After Christchurch:
Hate, harm and the limits of censorship

5. Arguments for and against restricting freedom of expression

David Bromell

Working Paper 21/06

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
21/06

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).

This is paper five 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

Arguments for and against restricting freedom of expression

Abstract

Public policy decisions about whether, when and how to regulate harmful communication in a free, open and democratic society necessarily involve values and moral principles. Political philosophy can shed light on these, to inform and guide decision making about the right thing to do, or not do.

In light of the terrorist attack on Christchurch mosques on March 15, 2019 and the subsequent Christchurch Call to eliminate terrorist and violent extremist content online (Working paper 21/02), this paper presents arguments both for and against the state restricting freedom of expression. It builds on previous working papers that define the problem of 'hate speech' and other key terms (Working paper 21/03), identify challenges in regulating online content (Working paper 21/04), and summarise current legal frameworks for regulating harmful communication (Working paper 21/05).

The paper analyses arguments for restricting freedom of expression (whether offline or online) in terms of the respective interests of 'protagonists', 'antagonists' and 'the audience' (or society). The proposed 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, **Striking a fair balance in regulation of harmful communication**, introduces further considerations to inform prudential balancing of freedom of expression, protection from harm, promotion of social cohesion, maintenance of public order and ensuring that the law can practically be enforced.

Working paper 21/08, **Counter-speech and civility as everyone's responsibility**, discusses use of the state's expressive and not only coercive powers, and counter-speech strategies as alternatives or complements to prohibition and censorship.

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

Introduction

Public policy making in a free, open and democratic society involves incremental social problem solving as we work out the right thing to do, or not do, through agreed processes. It demands prudential balancing of divergent, competing and conflicting interests, exercising principled pragmatism in local contexts (Bromell, 2019, p. 180).

Should the state ban or restrict harmful communication, or is it protected by the right to freedom of opinion and expression? Jeffrey Howard (2018, p. 20) argues that 'this is, at root, a question of political philosophy, which requires that we reflect on the moral principles that should guide the decisions of citizens and policy-makers.'

This paper summarises arguments both for and against restricting freedom of expression and discusses some principles expressed or implied by these arguments. While many of these arguments have a significant history that pre-dates the internet and opportunities for harmful digital communication, they continue to be relevant to decision making about whether and how to regulate any public communication, offline or online, that stirs up or incites discrimination, hostility or violence.[1]

Arguments for restricting freedom of expression

Thomas Scanlon (1979, pp. 520–528) considers the individual interests with which freedom of expression is concerned: the interests we have in communicating ('participant interests'), the interests we have in being exposed to what others have to say ('audience interests'), and the interests we have as bystanders who are affected by expression in other ways ('bystander interests').

Scanlon's proposal to consider different interests is useful. An 'interest' denotes a relationship between a *subject* (some specified individual or group of individuals) and a substantive *object* (an action, event, process or continuous state) that affects the situation, needs, wants, beliefs, aspirations, values, emotions and preferences of the subject over time (Bromell, 2019, p. 54). Our interests include our values, norms, ideas and ideals.

Our interests are comparative. An 'interest' is always *someone's* interest in *something* that puts that someone in a better position over time to get what they want or value, *compared to something else* (cf. Barry, 1964, p. 4; 1965, pp. 175–186).

Each of us has individual interests, but we may hold them in common with others in an 'interest group' in ways that shape our social identities and belongings. Individuals have plural interests and are likely to affiliate with more than one interest group. Within an interest group, and within a society, we may hold some interests in common; others will be different, competing or conflicting. Your interests are not necessarily the same as my interests (Bromell, 2019, pp. 27–35, 54–62).[2]

A not-so-hypothetical case

A woman is walking her young son to school. She wears a hijab / headscarf. A man yells at her from the other side of the street: 'Hey, raghead! Go back to where you came from!'

Traffic on the street separates the man from the woman and her son; people are walking on the footpath on both sides of the street, including other parents who are also walking their children to school.

Although spoken in a public place, the man's comments are addressed to the woman, not to anyone else who is on the street at the time, or to any other third party (for example, by livestreaming on social media).

If we were observing this incident from the outside looking in, how might we react and reflect on it?

[1] On particular challenges in regulating online content, see Working paper 21/04.

[2] On liberties and interests, see also Michelman (1992).

The woman and her son are likely to feel frightened and intimidated by the confrontation. Communicating that they do not belong here, even if it is just random 'mouthing off', is a public attack on the dignity and equal citizenship of the woman and her son, and as such it is a reprehensible abuse of freedom of expression. Perhaps it will prompt the woman to retreat to private space and withdraw from active participation in and contribution to public life. It may impact on the boy's self-image and confidence, and his engagement and achievement in education.

The man who shouts the abuse has exercised his right to freedom of expression and vented his private feelings publicly. Does it make a difference that he is a man shouting at a woman, whether he is young or old, what his own ethnic background is, where he fits in the socio-economic scale of things, or if he is mentally unwell or under the influence of drugs or alcohol?

While it would be difficult to prove beyond reasonable doubt in a court of law that his shouted comments intended to incite or were likely to incite others to discrimination, hostility or violence, his comments could reasonably be taken as intended to frighten or intimidate the woman, which is a summary offence in New Zealand law liable to imprisonment for a term not exceeding three months or a fine not exceeding $2,000.[3]

Members of the public who witness the man's outburst may variously feel protective towards the woman and her child ('Are you OK, and would you like me to walk the rest of the way with you?), sympathetic to the man's outburst ('I wouldn't say it, but he's not wrong'), intimidated ('If it's her turn today, is it my turn tomorrow?), frightened ('Should I intervene, but will I get attacked for it?'), angry ('How dare he!' / 'How dare you!') or ashamed ('I never thought I'd hear that in our town!').

All the world's a stage

Using the metaphor of the theatre, we can adapt Scanlon's analysis by thinking of the different interests that play out in time and space as the interests of the protagonist (for example, the woman walking with her son), the antagonist (for example, the man shouting abuse), and the audience (for example, members of the public who witness the incident).[4] When thinking about restricting

[3] Summary Offences Act 1981, s21(1):

> (1) Every person commits an offence who, with intent to frighten or intimidate any other person, or knowing that his or her conduct is likely to cause that other person reasonably to be frightened or intimidated,—
> (a) threatens to injure that other person or any member of his or her family, or to damage any of that person's property; or
> (b) follows that other person; or
> (c) hides any property owned or used by that other person or deprives that person of, or hinders that person in the use of, that property; or
> (d) watches or loiters near the house or other place, or the approach to the house or other place, where that other person lives, or works, or carries on business, or happens to be; or
> (e) stops, confronts, or accosts that other person in any public place.

[4] I am thinking here of ancient Greek tragedy as a metaphor for interests at play in harmful communication but see also Erving Goffman's (1959) dramaturgical approach to sociology, which uses theatre as a metaphor to represent how people present themselves to one another based on cultural values, norms and beliefs, playing various roles and strategies of concealment and exposure. Also relevant is Stephen Karpman's (1968, 1973) 'drama triangle' social model of human interaction, and 'bystander theory' (Latané & Darley, 1970), which suggests that the more an incident appears to be an emergency, the more likely individuals are to help the victims (Kunst, Porten-Cheé, Emmer & Eilders, 2021, p. 3; Leonard, Rueß, Obermaier, & Reinemann, 2018).

freedom of expression to regulate harmful communication, we are concerned with what happens 'onstage', in public space, rather than with whatever may happen 'backstage' or 'offstage' in private space.[5]

The interests of the protagonist

The protagonist is at the centre of the story—a character or group of characters trying to go about their lives despite opposition and obstacles put in their way by the antagonist.

Harmful communication that stirs up discrimination, hostility or violence against protagonists because of their actual or supposed social group identity impacts on what the Universal Declaration of Human Rights, Article 3, defines as the right to 'life, liberty and security of person' and 'all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status' (Article 2) (UN General Assembly, 1948).

As Jeremy Waldron (2012a, p. 59) argues, harmful expression is an assault first and foremost on the *dignity* of the individuals affected, 'in the sense of their basic social standing, the basis of their recognition as social equals and as bearers of human rights and constitutional entitlements'.

By 'dignity', Waldron does not mean 'a sort of shimmering Kantian aura' (2012a, p. 219), but rather the ability to go about everyday life as an ordinary member of society entitled to the same liberties, protections and powers that everyone else has:

> The guarantee of dignity is what enables a person to walk down the street without fear of insult or humiliation, to find the shops and exchanges open to him, and to proceed with an implicit assurance of being able to interact with others without being treated as a pariah (Waldron, 2012a, p. 220).

Harmful communication promotes and perpetuates negative stereotypes that fuel and seek to justify discrimination, hostility or violence. It is a form of group defamation that denigrates 'that basic standing and some characteristic associated more or less ascriptively[6] with all members of the group' (Waldron, 2012a, p. 57).

Waldron insists, however, in ways that align with international human rights standards,[7] that:

> … though we are talking about group dignity, our point of reference is the individual members of the group, not the dignity of the group as such or the dignity of the culture or social structure that holds the group together. The ultimate concern is what happens to individuals when defamatory imputations are associated with shared characteristics such as race, ethnicity, religion, gender, sexuality, and national origin (Waldron, 2012a, p. 60).

---

[5] On the distinction between public and private communication, see Working paper 21/07, **Striking a fair balance when regulating harmful communication**, and Working paper 21/03, **'Hate speech': Defining the problem and some key terms**.

[6] Waldron means here the practice of 'ascribing' or designating the status of a social group and anyone who belongs to it on the basis of some factor like age, sex, or race, rather than on the individual characteristics or achievements of members of the group.

[7] See Working paper 21/07, **Striking a fair balance when regulating harmful communication**, on Persons and groups.

The interests of the antagonist

In a free, open and democratic society, the state has no business either prescribing or proscribing what citizens are to feel, think, believe or value. Neither can the state justifiably require its citizens to like, approve of or refrain from criticising a religion, culture, belief, set of values or way of life.

The state can, however, justifiably require citizens to observe certain limits to freedom of expression—refraining from harmful communication in public space that incites or is likely to incite discrimination, hostility or violence against a social group with a common 'protected characteristic' such as nationality, race or religion.

Many people who engage in harmful communication are angry, and this anger is often driven by a feeling that something important to them and to their social group is threatened or at risk (Schmitt, 2017, p. 53).

Feeling angry is not and should not be a crime. Neither does anger equate to hate, fear or intent to cause harm. If an antagonist criticises a protagonist's religion, culture, beliefs, values or way of life, and expresses strong feelings about this, it does not necessarily mean that the antagonist 'hates' the protagonist, has a 'phobia' about them and what they stand for, or is seeking to incite discrimination, hostility or violence against them. Antagonists are well within their rights to argue strenuously and with feeling for their own values, beliefs and way of life, and against others' values, beliefs and ways of life.

Even if an antagonist does feel dislike or hatred of an individual or social group, it is not and should not be a crime to feel those emotions. And equally, if a protagonist responds to criticism and opposition by feeling alarmed, offended, angry or resentful, the protagonist does not have a right to retaliate with discrimination, hostility or violence towards the antagonist (Cohen, 1993, p. 253). The state cannot justifiably restrict the antagonist's freedom of expression simply to protect the protagonist against feeling offended (Waldron, 2012a, p. 126):[8]

> Some have said that toleration makes no sense except against a background of strong disagreement. We do not tolerate those of whom we approve or those to whom we are indifferent. We don't tolerate those whom we suspect might have the truth or part of the truth in a pluralistic world. We tolerate those whom we judge wrong, mistaken, or benighted. And surely toleration must permit us to give voice to those judgments. Otherwise it demands too much (Waldron, 2012a, p. 228).[9]

[8] In distinguishing between indignity and offence, Waldron (2012a, pp. 106, 107) explains: 'The distinction is in large part between objective or social aspects of a person's standing in society, on the one hand, and subjective aspects of feeling, including hurt, shock, and anger, on the other … So, to protect people from offense or from being offended is to protect them from a certain sort of effect on their feelings. And that is different from protecting their dignity and the assurance of their decent treatment in society.' This parallels the distinction between 'hate' and 'harm' in preferring 'harmful communication' to 'hate speech' (Working paper 21/03**, 'Hate speech': Defining the problem and some key terms**).

[9] Joseph Raz (1986, p. 401) defines toleration as 'the suppression or containment of an inclination or desire to persecute, harass, harm or react in an unwelcome way to a person.' He elaborates (pp. 401–407) on four features of tolerance: (1) only behaviour that is unwelcome to the person towards whom it is addressed or behaviour that is normally seen as unwelcome is intolerant behaviour; (2) one is tolerant only if one inclines or is tempted not to be; (3) that inclination is based on dislike of or antagonism to the behaviour, character or some feature of the existence of its object; and (4) the intolerant inclination is in itself, at least in the eyes of the person experiencing it, worthwhile or desirable.

In public space, however, the antagonist's right to freedom of expression is *a qualified right*, because freedom is not the only value, and in our life together freedom is in tension with other important values like equality, fairness and community (Bromell, 2019). In weighing freedom of expression against prevention of harm, a society may choose to assign a greater value to communication that contributes to public debate and the formation of public opinion than to an emotional outburst directed, or mis-directed, at an individual person or social group (Gärditz, 2020). As Bhikhu Parekh (2012, p. 49) argues:

> There is no obvious virtue in allowing everyone to say anything that comes into his or her head and requiring those affected to put up with it. Speech can have social consequences that need to be balanced against its benefits; it cannot be absolutized or allowed to override any and all other values.[10]

The interests of the audience

As the drama unfolds between protagonist and antagonist, the audience may reflect that if, despite our differences, we are to live together without recourse to violence, then a public good is at stake:

> Each group must accept that the society is not *just* for them; but it *is* for them too, along with all of the others. And each person, each member of each group, should be able to go about his or her business, with the assurance that there will be no need to face hostility, violence, discrimination, or exclusion by others. When this assurance is conveyed effectively, it is hardly noticeable; it is something on which everyone can rely, like the cleanness of the air they breathe or the quality of the water they drink from a fountain. This sense of security in the space we all inhabit is a public good, and in a good society it is something that we all contribute to and help sustain in an instinctive and almost unnoticeable way (Waldron, 2012a, p. 4).

The audience ('the public') has an interest, not just in preventing discrimination, hostility and violence that disrupt public order, but also in restricting harmful communication from becoming embedded as a feature of our environment that becomes 'world-defining' for all of us (Waldron, 2012a, p. 74).

Cohen (1993, p. 231) analyses this as an *environmental cost* of harmful communication that 'may help to constitute a degraded, sickening, embarrassing, humiliating, obtrusively moralistic, hypercommercialized, hostile, or demeaning environment.' In the same vein, Carolin Emcke (2019, p. xiv) writes:

> I … do not think uninhibited shouting, slandering and insulting represents an advancement of civilization. I do not consider it a sign of progress that every inner baseness may be turned outwards just because exhibiting resentments is now supposed to have some public or even political relevance. Like many other people, I refuse to get used to it. I do not want to see the new, unbridled appetite for hatred becoming normal.

[10] Parekh (2012, p. 50) adds: 'Beyond a certain point, the moral neutrality of the state is itself problematic. A liberal state should not enforce a particular view of the good life on its citizens and should allow a free flow of ideas, but some values are so central to its moral identity that it cannot remain neutral with respect to them.' On state neutrality and perfectionism, see Bromell, 2019, pp. 157–159.

This suggests a political and social objective of balancing rights and responsibilities in ways that create and maintain a civil, well-ordered society:

> No area of our social compact agitates us so much as how we speak to each other in conversation and in the public forum. Here our general and specific values are placed on view for others to see, and here we are able to assess the character of the compact we have with each other to maintain society for both our own good and that of the human race (Asante, 1995).

Harmful communication, whether offline or online, seeks to dominate and exercise power over minority social groups and interfere with their speech rights (Fiss, 1995, p. 287). It tends to silence the 'voice' of those targeted and to discourage them from active participation in and contribution to democratic deliberation and public life (Munger, 2017).[11]

The fundamental issue here is that harmful communication seeks to undermine freedom and equality, opposing the core values of liberal democracy: 'A state is not fully democratic if it formally guarantees rights and democratic procedures, while failing to endorse the underlying values of self-government in its broader culture' (Brettschneider, 2012, p. 15; cf. Waldron, 2012b, p. 336–337).[12] A democratic society requires public discourse between individuals who recognize one another as free and equal persons and members of a community (Heyman, 2009, p. 174) in which all citizens can belong, participate and contribute—economically, socially, culturally and politically.

While there are limits to what laws can achieve, regulation of harmful communication is at least part of the solution: 'It affirms the community's commitment to equality and civility, sets standards of good behavior, reassures vulnerable groups, and prevents the normal intergroup conflicts and prejudices of a multiethnic society from getting out of control' (Parekh, 2012, p.54).

Provided legislation is framed so that it can be effectively enforced, regulation of harmful communication can play an important role in a free, open and democratic society:

> Well-targeted prosecution of hatred and incitement ultimately serves to protect the basic functional conditions of a liberal democracy, which relies on a culture of debate that, despite all differences and disagreements, respects the dignity and recognition claims of all people (Gärditz, 2020).[13]

[11] Eric Heinze (2009a, p. 197) argues, to the contrary, that no serious empirical research has been undertaken to demonstrate a causal link between 'hate speech' and participation in public discourse—either to show that 'hate speech' deters participation, or that 'hate speech' bans help achieve aims of social inclusion.

[12] Corey Brettschneider (2012) argues, however, against both neutrality and prohibitionism in relation to discriminatory or hateful viewpoints, and for a third way (democratic persuasion) that distinguishes between a state's coercive power (its ability to place legal limits on hate speech), and its expressive power (its ability to influence beliefs and behaviour): 'On my view, the state should simultaneously protect hateful viewpoints in its coercive capacity and criticize them in its expressive capacity' (p. 3). See further Working paper 21/07, **Striking a fair balance in regulation of harmful communication**. Implications of this are explored in Working paper 21/08, **Counter-speech and civility as everyone's responsibility**.

[13] My translation of the original: 'Gezielte Strafverfolgung von Hass und Hetze dient dann letztlich dem Schutz basaler Funktionsbedingungen einer freiheitlichen Demokratie, die auch von einer Debattenkultur abhängt, die bei aller Differenz in der Sache die Würde und den Achtungsanspruch aller Menschen respektiert.'

Arguments against restricting freedom of expression

There are seven main arguments against restricting freedom of expression, which concern:

- Individual autonomy;
- Legal responsibility;
- Reason and 'the marketplace of ideas';
- Political legitimacy and representative democracy;
- Restraining the state;
- A dilemma in human rights law; and
- Legal efficacy.

Individual autonomy

The first argument concerns human autonomy, 'the Kantian right of each individual to be treated as an end in himself, an equal sovereign citizen of the kingdom of ends with a right to the greatest liberty compatible with the like liberties of all others' (Fried, 1992, p. 233). John Stuart Mill wrote in his essay *On Liberty* that:

> If all mankind minus one, were of one opinion, and only one person were of the contrary opinion, mankind would be no more justified in silencing that one person, than he, if he had the power, would be justified in silencing mankind (Mill, 1859, p. 33).

Thomas Nagel (2002, p. 45) similarly insists:

> To admit the right of the community to restrict the expressions of convictions or attitudes on the basis of their content alone is to rob everyone of authority over his own mental life. It makes us all, equally, less free.

No one, least of all the state, can justifiably limit the use of our rational powers or tell us what to feel, think, believe or value. As Edwin Baker (2012, p. 64) states the argument:

> Typically racist hate speech embodies the speaker's at least momentary view of the world and, to that extent, expresses her values. Of course, her speech does not respect others' equality or dignity. It is not, however, the speaker but the state's legitimacy that is at stake in evaluating the content of the legal order. Law's purposeful restrictions on such racist or hate speech violate the speaker's formal autonomy, whereas her hate speech does not interfere with or contradict anyone else's formal autonomy even if such speech does cause injuries that sometimes include undermining others' substantive autonomy. For this reason, prohibitions on racist or hate speech should generally be impermissible—even if arguably permissible in special, usually institutionally bound, limited contexts where the speaker has no claimed right to act autonomously—such as when, as an employee, one has given up one's autonomy to meet role demands inconsistent with expressions of racism.

The problem is compounded by the 'intractable vagueness and over-breadth' of 'hate speech' regulation, 'thus necessitating enforcement according to the subjective standards of complainants and enforcing authorities' (Strossen, 2018, p. 13). And while he does present a case for regulation of harmful communication, Waldron reminds us that:

> Defenders of hate speech regulation need to face up honestly to the moral costs of their proposals. Obviously, restrictions of the kind we are considering are designed to stop people from printing, publishing, distributing, and posting things that they would like to say and that they would like others to read or hear … When people speak, they are disclosing important

> aspects of themselves to the world, staking out their own place in a society that consists of millions of distinctive individuals, each defined by his or her principles, values, convictions, and beliefs (Waldron, 2021a, pp. 148, 161).

Human agency and legal responsibility

Autonomy intersects with questions of legal responsibility. Thomas Scanlon argues that if I tell someone they should rob a bank and they subsequently act on this advice, I am not legally responsible for that person's act and neither could my advice legitimately be made a separate crime: 'A person who acts on reasons he has acquired from another's act of expression acts on what *he* has come to believe and has judged to be a sufficient basis for action' (Scanlon, 1972, p. 212, emphasis his).

Scanlon's argument has not altogether stood up to subsequent criticism (Barendt, 2009; Amdur, 1980) and he himself no longer endorses the 'Millian Principle' he argued for in the 1972 article (Scanlon, 1979). The person who inspires someone else to rob a bank bears moral responsibility for the robbery— and perhaps legal responsibility ought to reflect moral responsibility (Amdur, 1980, p. 297). But there is still some merit in Scanlon's argument that a crime of incitement 'has to be something more than merely the communication of persuasive reasons for action' (Scanlon, 1972, p. 212).

Seana Shiffrin (2003, pp. 1136–1137) elaborates on this question:

> A speaker may advocate illegal action even though this may inspire some audience members to perform illegal actions. So long as the incendiary speech does not meet the *Brandenburg v. Ohio* standard[14] (that is, it is not likely *and* not intended to incite or produce imminent lawless action), the speaker may not be prevented from giving this speech or criminally penalized for it afterward, even if it is quite foreseeable that some member of the audience will at some point be persuaded by the advocacy and proceed to break the law or commit violence. In this circumstance, we hold the noncompliant agent solely legally responsible for the harm. It is his legal responsibility not to spring to harmful action on account of the speech. The speaker herself does no legal wrong; merely advocating wrong is not itself a legal wrong. The speaker's liberty interest is not overcome or superseded by the interest we have in preventing actors from harming others.

Shiffrin notes (p. 1156) the risk of exaggerating 'the moral distinction between intended and merely foreseeable consequences' and overestimating 'the distance between the degree of responsibility associated with intended, as opposed to merely foreseen, consequences.' She nevertheless maintains that:

> The primary motivation here is not one about causation. That is, speaker liability is rejected not because there is a sense that the speaker does not directly or indirectly cause the harm. Rather, speaker liability is rejected primarily because the point of the activity and our protection of it depend on a separation of responsibility between speakers and audiences (Shiffrin, 2003, p. 1161).

[14] *Brandenburg v. Ohio* was a landmark decision (1969) of the US Supreme Court, holding that government cannot punish inflammatory speech—in this case at a Ku Klux Klan rally in Hamilton County in rural Ohio in 1964—unless that speech is 'directed to inciting or producing imminent lawless action and is likely to incite or produce such action'.

Kenan Malik (in Molnar, 2012a, p. 85), arguing that lines should not be blurred between human agency and responsibility, or between attitudes and actions, notes, however, that:

> … there are clearly circumstances in which there is a direct connection between speech and action, where someone's words have directly led to someone else taking action. Such incitement should be illegal, but it has to be tightly defined. There has to be both a direct link between speech and action and intent on the part of the speaker for that particular act of violence to be carried out.

Reason and 'the marketplace of ideas'

A third argument against restricting freedom of expression focuses on 'the marketplace of ideas' (Fraleigh & Tuman, 2011) in which even dangerous ideas are free to be expressed—and defeated in argument.[15] Thomas Jefferson wrote in a letter about 'the illimitable freedom of the human mind, for here we are not afraid to follow truth wherever it may lead, nor to tolerate any error so long as reason is left free to combat it' (Jefferson, 1820).

John Stuart Mill argued in his essay *On Liberty*:

> But the peculiar evil of silencing the expression of an opinion is, that it is robbing the human race; posterity as well as the existing generation; those who dissent from the opinion, still more than those who hold it. If the opinion is right, they are deprived of the opportunity of exchanging error for truth: if wrong, they lose, what is almost as great a benefit, the clearer perception and livelier impression of truth, produced by its collision with error (Mill, 1859, p. 33).

As Joshua Cohen (1993, pp. 232–233) explains reasonable persuasion, 'people have the capacity to change their minds when they hear reasons presented, and sometimes they exercise that capacity.' He adds: 'This is the assumption of Brandeis's remark [Whitney v. California, 274 U.S. (1927) 377] that "if there be time to expose through discussion the falsehood and fallacies, to avert the evil by the process of education, the remedy to be applied is more speech, not enforced silence"' (Cohen, 1993, p. 233).

Along these lines, Shiffrin (2003, p. 1160) argues that:

> Confrontations with misguided views provoke audiences to reconsider their judgments and to reassess the foundations of their convictions. Negative audience reactions to them provoke reconsideration and reassessment by *speakers* … Ongoing public confrontation and reaction to other citizens' good-faith visions of how we should live together is central to our way of both discovering and understanding our convictions of how we should go on.

---

[15] The 'marketplace of ideas' metaphor, and its assumption that people can be argued out of hostility, has been much debated. See Howard (2019, pp. 10–11), who cites Brietzke (1997) as a good summary of criticisms of the metaphor. Mengistu (2012, p. 358) comments:

> The very metaphor of the 'marketplace of ideas' implies that government should regulate freedom of speech in certain circumstances, viz ., where 'market failures' or irregularities exist. Principles of substantive equality point to a market failure requiring government intervention if one group cannot get its ideas heard in the marketplace not because of the merits of those ideas but because of structural imbalances in society … The prohibition of hate speech is the equivalent of an antitrust law that removes from the marketplace a cartel and the resulting abuse of monopoly that squelches competition.

For a contrary argument, see Strossen in Molnar, 2012b.

In the marketplace of ideas, Waldron (2012a, p. 198) writes, 'everything must be up for grabs—or, more soberly, everything must be open to debate and challenge in a free and democratic society, no matter how important the objects of challenge seem to be to the culture and identity of our community.' No individual or social group has a right to be free from expression that challenges their ideas or beliefs (Heyman, 2009, p. 180), and there is a genuine public interest in hearing extremist views in the public square:

> … because it is vital for us to know that they are held and held sufficiently strongly that some people wish to communicate them to others. We also need to know who holds these views and why they are held. We can only respond intelligently to undesirable extremist attitudes, and remove or reduce the reasons why they are held, if we allow them, to some extent, to be disseminated (Barendt, 2009, p. 453).

Political legitimacy and representative democracy

A fourth argument is that freedom of expression is essential to political deliberation and legitimacy in a well-functioning representative democracy (Meiklejohn, 1960; Raue, 2020), so the state should restrict it very little, if at all. Particularly in new democracies, with long histories of suppression, freedom of speech needs maximum protection (Molnar, 2009, p. 263).

As Ronald Dworkin (2009, p. vii) has stated the argument:

> Fair democracy requires what we might call a democratic background: it requires, for example, that every competent adult have a vote in deciding what the majority's will is. And it requires, further, that each citizen have not just a vote but a voice: a majority decision is not fair unless everyone has had a fair opportunity to express his or her attitudes or opinions or fears or tastes or presuppositions or prejudices or ideals, not just in the hope of influencing others (though that hope is crucially important), but also just to confirm his or her standing as a responsible agent in, rather than a passive victim of, collective action. The majority has no right to impose its will on someone who is forbidden to raise a voice in protest or argument or objection before the decision is taken.

Banning 'hate speech' undermines democracy in two ways. First, democracy can only work if *every* citizen believes that their voice counts, including the voices of 'antagonists' as well as those of 'protagonists'. Secondly, it absolves the rest of us ('the audience') of the responsibility of politically challenging it. Kenan Malik (in Molnar, 2012a, pp. 89–90) comments: 'Where once we might have challenged obnoxious or hateful sentiments politically, today we are more likely simply to seek to outlaw them.'

When the state silences speech, it interferes with the freedoms of both the speaker and the audience: 'It stops both mouth and ears. It prevents a transaction between citizens' (Fried, 1992, p. 236).[16] Corey Brettschneider (2012, p. 16) highlights the trade-off involved here:

> Militant democrats contend that liberals offer no way to prevent the collapse of liberal democratic regimes. On this view, the rights protections afforded to illiberal groups might result in the spread of hateful viewpoints and the demise of liberal democratic protections, as in the case of Weimar Germany. But it is important to point out here that a clear tradeoff

[16] The question, as Jeremy Waldron puts it, is 'whether hate speech laws do actually exclude people from the political process, whether they do actually insulate certain norms of civility from challenge' (Waldron, 2012a, p. 198).

> would come from abandoning rights in order to coercively suppress hateful viewpoints. It would result in a loss to a major aspect of democratic legitimacy.[17]

In an interview, Kenan Malik (in Molnar, 2012a, p. 86) illustrated what this means for democratic governance:

> I support laws against discrimination in the public sphere. But I absolutely oppose laws against the advocacy of discrimination. Equality is a political concept, and one to which I subscribe. But many people don't. It is clearly a highly contested concept. Should there be continued Muslim immigration into Europe? Should indigenous workers get priority in social housing? Should gays be allowed to adopt? These are all questions being keenly debated at the moment. I have strong views on all these issues, based on my belief in equality. But it would be absurd to suggest that only people who hold my kind of views should be able to advocate them. I find arguments against Muslim immigration, against equal access to housing, against gay adoptions unpalatable. But I accept that these are legitimate political arguments. A society that outlawed such arguments would, in my mind, be as reactionary as one that banned Muslim immigration or denied gays rights.

Restraining the state

A fifth argument against restricting freedom of expression, as Jeffrey Howard (2018, p. 20) has summarised it, is that 'it is dangerous to give the state the power to restrict dangerous speech, since it is likely to misuse that power (either by making mistakes about what is genuinely dangerous, or by abusing the power for political purposes)'. The cure may indeed well be worse than the disease (Strossen, 2018, p. 14). Strossen (1996) had earlier argued that censorship measures, despite being allegedly designed to protect disempowered groups, are most often disproportionately used against them.

The concern here is rightful distrust of the coercive power of the state. Waldron (2012a, p. 26) acknowledges '… the massive power that the government can deploy—that the government of this country has deployed in the past and that governments all over the world continue to deploy—to suppress dissent, deflect criticism, and resist exposure of its malfeasances.' Hate speech law in apartheid South Africa, for example, was used to criminalise criticism of white domination. And

[17] Karl Popper's 'paradox of tolerance' is often quoted selectively on (in)tolerance of intolerance, omitting his elaboration of the initial proposition. His note on this is worth reading in full, as it sets a high bar for suppression of intolerance:

> Unlimited tolerance must lead to the disappearance of tolerance. If we extend unlimited tolerance even to those who are intolerant, if we are not prepared to defend a tolerant society against the onslaught of the intolerant, then the tolerant will be destroyed, and tolerance with them.—In this formulation, I do not imply, for instance, that we should always suppress the utterance of intolerant philosophies; as long as we can counter them by rational argument and keep them in check by public opinion, suppression would certainly be most unwise. But we should claim the right to suppress them if necessary even by force; for it may easily turn out that they are not prepared to meet us on the level of rational argument, but begin by denouncing all argument; they may forbid their followers to listen to rational argument, because it is deceptive, and teach them to answer arguments by the use of their fists or pistols. We should therefore claim, in the name of tolerance, the right not to tolerate the intolerant. We should claim that any movement preaching intolerance places itself outside the law, and we should consider incitement to intolerance and persecution as criminal, in the same way as we should consider incitement to murder, or to kidnapping, or to the revival of the slave trade, as criminal (Popper, 1966, p. 265 n.4).

governments in East Africa regularly shut down internet access or manipulate online conversations to control dissent—Uganda did both before the presidential election in January 2021 (Ovide, 2021).

Shiffrin (2003, p. 1159) reflects that the state, even if not dangerous, is 'a poor and deeply biased judge about what visions for stability and change are defective'. Given the state's coercive powers over anyone and everyone who lives within its jurisdiction, governmental regulation requires restraint, justification and rights of review and appeal:

> Freedom of speech is not the same as an uninhibited license to speak—to lie, to deceive, to molest, to coerce. So the fundamental postulate of distrust of government does not translate into a total ban against all government regulation of all forms of speech, but into a strong presumption that can be overridden only by establishing some compelling government interest (Epstein, 1992, p. 45).

If a government restricts speech that it thinks might persuade people to formulate 'wrong' public policy, then the government has usurped the critical democratic principle of popular sovereignty (Weinstein, 2009, p. 26). Weinstein cites James Madison (1794): 'If we advert to the nature of Republican Government, we shall find that the censorial power is in the people over the Government, and not in the Government over the people.' As Kenan Malik (2012, p. 49) puts it:

> The moral of the story is that one should be careful what one wishes for. If we invite the state to define the boundaries of acceptable speech, we should not be surprised if it is not just speech to which we object that gets curtailed.

A dilemma for human rights law

A sixth argument against restricting freedom of expression is that 'hate speech' bans can create a dilemma for human rights law because of under- and over-inclusion in drafting legislation that specifies 'protected characteristics'. Nadine Strossen (2018, pp. 106–107) asks:

> How can an enumeration of protected personal characteristics successfully steer between these two problematic alternatives: being unconstitutionally underinclusive, thus impermissibly punishing speech about some groups but not others; and being unconstitutionally overinclusive, thus unjustifiably prohibiting speech that is valuable or at least poses no realistic danger of contributing to serious harm?

In other words, legislation framed in a particular context at a particular point in time defines a list of protected categories, which may implicitly sanction discrimination against categories that are not so protected. Eric Heinze (2009b, pp. 280–281) explains:

> The premise of international human rights since the Universal Declaration of Human Rights … has been that norms of the human rights corpus can claim universal legitimacy only insofar as they can, in principle, be framed and applied so as to encompass all human beings. For sexual minorities, or any other actual or potential beneficiary group, to claim protection of norms that cannot be extended to other equally vulnerable groups should prompt the gravest ethical concerns about whether such norms properly belong within the international human rights corpus (except, again, as temporary measures in unstable states), despite the fact that hate speech bans have been endorsed within international human rights law.

This is the situation New Zealand currently finds itself in—existing human rights legislation provides sanctions against 'incitement of disharmony' on the grounds of colour, race, or ethnic or national origins but not, for example, on grounds of religion, gender, disability, sexual orientation, age or obesity.

Presumably, some empirically established threshold of hostility might be adopted to narrow the field of protected categories, but this risks 'more-victim-than-thou' jurisprudence that would only entrench discrimination (Heinze, 2009b, p. 278). On the other hand, if 'hate speech' bans were extended to include all similarly situated categories, it would involve significant and possibly draconian limitations on fundamental rights of speech and expression, with broad censorship of films, videos and publications as well as everyday speech (Heinze, 2009b, p. 279). Heinze concludes:

> Hate speech bans have no place within longstanding, stable, and prosperous democracies, which have ample means at their disposal to protect sexual minorities and other vulnerable groups from hate crime and discrimination, without having to impose inevitably arbitrary limits on speech (p. 285).

Legal efficacy

A seventh argument against restricting freedom of expression is that 'hate speech' laws do not effectively curb the harms they are intended to prevent (Strossen, 2018, Chap. 6). Reflecting on the UK's Terrorism Act 2006, Tufyal Choudhury notes that radicalisation is largely a private process:

> Public statements that encourage acts of terrorism may contribute to this process but are not central to it. Furthermore, the provisions in the legislation that aim to proscribe such statements are drafted with a degree of breadth and vagueness that increases the risks of the legislation becoming counterproductive (Choudhury, 2009, p. 464).

Proposals to regulate harmful communication therefore need to include detailed, evidence-informed discussion of whether criminalising 'hate speech' is an effective intervention in the causal chain (Baker, 2012, p. 69; M. Malik, 2009, p. 103), and whether prohibitions divert energy from the more vital activity of responding expressively to harmful communication (Baker, 2009, p. 150; 2012, pp. 72, 75).[18] Baker explains: 'Legal prosecutions focus on the wrong issues—legal requirements, legal line drawing, propriety of prosecution of this rather than other cases' (Baker, 2009, p. 151). Further, prohibiting verbal expression of ideas, values and beliefs—even if they are abhorrent—may reduce democratic self-understanding that conflict and disagreement are to be worked out politically, and not through recourse to violence (Baker, 2009, p. 152; 2012, p. 74).[19]

A law prohibiting harmful communication would not, for example, have deterred the Christchurch mosque shooter. His manifesto reportedly states explicitly that his murderous act will hopefully beget more chaos due to the illiberal policies the government would likely implement in response to it (Giraud, n.d.; Giraud, 2020).

Besides, a law is effective only to the extent that there is a public will to comply with it and state agencies are able to enforce it effectively. Special Rapporteur on Minority Rights, Rita Izsák, commented in a 2015 report to the Human Rights Council that even where 'hate speech' laws exist, implementation of the law is often poor and court cases are rare (UN General Assembly, 2015, para. 33).[20]

---

[18] see further Working paper 21/08, **Counter-speech and civility as everyone's responsibility**.

[19] The idea of democracy is not to remove inevitable disagreements and conflicts from public life but to enable us to manage them politically without recourse to domination, humiliation, cruelty or violence (Bromell, 2019).

[20] The Royal Commission of Inquiry companion report (2020, pp. 21–27) briefly discusses this issue and the three decisions applying sections 61 and 131 of the Human Rights Act 1993 and equivalent provisions in earlier legislation in New Zealand.

The vaguer the definition of 'hate speech', the less laws are likely to be enforced, because of the inevitable need to exercise discretionary judgements (Strossen, 2018, p. 140). And harmful digital communication is in any case difficult to police, for reasons discussed in Working paper 21/04, **Challenges in regulating online content**. Joshua Cohen (1993, p. 262) asks:

> How much injurious expression would actually be avoided? Would the regulation be at all effective in combating the underlying problems reflected in hate speech? Furthermore, apart from addressing these questions about the regulation itself, we need to consider the wisdom of focusing energy and attention on regulating hate speech … rather than on taking more affirmative measures to combat the harms that the regulation aims to avoid.

Conclusion

Public policy decisions about whether, when and how to regulate harmful communication in a free, open and democratic society involve values and moral principles. Political philosophy can shed light on these, to inform and guide principled and pragmatic decision making about the right thing to do, or not do.

There are arguments both for and against restricting freedom of expression, whether online or offline. Public communication that incites discrimination, hostility or violence is relatively straightforward to deal with and is in most cases already illegal under existing criminal law.[21] Casual abuse and 'mouthing off' also cause harm, and when this involves harassment or intimidation it may also be illegal under existing criminal or civil law. But in urging additional state prohibition and censorship that restrict freedom of opinion and expression, we should be careful what we wish for.

Working paper 21/07, **Striking a fair balance in regulation of harmful communication**, introduces further considerations to inform and encourage prudential balancing of freedom of expression, protection from harm, promotion of social cohesion, maintenance of public order and ensuring that the law can practically be enforced (Royal Commission of Inquiry, 2020, p. 701).

A longstanding political principle is that if it is not necessary to make a law, then it is necessary not to make a law.[22] Maleiha Malik (2009, p. 105) argues that 'we need to think more imaginatively about a range of non-legal responses to hate speech rather than relying exclusively on the criminal law.' Working paper 21/08, **Counter-speech and civility as everyone's responsibility**, discusses the use of the state's expressive and not only coercive powers, and introduces counter-speech strategies as alternatives, or complements, to prohibition and censorship.

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

[22] This saying is sometimes attributed to Charles de Secondat, Baron de Montesquieu, but I have not been able to verify this. Montesquieu certainly had some sensible things to say in *The spirit of laws* about the drafting of legislation, including the observation that: 'As useless laws debilitate such as are necessary, so those that may be easily eluded, weaken the legislation' (Secondat, 1752, Vol. 2, Book 29, Chap. 16, p. 345).

References

Amdur, R. (1980). Scanlon on freedom of expression. *Philosophy & Public Affairs, 9*(3), 287–300. Accessed December 2, 2020, from https://www.jstor.org/stable/2265118

Asante, M. (1995). Unraveling the edges of free speech. *National Forum*, *75*(2), 12.

Baker, C. (Edwin). (2009). Autonomy and hate speech. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 139–157). Oxford: Oxford University Press.

Baker, C. (Edwin). (2012). Hate speech. In M. Herz & P. Molnar (Eds), *The content and context of hate speech: Rethinking regulation and responses* (pp. 57–80). Cambridge: Cambridge University Press.

Barendt, E. (2009). Incitement of, and glorification of, terrorism. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 445–462). Oxford: Oxford University Press.

Barry, B. (1964). The public interest. In B Barry & W. Rees (Eds.*), Symposium: The public interest. Proceedings of the Aristotelian Society*, Supplementary Volumes, 38, 1–18. Accessed November 9, 2020, from http://www.jstor.org/stable/4106601

Barry, B. (1965). *Political argument.* London: Routledge & Kegan Paul.

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

Brietzke, P. (1997). How and why the marketplace of ideas fails. *Valparaiso University Law Review*, *31*(3), 951–970. Accessed November 18, 2020, from HeinOnline.

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

Choudhury, T. (2009). The Terrorism Act 2006: Discouraging terrorism. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 463–487). Oxford: Oxford University Press.

Cohen, J. (1993). Freedom of expression. *Philosophy & Public Affairs, 22*(3), 207–263. Accessed November 3, 2020, from https://www.jstor.org/stable/2265305

Dworkin, R. (2009). Foreword. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. v–ix). Oxford: Oxford University Press.

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

Epstein, R. (1992). Property, speech, and the politics of distrust. *University of Chicago Law Review, 59*(1), 41–89. Accessed November 11, 2020, from https://www.jstor.org/stable/1599933

Fiss, O. (1995). The supreme court and the problem of hate speech. *Capital University Law Review, 24*(2), 281-292. Accessed November 3, 2020, from HeinOnline.

Fraleigh, D., & Tuman, J. (2011). *Freedom of expression in the marketplace of ideas*. Thousand Oaks, CA: SAGE. Accessed November 3, 2020, from https://doi.org/10.4135/9781452275215

Fried, C. (1992). The new First Amendment jurisprudence: A threat to liberty. *University of Chicago Law Review, 59*(1), 225–253. Accessed November 3, 2020, from https://www.jstor.org/stable/1599937

Gärditz, K. (2020). Die Grenze des Sagbaren. *Legal Tribune Online,* June 22, 2020. Accessed November 13, 2020, from https://www.lto.de/recht/hintergruende/h/bverfg-beschluss-1-bvr-2459-19-grundrechte-meinungsfreiheit-beleidigung-grenze/

Giraud, D. (n.d.). Prof. Paul Spoonley doesn't believe in the multicultural society. Blog post, Free Speech Coalition. Accessed October 19, 2020, from https://www.freespeechcoalition.nz/blog

Giraud, D. (2020). Who exactly would speech restrictions be protecting? *Stuff*, January 29, 2020. Accessed October 19, 2020, from https://www.stuff.co.nz/national/politics/opinion/119089031/who-exactly-would-speech-restrictions-be-protecting

Goffman, E. (1959). *The presentation of self in everyday life.* Garden City, NY: Doubleday.

Heinze, E. (2009a). Wild-West cowboys versus cheese-eating surrender monkeys: Some problems in comparative approaches to hate speech. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 182–203). Oxford: Oxford University Press.

Heinze, E. (2009b). Cumulative jurisprudence and hate speech: Sexual orientation and analogies to disability, age, and obesity. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 265–285). Oxford: Oxford University Press.

Heyman, S. (2009). Hate speech, public discourse, and the First Amendment. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 158–181). Oxford: Oxford University Press.

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

Jefferson, T. (1820). Letter to William Roscoe, December 27, 1820. Accessed November 2, 2020, from http://tjrs.monticello.org/letter/387

Karpman, S. (1968). Fairy tales and script drama analysis. *Transactional Analysis Bulletin, 26*(7), 39–43. Accessed November 19, 2020, from https://calisphere.org/item/b18aab11-07dd-49b1-8a7c-29bad8a18bd9/

Karpman, S. (1973). 1972 Eric Berne Memorial Scientific Award Lecture. *Transactional Analysis Journal, 3*(1), 73–76. https://doi.org/10.1177/036215377300300118

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

Latané, B., & Darley, J. (1970). *The unresponsive bystander: Why doesn't he help?* Englewood Cliffs, N.J. : Prentice-Hall.

Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018). Perceiving threat and feeling responsible: How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *Studies in Communication and Media*, *7*(4), 555–579. https://doi.org/10.5771/2192-4007-2018-4-555

Madison, J. (1794). James Madison to James Monroe, December 4, 1794. James Madison Papers at the Library of Congress. Accessed November 30, 2020, from http://hdl.loc.gov/loc.mss/mjm.05_0799_0804

Malik, K. (2012). Enemies of free speech. *Index on Censorship, 41*(1), 40–53. https://doi.org/10.1177/0306422012440233

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

Meiklejohn, A. (1960). *Political freedom: The constitutional powers of the people.* New York: Harper.

Mengistu, Y. (2012). Shielding marginalized groups from verbal assaults without abusing hate speech laws. In M. Herz & P. Molnar (Eds), *The content and context of hate speech: Rethinking regulation and responses* (pp. 352–377). Cambridge: Cambridge University Press.

Michelman, F. (1992). Liberties, fair values, and constitutional method. *University of Chicago Law Review, 59*(1), 91–114. Accessed November 11, 2020, from https://www.jstor.org/stable/1599934

Mill, J. (1859). *On liberty.* Cambridge: Cambridge University Press, 2011. Accessed November 2, 2020, from https://doi.org/10.1017/10.1017/CBO9781139149785

Molnar, P. (2009). Towards improved law and policy on 'hate speech': The 'clear and present danger' test in Hungary. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 237–264). Oxford: Oxford University Press.

Molnar, P. (2012a). Interview with Kenan Malik. In M. Herz & P. Molnar (Eds), *The content and context of hate speech: Rethinking regulation and responses* (pp. 81–91). Cambridge: Cambridge University Press.

Molnar, P. (2012b). Interview with Nadine Strossen. In M. Herz & P. Molnar (Eds), *The content and context of hate speech: Rethinking regulation and responses* (pp. 378–398). Cambridge: Cambridge University Press.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior, 39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5

Nagel, T. (2002). *Concealment and exposure, And other essays*. Oxford: Oxford University Press.

Ovide, S. (2021). What internet censorship looks like. *New York Times,* January 21, 2021. Accessed January 25, 2021, from https://nyti.ms/3ix2p3d

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.

Popper, K. (1966). *The open society and its enemies: Vol. I. The spell of Plato* (5th ed.). Princeton, NJ: Princeton University Press.

Raue, S. (2020). Rettet die Meinung! *Frankfurter Allgemeine Zeitung,* October 22, 2020. Accessed November 16, 2020, from https://www.faz.net/-gsb-a4ney

Raz, J. (1986). *The morality of freedom*. Oxford: Clarendon Press.

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

Scanlon, T. (1972). A theory of freedom of expression. *Philosophy & Public Affairs, 1*(2), 204–226. Accessed November 3, 2020, from https://www.jstor.org/stable/2264971

Scanlon, T. (1979). Freedom of expression and categories of expression. *University of Pittsburgh Law Review, 40*, 519–550. Accessed November 3, 2020, from HeinOnline.

Schmitt, J. (2017). Online hate speech: Definition und Verbreitungsmotivationen aus psychologischer Perspektive. In K. Kasper, L. Gräßer, & A. Riffi (Eds.), *Online hate speech: Perpektiven auf eine neue Form des Hasses* (pp. 51–56). 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

Secondat, C. de (1752). *The spirit of laws. Translated from the French of M. de Secondat, Baron de Montesquieu. In 2 volumes*. 2nd ed., transl. Nugent. London: J. Nourse & P. Vaillant.

Shiffrin, S. (2003). Speech, death, and double effect. *New York University Law Review, 78*(3), 1135–1185.

Strossen, N. (1996). Hate speech and pornography: Do we have to choose between freedom of speech and equality? *Case Western Reserve Law Review, 46*(2), 449–478. Accessed December 16, 2020, from https://core.ac.uk/download/pdf/214114548.pdf

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

UN General Assembly. (1948). Universal Declaration of Human Rights. Accessed October 30, 2020, from https://www.un.org/en/universal-declaration-human-rights/

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

Waldron, J. (2012b). Hate speech and political legitimacy. In M. Herz & P. Molnar (Eds), *The content and context of hate speech: Rethinking regulation and responses* (pp. 329–340). Cambridge: Cambridge University Press.

Weinstein, J. (2009). Extreme speech, public order, and democracy: Lessons from *The Masses*. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 23–61). Oxford: Oxford University Press.