understood as discrimination, hostility and violence.[16]

An assessment of 'harm' is relevant, therefore, in weighing up whether a regulatory proposal to restrict freedom of expression is reasonable and demonstrably justifiable.

Of course, a public policy proposal may be justifiable without necessarily being justified. Whether or not it is justified may only become clear through a review and appeal process and/or the settled agreement of the public over time.

Summary of definitions

A **'hate crime'** involves the commission of a criminal offence, for example assault and injury to another person, or damage to property, associated with a motivation and/or demonstration of hostility to the victim as a member of a social group with a common 'protected characteristic' such as nationality, race or religion.

'Hate speech' (better, 'harmful communication') is public communication that incites discrimination, hostility or violence against members of a social group with a common 'protected characteristic' such as nationality, race or religion.

Conclusion: Keep the focus on harm, not hate

We need to maintain clear distinctions between 'hate speech' and 'hate crime', and between 'hate speech' and criticism, insult and 'hurtful' remarks that cause offence.

A democratic state can justifiably use its coercive powers to protect its citizens from harmful public communication that incites discrimination, hostility or violence against them based on their actual or supposed membership of a social group with a common 'protected characteristic'.

A democratic state cannot justifiably restrict freedom of opinion and expression by prohibiting criticism, satire, offensive or 'hurtful' comments, disapproval, dislike—or even hatred.

This keeps the focus, not on the *emotion* of 'hate', but on the *effect* of harm (discrimination, hostility or violence). 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.

The remaining five working papers in this series develop this argument further and elaborate on challenges in regulating online content (Working paper 21/04), current legal frameworks for

[15] Cf. Films, Videos, and Publications Classification Act 1993 s3(1), and see the discussion in Bromell & Shanks, 2021, pp. 46–47.

[16] On the distinction between harm and offence, see further Working paper 21/07, **Striking a fair balance when regulating harmful communication**, pp. 8–10.

regulating harmful communication (Working paper 21/05), arguments for and against restricting freedom of expression (Working paper 21/06), striking a fair balance when regulating harmful communication (Working paper 21/07), counter-speech as an alternative or complement to prohibition and censorship, and civility as everyone's responsibility (Working paper 21/08).

References

Adams, G. (2021). Will Ardern back away from new 'hate speech' laws? January 29, 2021. Accessed January 29, 2021, from https://democracyproject.nz/2021/01/29/graham-adams-will-ardern-back-away-from-new-hate-speech-laws/

Ardern, J. (2020). Prime Minister's comments on Royal Commission of Inquiry into Christchurch Terror Attack. Media release, December 8, 2020. Accessed December 8, 2020, from https://www.beehive.govt.nz/speech/prime-minister%E2%80%99s-comments-royal-commission-inquiry-christchurch-terror-attack

Article 19. (2012). *Prohibiting incitement to discrimination, hostility or violence: Policy brief.* London: Article 19. Accessed November 2, 2020, from https://www.article19.org/data/files/medialibrary/3548/ARTICLE-19-policy-on-prohibition-to-incitement.pdf

Berentson-Shaw, J., & Elliott, M. (2019). *Online hate and offline harm.* The Workshop, May 2019. Accessed December 10, 2020, from https://www.theworkshop.org.nz/publications/online-hate-and-offline-harm-2019

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

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

Devlin, C. (2020a). Justice Minister forges ahead with hate speech laws for New Zealand. *Stuff,* March 13, 2020. Accessed October 15, 2020, from https://www.stuff.co.nz/national/politics/120264595/justice-minister-forges-ahead-with-hate-speech-laws-for-new-zealand

Devlin, C. (2020b). Hate speech law stalled until after election—no support yet from NZ First. *Stuff,* June 24, 2020. Accessed October 15, 2020, from https://www.stuff.co.nz/national/politics/121922974/hate-speech-law-stalled-until-after-election--no-support-yet-from-nz-first

Duff, M. (2019). Hate crime law review fast-tracked following Christchurch mosque shootings. *Stuff*, March 30, 2019. Accessed October 15, 2020, from https://www.stuff.co.nz/national/christchurch-shooting/111661809/hate-crime-law-review-fasttracked-following-christchurch-mosque-shootings

Ensor, J. (2020). Hate-motivated crime data collection being strengthened as Muslim leaders demand action. *Newshub,* March 15, 2020. Accessed October 19, 2020, from https://www.newshub.co.nz/home/new-zealand/2020/03/hate-motivated-crime-data-collection-being-strengthened-as-muslim-leaders-demand-action.html

Erjavec, K., & Kovačič, M. (2012). 'You don't understand, This is a new war!': Analysis of hate speech in news web sites' comments. *Mass Communication & Society, 15*(6), 899–920. https://doi.org/10.1080/15205436.2011.619679

Ernst, J., Schmitt, J., Rieger, D., Beier, A., Vorderer, P., Bente, G., & Roth, H.-J. (2017). Hate beneath the counter speech? A qualitative content analysis of user comments on YouTube related to counter speech videos. *Journal for Deradicalization, 10*, 1–49. Accessed November 26, 2020, from https://journals.sfu.ca/jd/index.php/jd/article/view/91

eSafety Commissioner. (2020). *Online hate speech: Findings from Australia, New Zealand and Europe*, February 2, 2020. Accessed October 19, 2020, from https://www.esafety.gov.au/about-us/research/online-hate-speech

Faafoi, C. (2020). Making New Zealand safer for everyone. Joint media statement by Hons P. Williams, C. Faafoi & P. Radhakrishnan, December 8, 2020. Accessed December 8, 2020, from https://www.beehive.govt.nz/release/making-new-zealand-safer-everyone

Feinberg, J. (1973). *Social philosophy*. Englewood Cliffs, NJ: Prentice-Hall.

Feinberg, J. (1980). *Rights, justice and the bounds of liberty: Essays in social philosophy.* Princeton, NJ: Princeton University Press.

Ferguson, C. (2020). Resisting the Mourner's Veto. *Quillette,* December 3, 2020. Accessed February 15, 2021, from https://quillette.com/2020/12/03/resisting-the-mourners-veto/

FIANZ. (2020). FIANZ submission to the Royal Commission of Inquiry into the Attack on Christchurch Mosques. Federation of Islamic Associations of New Zealand Inc., February 2020. Accessed December 14, 2020, from https://fianz.com/christchurch-mosques-attach/

FIANZ. (2021). *The engagement process: Submission to Hon. Andrew Little, Lead Coordination Minister for the Government's response to the Royal Commission's report into the terrorist attack on the Christchurch mosques*, February 2021. Accessed March 2, 2021, from https://assets.documentcloud.org/documents/20493407/fianz-hui-report-march-2021.pdf

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

Geschke, D., Klaßen, A., Quent, M., & Richter, C. (2019). *Hass im Netz: Eine bundesweite repräsentative Untersuchung*. Institut für Demokratie und Zivilgesellschaft, June 2019. Accessed November 24, 2020, from https://www.idz-jena.de/forschung/hass-im-netz-eine-bundesweite-repraesentative-untersuchung-2019/

Hansard. (2020). Parliamentary Debates (Hansard) for Tuesday, 8 December 2020. House of Representatives, 53/749, corrected. Accessed January 29, 2021, from https://www.parliament.nz/en/pb/hansard-debates/rhr/combined/HansD_20201208_20201208

Herz, M., & Molnar, P. (2012). Introduction. In M. Herz & P. Molnar (Eds), *The content and context of hate speech: Rethinking regulation and responses* (pp. 1–7). Cambridge: Cambridge University Press.

Human Rights Commission. (2019a). *It happened here: Reports of race and religious hate crime in New Zealand 2004–2012.* Accessed October 16, 2020, from https://www.hrc.co.nz/news/commission-issues-2004-2012-hate-crime-summary/

Human Rights Commission. (2019b). *Whakamauāhara Hate Speech. An overview of the current legal framework*. December 17, 2019. Accessed October 15, 2020, from https://www.hrc.co.nz/news/resource-hate-speech-legal-framework-published/

Kenny, K. (2020a). Hate speech up since Christchurch terror attack; Government considers law changes. *Stuff*, January 27, 2020. Accessed October 16, 2020, from https://www.stuff.co.nz/national/christchurch-shooting/119028862/hate-speech-up-since-christchurch-terror-attack-government-considers-law-changes

Kenny, K. (2020b). Police say they're working to improve monitoring of hate crimes in New Zealand. *Stuff*, February 14, 2020. Accessed October 16, 2020, from https://www.stuff.co.nz/national/119506044/police-say-theyre-working-to-improve-monitoring-of-hate-crimes-in-new-zealand

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

Landesanstalt für Medien NRW. (n.d.). Forsa-Befragung zur Wahrnehmung von Hassrede. Accessed November 24, 2020, from https://www.medienanstalt-nrw.de/themen/hass/forsa-befragung-zur-wahrnehmung-von-hassrede.html

Little, A. (2019). Hate speech threatens our right to freedom of speech. *NZ Herald,* April 28, 2019. Accessed October 30, 2020, from https://www.nzherald.co.nz/nz/andrew-little-hate-speech-threatens-our-right-to-freedom-of-speech/2II6E7A5AZHQRG2HM4ZQXNFA4M/

Mills, C., Freilich, J., & Chermak, S. (2017). Extreme hatred: Revisiting the hate crime and terrorism relationship to determine whether they are 'close cousins' or 'distant relatives'. *Crime & Delinquency, 63*(10), 1191–1123. https://doi.org/10.1177/0011128715620626

Netsafe. (2019). *2019 online hate speech insights*, December 12, 2019. Accessed October 16, 2020, from https://www.netsafe.org.nz/2019-online-hate-speech-insights/

NZ Ministry of Justice. (2020). *The New Zealand Crime and Victims Survey. Key findings, Cycle 2, October 2018 to September 2019.* Accessed December 10, 2020, from https://www.justice.govt.nz/assets/Documents/Publications/NZCVS-Y2-A5-KeyFindings-v2.0-.pdf

Parekh, B. (2012). Is there a case for banning hate speech? In M. Herz & P. Molnar (Eds), *The content and context of hate speech: Rethinking regulation and responses* (pp. 37–56). Cambridge: Cambridge University Press.

Post, R. (2009). Hate speech. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 123–138). Oxford: Oxford University Press.

RNZ. (2020). Islamic Women's Council say police and spy agencies 'failed' to protect Muslims. March 9, 2020. Accessed October 15, 2020, from https://www.stuff.co.nz/national/120132938/islamic-womens-council-say-police-and-spy-agencies-failed-to-protect-muslims

Royal Commission of Inquiry. (2020a). *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/

Royal Commission of Inquiry. (2020b). *Hate speech and hate crime related legislation.* [Companion paper to 2020a]. Accessed December 10, 2020, from https://christchurchattack.royalcommission.nz/publications/comp/introduction/

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

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

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

Williams, M., Burnap, P., Javed, A, Liu, H., & Ozalp, S. (2020). Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *British Journal of Criminology, 60*(1), 93–117. https://doi.org/10.1093/bjc/azz049